Evaluating the Psychometric Properties of the Early Literacy and Numeracy Observation Tool (ELNOT)

by

Humeyra Nur Celebi

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Arts

in

Psychology

Carleton University

Ottawa, Ontario

Abstract

The present study evaluated the psychometric properties of the Early Literacy and Numeracy Observation Tool (ELNOT), a tool aimed at identifying children who are at-risk or gifted, and monitoring children's academic progress. Kindergarteners ($N = 356$; 173 junior kindergarteners) were assessed at two time points (fall and spring). ELNOT consists of four literacy and six numeracy tasks; some tasks were eliminated from the analyses due to floor or ceiling effects. The measurement precision of three tasks (*Print Concepts*, *Letter Identification*, *Number Recognition*) was examined through item analyses. Overall, these tasks demonstrated sufficient measurement precision for most junior kindergarteners and some senior kindergarteners. The measurement invariance of five tasks (*Print Concepts*, *Letter Identification*, *Early Spelling*, *Number Naming*, *Magnitude Comparison*) was examined across gender, kindergarten level, and time through multiple-group confirmatory factor analyses; the model structure was invariant across all comparisons. ELNOT can accurately identify junior kindergarteners who are potentially at-risk.

Acknowledgements

There are so many amazing people for whom I am grateful. My supervisors Dr. Monique Sénéchal and Dr. Jo-Anne LeFevre: I will never forget your kindness and support. Thank you for guiding me in every step of this journey. It was an honour to be your student.

My committee members Dr. Deepthi Kamawar and Dr. Tamara Sorenson Duncan: For the few times we have seen each other, I felt nothing else but warmth. Thank you for your valuable feedback and support.

Dr. Craig Leth-Steensen: Thank you for sharing your knowledge. I never once felt anxious when I was learning new analytic procedures with you. Thank you for walking me though each and every step of the analyses.

Stephanie Pagan, the research officer of the Ottawa-Carleton District School Board: Thank you for giving me the opportunity to work on this project and patiently waiting for this thesis to be completed.

The Alaca family, especially my loves Betul and Zahide: Thank you for all your support, which I do not have enough space to count. You will always be in my duas.

My family: Dostum, annem, my dearest mother Gulten Celebi; my mentor, babam, caAnım daddy the best Selahattin Celebi. Thank you for always being there for me and believing in me. I am forever grateful for your support. Sizler benim “iyi ki”lerimsiniz. May Allah be pleased with you and bless you in this world and the hereafter. Ameen. My sister, ablam Seyda Nur Celebi. Thank you for being with me all through my university studies and being a source of support that I could always count on. My sister, ablam Saliha Celebi. Thank you for being my sir arkadasi. And my roommate, little miss Betul Beyza Celebi. Thank you for being understanding even when you could not yet comprehend what this thing called thesis is. Sizi seviyorum.

Table of Contents

List of Tables

List of Illustrations

List of Appendices

Evaluating the Psychometric Properties of the Early Literacy and Numeracy Observation Tool (ELNOT)

Early literacy and numeracy skills serve as the building blocks for later academic success (Aunio & Räsänen, 2016; Desoete & Grégoire, 2006; Krajewski & Schneider, 2009; Lonigan et al., 2016; Schatschneider & Westberg, 2008; Storch & Whitehurst, 2002). Early literacy refers to children's conceptual and procedural knowledge about reading and writing (Sénéchal et al., 2001), and early numeracy refers to children's knowledge of symbolic numbers and their associations to quantities (LeFevre et al., 2010). The skills that children develop before they start formal education in grade 1 are fundamental for developing proficient reading, writing, and mathematical skills. However, in kindergarten, there are only a few instruments that assess both early literacy and numeracy that are available to educators. The goal of this thesis is to evaluate the psychometric properties of a teacher-administered observation tool that would allow teachers to track kindergarten children's progress in early reading, writing, and mathematics.

Early Literacy

Sénéchal and colleagues (2001) proposed a model of early literacy that distinguished conceptual and procedural literacy knowledge. Conceptual knowledge about literacy encompasses children's understanding of the functions of print, whereas procedural knowledge about literacy encompasses children's understanding of the mechanics of reading and writing. The most studied examples of conceptual knowledge include the concepts of knowing that print conveys the message, as well as the understanding of word and letter concepts. The most studied examples of procedural knowledge include children's knowledge of the alphabet, their awareness of the phonological structure of oral language, and their early attempts at spelling.

Individual differences in reading skills are established early (e.g., Francis et al., 1996; Torgesen & Burgess, 1998). For instance, reading skills in grade 1 uniquely predicted 27% of the variance in reading comprehension 10 years later, after accounting for general cognitive ability (Cunningham & Stanovich, 1997). Importantly, differences in children's reading skills in grade 1 seem to result from differences in the building blocks of reading already apparent during the kindergarten years or earlier (Ortiz et al., 2012; Ozernov-Palchik & Gaab, 2016). Both conceptual and procedural knowledge have been shown to be significant and independent predictors of later reading achievement for children in preschool and kindergarten. The following section is a review of these building blocks with special attention to longitudinal research. Specifically, I will review the literature on the three foundational early literacy skills: learning the concepts of print, knowing the names and sounds of letters, and distinguishing between the different sounds of the oral language, as well as young children's early spelling and reading.

Print Concepts

From very early on, children encounter print in different settings (e.g., on signs and labels, in books) and observe adults use print for various purposes (e.g., making lists, writing letters). They begin to develop their own concepts about print (Clay, 1991). These concepts about print represent their perception of what reading and writing mean, as well as their knowledge of the functions and conventions of written language. Children's perception of the meaningfulness of print includes recognizing that the labels on a cereal box or the print on a stop sign convey meaning (i.e., identifying an object or action), understanding that speech can be written down and in turn print represents speech, and knowing that written messages can be reread to share with others. Children's knowledge of the functions and conventions of print include understanding the terminology used for written language, knowing the behaviours involved in book orientation, and differentiating between written text and drawing (Sulzby et al., 1989). Children continue to develop print concepts during or after kindergarten. For example, children can differentiate between a picture and written text earlier than they can identify the function of punctuation marks. In fact, children typically master print concepts after second or third grade (Beach & Robinson, 1992), which is around the same time that children become relatively proficient writers and fluent readers, according to provincial standards (i.e., Ontario curriculum expectations; Ontario, 2008).

Clay's *Concepts about Print* (CAP; Clay, 2013) instrument has been used extensively to assess children's knowledge of print-related concepts (e.g., Dickinson & Snow, 1987; Haden et al., 1996; Hecht & Close, 2002; Neumann, 2018). For example, it assesses children's understanding that it is the text, not the pictures, that are being read. Print concepts are predictive of early reading achievement. Children's performance on the CAP task has been shown to be indicative of their developmental progress in reading and writing (Clay, 2013). Children's performance on the CAP task at four years of age was correlated with their reading achievement at ages 7 and 11 (Tizard et al., 1988). A meta-analysis by the National Early Literacy Panel (NELP; 2008) showed that print concepts have moderate ($.30 < r < .50$) to strong ($r \geq .50$) predictive relations with later literacy outcomes. Across 12 studies with 2,604 children in preschool or kindergarten, print concepts yielded an average moderate relation with later decoding skills ($r = .34$). Similarly, across four studies involving 534 children, print concepts had an average moderate correlation with later spelling ($r = .43$). Furthermore, across three studies with 535 children, print concepts were, on average, strongly related to later reading comprehension ($r = .54$). Print concepts continued to be a significant predictor when age and language were controlled. Nevertheless, when variables such as alphabet knowledge and phonological awareness are controlled, print concepts did not uniquely predict reading performance (NELP, 2008).

Alphabet Knowledge

Alphabet knowledge is the knowledge of the names and sounds of the letters in one's language (Puranik et al., 2011). Children first learn the names and shapes of letters that they are most exposed to, such as the letters in their names, and build upon this knowledge between the ages of 3 and 5 (Bloodgood, 1999; Treiman & Broderick, 1998). The assessment of letter-name and letter-sound knowledge typically involves untimed recognition or production of letter names and sounds, respectively (Piasta & Wagner, 2010). For example, children may be asked to produce the associated names of letters that are presented on cards. It is important to note that producing letter names (such as singing the alphabet song) is not the same as being able to recognize or label them out of sequence. As such, letters are usually presented to children in a random order on flashcards to differentiate those who merely recite the letters in the alphabet song from those who actually recognize them. There are also assessment tools that incorporate both uppercase and lowercase letters, which tests children's ability to identify both forms of the letters. However, in an attempt to shorten an assessment of letter name knowledge, it could be argued that including just lowercase letters as opposed to presenting both forms can provide an equally effective assessment of children's ability (Bowles et al., 2014).

In alphabetic languages, letters are the building blocks of literacy skills, and thus, learning about letters and their properties is an important foundation for reading and writing (Snow et al., 1998). Accordingly, knowledge of letter names and sounds is a consistent predictor of children's later reading and spelling abilities (Hammill, 2004; Scarborough, 1998; Schatschneider et al., 2004), and is independent of other important predictors including phonological awareness and oral language (Burgess & Lonigan, 1998; McBride-Chang, 1999; Wagner et al., 1994). In fact, a meta-analysis reveals that alphabet knowledge has moderate ($.30 < r < .50$) to strong ($r \geq .50$) predictive relations with later literacy outcomes, and maintains its predictive power even after variables such as age, socioeconomic status, oral language, phonological awareness, and IQ are controlled (NELP, 2008). On average, alphabet knowledge was moderately correlated with later reading comprehension in 17 studies ($r = .48$) and highly correlated with later decoding in 52 studies ($r = .50$) and with later spelling in 18 studies ($r = .54$). Furthermore, children who have poor alphabet knowledge at preschool and kindergarten are more likely to experience difficulties with learning to read and be identified as having reading disabilities (Gallagher et al., 2000; O'Connor & Jenkins, 1999; Torppa et al., 2006).

Understanding the link between letters and their sounds seems to be particularly important – knowledge of letter-sounds is crucial for decoding of words, and individual differences among children in reading comprehension are often related to skill differences at the word level (Perfetti, 2007). Furthermore, children who are poor at identifying the sounds of visually presented letters are at higher risk of having reading problems (e.g., Hulme et al., 2012; Schatschneider et al., 2004). In contrast, children's reading performances improve when the links between letters and sounds are emphasized during instruction (e.g., McArthur et al., 2018).

Phonological Awareness

Phonological awareness is the ability to identify and manipulate units of speech, at the word, syllable, or phoneme level. It typically begins to develop between the ages of 4 and 5 and continues to improve over time (Burgess & Lonigan, 1998). Children first recognize larger units such as syllables and onset-rime before smaller units such as individual phonemes. In addition, with respect to phonemes, children are able to identify beginning and ending sounds of words before medial sounds (Anthony & Francis, 2005; Blachman, 2000; Carroll et al., 2003). For example, children can identify that the word "cat" begins with /k/ and ends with /t/ before being able to identify that the middle sound is /a/.

A wide range of tasks are used to assess children's phonological awareness, including rhyming, isolating the sounds in the beginning, middle, and end of words, sounding out words (e.g., cat – /k/ /æ/ /t/), and saying words with certain sounds deleted (e.g., cat without saying /k/ – at; Blachman, 2000; Spector, 1992). The common element in all these tasks is that they all require the child to focus on the underlying phonological structure of the spoken word, not the meaning of the word (Blachman, 2000).

There is extensive evidence showing that children's performance on measures of phonological awareness predict their future reading and writing achievement. As reported in the meta-analysis by NELP (2008), similar to alphabet knowledge, phonological awareness maintains its predictive power even after individual differences among children such as IQ and socioeconomic status are controlled. On average, when assessed in preschool or kindergarten, phonological awareness was moderately correlated with later decoding across 69 studies involving 8,443 children (r = .40), reading comprehension across 20 studies involving 2,461 children (r = .44), and spelling across 21 studies involving 2,522 children (r = .40). Catts and colleagues (2017) found that children who lacked phonological awareness skills in kindergarten were five times more likely to become poor readers in grade 2 compared to children who have adequate phonological awareness skills.

Early Spelling

Before children learn to spell, that is, to map letter sounds to symbols using orthographic conventions of their language, young children's attempts at spelling words may yield phonologically accurate but orthographically inaccurate outputs (e.g., “ez” for “easy”). These attempts are often referred to as invented spelling (Read, 1971). Children’s invented spellings demonstrate their level of understanding of basic early literacy skills such as print concepts, alphabet knowledge, and phonological awareness (Bear & Templeton, 1998; Gentry & Gillet, 1993). At the beginning, most children capture only the first sounds in words correctly, for example, writing “b” for bike (Gentry & Gillet, 1993). Then, they typically learn to correctly spell the final sounds in words (e.g., writing “bk” for bike) before gradually learning to spell the medial sounds. Finally, as they receive training in reading and spelling, they achieve conventional spelling skills.

Accordingly, a meta-analysis found that invented spelling has strong ($r \geq .50$) predictive relations with later literacy outcomes (NELP, 2008). Ten studies that assessed invented spelling in preschool or kindergarten children ($N = 778$) showed an average correlation of .58 with later decoding skills. Across three studies with 354 children, invented spelling had an average correlation of .69 with later spelling skills. Intervention studies have found that facilitating invented spelling had an additional impact on children’s word reading compared to phoneme awareness instruction alone (Ouellette & Sénéchal, 2008; Sénéchal et al., 2012). For example, studies have found that English-speaking kindergarten non-readers (Ouellete & Sénéchal, 2008) and at-risk kindergarten children (Sénéchal et al., 2012) learned more novel words when assigned to an invented spelling group compared to children in other conditions, such as phoneme awareness training. Invented spelling integrates phoneme and orthographic representation of words and can support at-risk children in developing their reading skills (Sénéchal et al., 2012).

Children's ability to spell is foundational for both writing and reading (Graham et al., 2002; Treiman, 1998). As reported in the meta-analysis by NELP (2008), spelling skills assessed in preschool and kindergarten have been found to have strong ($r \geq .50$) predictive relations with later literacy outcomes. Across seven studies with 1,184 children in preschool or kindergarten, spelling had an average of .60 correlation with later decoding skills. Similarly, across four studies involving 398 children, spelling in kindergarten or earlier had, on average, a strong correlation with later spelling ($r = .78$). However, Sénéchal (2017) found that accurate spelling at the beginning of grade 1 was no longer predictive of reading at the end of grade 1 after accounting for children's early word reading skills at beginning of grade 1. She argued that appropriate reading skills are required for accurate orthographic spelling.

As briefly mentioned, children's invented spelling can become more sophisticated as they develop the skill. As such, studies have examined whether the sophistication of kindergarten students' invented spelling is predictive of their reading and spelling in grade 1. In one longitudinal study, Ouellette and Sénéchal (2016) used path modelling to determine whether there was a causal role of invented spelling in kindergarten students' later literacy outcomes. Findings demonstrated that invented spelling influenced children's subsequent reading and alphabetic knowledge, while also mediating the connection between phonological awareness and early reading. Similarly, Sénéchal (2017) found that the sophistication of students' invented spelling mediated the relationship between phoneme awareness and letter knowledge and early reading. These findings continue to support the understanding that invented spelling has a unique contribution to children's literacy outcomes that is separate from phonological awareness.

Early Reading

The goal of reading is to construct meaning from what has been read. The three stages of reading development that Chall (1983) proposed were pre-reading, decoding, and fluency stages. In the pre-reading stage, children develop an understanding of the nature of reading – that speech can be written down and read by someone else, which is the same as perception of the meaningfulness of print mentioned under the print concepts section. In the decoding stage, children learn letter-sound relations and phonological awareness. Teachers play an important role at this stage in terms of helping children understand the complex symbol system. In the fluency stage, children are able to recognize words with ease while understanding their meaning. Although it is dependent on the level of formal instruction children receive, they typically run through this stage in second and third grades. However, it is important to assess children's early reading before they enter elementary school so that children at risk of falling behind their peers when receiving formal instruction may be identified and provided with additional help.

Assessment of young children's reading typically involves decoding, sight word recognition, and oral reading fluency (Piasta & Wagner, 2010). Decoding words can be described as mapping letter sounds with their representations to read real words or as using orthographic knowledge to read sight words. Word decoding is typically assessed using a standardized test such as the word identification subtest of the Woodcock Reading Mastery Test. As reported in a meta-analysis, preschool and kindergarten children's performances in decoding words had moderate ($.30 < r < .50$) to strong ($r \geq .50$) predictive relations with later literacy outcomes (NELP, 2008). Across 21 studies involving 4,121 children, decoding words in kindergarten or earlier had an average correlation of .52 with later decoding skills. Across six studies involving 1,091 children, decoding words had an average correlation of .40 with later reading comprehension skills. Across six studies with 1,112 children, decoding words had an average correlation of .54 with later spelling skills.

A few children enter preschool, kindergarten, and first grade as readers (Clark, 1976; Durkin, 1966). In other words, without receiving any formal instruction, these children can decode words and read with comprehension at or above the level of second graders (Olson et al., 2006). Given that early readers make up only a small percentage of the population, research on these children is limited with respect to their later academic success. In a longitudinal study, Stainthorp and Hughes (2004) assessed children's literacy skills such as reading accuracy, speed and comprehension, as well as phonological awareness, at 5, 6, 7 and 11 years of age ($N = 28$; 14 early readers). Across all time points, the early readers maintained their significantly advanced abilities compared to the nonearly readers. Although the gap between the two groups on reading accuracy slightly narrowed from age 5 to 11, the early readers continued to have great advantage over the nonearly readers.

The goal of reading, as mentioned earlier, is to understand the meaning of written texts, which is referred to as reading comprehension. Children's reading comprehension skills are typically assessed with standardized tests such as the passage comprehension subtests of the Woodcock Reading Mastery Test (NELP, 2008). As reported in the meta-analysis, across five studies with 700 children in preschool or kindergarten, reading comprehension skills had an average strong correlation with later decoding skills ($r = .52$).

Conclusion

Children's early literacy skills are highly predictive of their future literacy skills; as such, it is essential to provide children with appropriate literacy instruction during their early years as well as to identify struggling readers at an early age for early intervention. Although there are existing tools that identify children's knowledge in the areas of print concepts, alphabet knowledge, phonological awareness (distinguishing between the different sounds of the oral language), as well as early spelling and reading skills, these well-established tools are costly and may take up too much time in the classroom for teachers and educators to administer. In conclusion, teachers and educators would benefit from a low cost, easily administered tool that assesses children's literacy skills to appropriately target instruction in the classroom to their individual needs.

Early Numeracy

Early numeracy skills form the basis for later math learning. Some of the key skills include counting skills and understanding number relations (Aunio & Niemivirta, 2010; Purpura & Reid, 2016), identifying number symbols (Göbel et al., 2014; Pinto et al., 2016), and differentiating numerical magnitudes (LeFevre et al., 2010; Merkley & Ansari, 2016; Muldoon et al., 2013). Individual differences in numeracy skills are established early (Klibanoff et al., 2006). For example, children’s ability to identify numbers and discriminate between quantities at the end of kindergarten strongly predicted their mathematics outcomes at the end of grade 1 (Chard et al., 2005; Clarke & Shinn, 2004). The predictive relationship between early numeracy skills and later math achievement continues in subsequent grades; higher levels of number competence in kindergarten predicted composite mathematics achievement at the end of grade 3 (Jordan et al., 2009). The following section is a review of the foundational precursor skills for mathematics development in children, with special attention given to longitudinal research. These precursor skills (verbal counting, cardinality, number identification, and magnitude comparison) have been proven extensively in research to be predictive of children's later mathematical skills.

Verbal Counting

Verbal counting, that is, reciting the numbers in order out loud, is an important skill for young children because it is their first introduction to the symbolic number system (Gelman & Gallistel, 1986). From 2- to 3-years of age, children recite number words through memorization without being aware of the relationship between those number words and their corresponding quantities (Wynn, 1992). Most 3-year-olds learn to count to ten in the correct order before they can recognize the symbolic representation of these numbers (Mix, 2002, 2009). Because numbers 0-19 have unique names in English, children can only learn them through memorization. However, once they figure out the counting system of up to 20, they learn the rest of the counting sequence relatively easily.

Gradually, children link number words to other representations such as quantities and digits. Verbal counting thus forms the foundation of numerical abilities such as the cardinality principle, that is, mapping number words to quantities (Jiménez Lira et al., 2017) and, more generally, counting aids children's understanding of the link between quantitative information of numbers and their symbolic representations (Xu, 2018).

Researchers typically assess verbal counting ability by asking children to count as high as they can starting from one (e.g., Clarke & Shinn, 2004; Purpura & Lonigan, 2015). As a core numeracy skill, verbal counting skills are one of the best predictors of later mathematics performance (Aubrey et al., 2006; Aubrey & Godfrey, 2003; Kavkler et al., 2003). Kindergarten children's more sophisticated counting skills, which require knowing the counting string, such as number-word correspondence and enumeration skills, have been found to predict basic arithmetic skills in the early years of elementary school (Aunola et al., 2004; Desoete et al., 2009; Jordan et al., 2007; Koponen et al., 2007; Kurdek & Sinclair, 2001; LeFevre et al., 2006).

For example, object counting plays an explicit role in arithmetic because children use counting strategies in their early calculations (Geary, 2004). Counting eventually becomes a more automatic process, allowing children to use their cognitive resources towards more complex mathematical tasks (Gersten & Chard, 1999; Resnick, 1989).

Cardinality

Cardinality means that the final number word in a counting sequence represents the quantity of the set (LeFevre et al, 2006). It takes quite a long time for children to learn how number words map onto exact, cardinal numbers (Briars & Siegler, 1984; Frye et al., 1989; Fuson, 1988; Le Corre et al., 2006; Sarnecka & Carey, 2008; Wagner & Walters, 1982). After learning to recite the counting sequence by rote, children learn the numerical meanings of the first few number words, one at a time, in sequential order (Le Corre et al., 2006; Sarnecka & Lee, 2009; Wynn, 1990, 1992). Their progress can be observed clearly through the Give-N task (Frye et al., 1989; Le Corre et al., 2006; Sarnecka & Lee, 2009; Wynn, 1990, 1992), where children are asked to form a set of a given number (e.g., “Can you give three apples to the puppet?’’). The predictive value of the Give-N task is supported by many studies, both longitudinal (Wynn, 1992) and cross-sectional (Condry & Spelke, 2008; Le Corre & Carey, 2007; LeCorre et al., 2006; Sarnecka & Gelman, 2004; Sarnecka et al., 2007; Schaeffer et al., 1974; Wynn, 1990).

By age 3, most English-speaking children only understand the meaning of the number “one” (Sarnecka & Carey, 2008). Thus, during the Give-N task, the “one-knowers” give one object when asked for one, and generally give two or more objects when asked for any other numeral. Several months later, children develop an understanding of “two”, at which point they give two objects when asked for two (i.e., “two-knowers”), and give three or more objects for higher numbers. They later learn the meaning of three (i.e., “three-knowers”), then four (i.e., “four-knowers”) over several months. Children at any of these levels are referred to as subset-knowers (Slusser et al., 2013) because even though they can recite the number words in order to ten or higher, they know the exact, numeral meanings of only a subset of those number words (Le Corre & Carey, 2007; Le Corre et al., 2006; Sarnecka & Carey, 2008).

By age 4, many children have developed an understanding of the cardinal principle for sets up to five or six (Le Corre & Carey, 2007; Le Corre et al., 2006; Sarnecka & Carey, 2008; Shusterman et al., 2010). At this point, they are referred to as cardinal principle knowers (CP-knowers). CP-knowers know the meaning of every number in their count list and, more generally, that the cardinal meaning of any number word is dependent on that word’s position in the counting list (Gelman & Gallistel, 1978). CP-knowers can demonstrate their knowledge by generating the set size that corresponds to any number they can count to. It is important to note that what separates the CP-knowers from the subset-knowers is not just the size of the sets that they can produce; rather, it is that CP-knowers understand the principle of counting, whereas subset-knowers do not (Sarnecka & Carey, 2008). In summary, by age 5 or 6, children understand how number words are connected to quantities (i.e., they have cardinal principle knowledge and enumeration skills). They can then learn to attach this quantity knowledge to Arabic symbols (Case et al., 1996) and shift their focus to the cardinality of a set, that is, the symbolic digit representation, as opposed to the number’s physical representation as a concrete object (Bialystok, 1992).

Another task that is used to assess children’s understanding of the cardinal principle is magnitude comparison of sets (Krajewski & Schneider, 2009; Sarnecka & Wright, 2013). Children need to have an understanding of both cardinality and equinumerosity, that is, “only sets whose items can be placed in one-to-one correspondence have the same number of items" in order to succeed in this task (Sarnecka & Wright, 2013). Therefore, this task requires more knowledge on children's part and differentiates CP-knowers from non-CP-knowers.

Acquiring cardinal knowledge is critical in the development of quantitative skills (Brainerd, 1979; Fuson, 1988; Gelman & Gallistel, 1978; Sarnecka & Carey, 2008; Wynn, 1990). For example, preschoolers' performance on the Give-N task in the fall predicted 48% of the variance in their mathematics achievement at the end of the school year (van Marle et al., 2014). Similarly, children who had not yet developed the cardinal value of at least three at the beginning of preschool performed significantly worse on a complex task that required the addition of two or more cardinal values (represented with symbolic or non-symbolic stimuli) in kindergarten compared to CP-knowers (Moore et al., 2016). This relationship was still in place after controlling for variables such as executive functions, intelligence, and number naming fluency. However, the Give-N task becomes less sensitive in detecting individual differences in cardinal knowledge as children get older. By age 4, most children are three-knowers (Moore et al., 2016), and by age 5, the majority are CP-knowers (Sarnecka & Carey, 2008). Therefore, a relatively more difficult task (e.g., point-to-x; Wynn, 1990) may be more appropriate in measuring the cardinal knowledge of children who are already CP-knowers.

Number Identification

Children build on their simple counting knowledge by learning the symbolic representation of numbers (1, 2, 3, …) that correspond to each number word (one, two, three, …) in the count sequence (Bialystok, 1992). Children's ability to map number words and digits is typically assessed through numeral identification tasks (e.g., Hornburg et al., 2018; Jiménez Lira et al., 2017; Purpura & Lonigan, 2015), where children are given the names of the numbers (e.g., "Show me seven") and asked to point to the corresponding Arabic digits (e.g., 7) or presented with Arabic digits and asked to say the corresponding names. Both tasks speak to the mapping of number words and digits, but in different directions: The former tests children's ability to map number words onto digits, while the latter tests their ability to map digits onto number words (Jiménez Lira et al., 2017). Typically, children are equally good at both directions of mapping (Hurst et al., 2017).

The beginning of children's numeracy development may be marked with their ability to differentiate between numerals and other signs and symbols (e.g., knowing that numbers and letters are different), which is followed by learning the names of numbers (Krajewski & Schneider, 2009; Sarama & Clements, 2009). At this point, they do not understand that digits hold quantitative meanings as well. Children develop an understanding of the link between number words and digits during preschool: Approximately 25% of children correctly identified numbers 1 to 9 by age four (Ginsburg & Baroody, 2003). In another study, four-year-old children had an average of 40% success rate in naming digits from 0 to 9, while five- and six-year-olds had approximately a 60% and 90% success rate, on average (Berteletti et al., 2010).

Studies have shown that, alongside counting, kindergarten children's numeral identification ability of numbers between 0-20 (Clarke & Shinn, 2004), 0-100 (Lembke & Foegen, 2009), or greater (Cirino, 2011) is important for their mathematics growth in grade 1. In addition, children's ability to identify one-, two-, and three-digit Arabic numerals at school entry were found to be a strong predictor of growth in arithmetic skills during the following year (Göbel et al., 2014).

Magnitude Comparison of Digits

Magnitude comparison of digits is the ability to map digits to quantities (Bialystok, 1992) in order to decide which digit represents the larger quantity. Numerical magnitude processing is frequently measured using number comparison paradigms in which participants choose the larger of the two numbers. Individuals respond more quickly and accurately when comparing numbers that are numerically more distant (e.g., 5 vs 9) than when they are closer (e.g., 5 vs 6; Moyer & Landauer, 1967).

Research across different studies and populations consistently shows positive relations between symbolic magnitude comparison skills and mathematics (e.g., De Smedt et al., 2013). Children who are accurate and efficient at magnitude comparison tend to also perform well in more complex mathematics tasks, including arithmetic (Nosworthy et al., 2013), word problems (Fuchs et al., 2010), fractions (Mou et al., 2016), and geometry (Lourenco & Bonny, 2017). In a recent meta-analysis, symbolic comparison and mathematical competence were moderately correlated (r = .30, averaged across 89 effect sizes; Schneider et al., 2017). This relationship is further supported through longitudinal findings of several studies (Bartelet et al., 2014; Matejko & Ansari, 2016; Vanbinst et al., 2018; Xenidou-Dervou et al., 2017). For example, children's symbolic comparison skills (digit range: 1–9) in kindergarten are a robust and consistent predictor of mathematics performance (i.e., number relations, mental arithmetic, applied calculations and measurement) in both grades 1 and 2, even after ruling out the individual differences related to IQ and working memory (Xenidou-Dervou et al., 2017).

Conclusion

Early numeracy skills serve as the building blocks of acquiring mathematics and have long term relations with children's future mathematical skills. It is therefore essential to identify struggling math learners at an early age for early intervention. Although there are existing tools that identify children's knowledge in the areas of verbal counting, cardinality, number identification, and magnitude comparison, a tool that combines both early literacy and numeracy measures would be valuable for teachers and educators. Teachers and educators could use this tool to identify children's areas of strengths and weaknesses to appropriately target instruction in the classroom to children's individual needs.

Relations Between Literacy and Numeracy in Kindergarten

The two main domains of school achievement can be identified as achievement in reading and math, and the close relationship between these domains has been supported across different levels of age (Chen & Chalhoub-Deville, 2016; Purpura & Napoli, 2015) and across gender (Abedi & Lord, 2001). Early literacy and numeracy skills seem to be highly related (Duncan et al., 2007; Hecht et al., 2001; Purpura et al., 2011; Welsh et al., 2010). A recent meta-analysis found an average large correlation (r = .55) between measures of reading and math (Singer & Strasser, 2017). For example, basic knowledge of print and mathematics (letter identification and number identification) were significantly associated with each other in preschool (Piasta et al., 2010) and kindergarten (Matthews et al., 2009). In conclusion, it is important to measure children's knowledge in both early literacy and numeracy to identify areas of need in the two domains to ensure academic achievement.

Gender

Do girls and boys differ in their early literacy or numeracy skills? There is no consensus. In terms of literacy achievement, girls seem to have an advantage in memorizing the alphabet (Lundberg et al., 2012), which is foundational to developing early literacy skills. They also perform better in some areas of reading development compared to boys (McGuinness et al., 1995). On the other hand, gender was not found to be a strong predictor of children's reading ability (Limbrick et al., 2011).

Several studies report identical primary numerical abilities between girls and boys (e.g., Dehaene, 1997, Nunes & Bryant, 1996); however, some studies indicate gender differences. For example, girls aged four to seven years outperformed boys in basic arithmetic tasks on the British National Curriculum Key Stage 1 measurements (Demie, 2001; Gorard et al., 2001; Strand, 1997, 1999). To be more precise, girls performed better than boys during Baseline measurements (i.e., 4-years-old) and Key Stage 1 measurement (i.e., 7-years-old); however, between the two measurement points, the boys showed higher progress (Strand, 1999). Similarly, Finnish girls aged 4 to 7 years (Aunio et al., 2008, Aunio et al., 2006) and Australian girls aged 5 to 6 years (Boardman, 2006) had better early numeracy skills than same aged boys. Contrary to these findings, Jordan et al. (2006) found that kindergarten boys had higher performances in the overall number sense, nonverbal calculation and estimation skills compared to girls. Yet, no clear gender differences have been found in mathematical performance in the early years of elementary school (Aubrey & Godfrey, 2003, Aunola et al., 2004, Carr & Jessup, 1997, Fennema et al., 1998).

Achievement studies where task scores are compared among different subgroups often do not assess whether the latent variables tested are functioning similarly across different subgroups (e.g., girls and boys; Dimitrov, 2010). It is important to check that each item in a task, as well as the task as a whole, have the same meaning across these groups for the results to be accurate; that is, the measurement is invariant across gender. Establishing measurement invariance for standardized tests such as the Wechsler tests is even more crucial considering their wide use around the world (Bowden, 2013). For example, some items in the Arithmetic subtest of the

Wechsler Intelligence Scale for Children – Third Edition (WISC-III; Wechsler, 1991) functioned differently across gender (Maller, 2001). These items were more difficult for girls than boys with the same latent ability level. In other words, the probability of children with the same underlying arithmetic skills passing or failing these items changes as a function of group membership (gender). Thus, it would not be accurate to state that boys have a higher level of arithmetic skills than girls based on their performances on these items. Measurement invariance across gender was established in the later versions of the WISC (e.g., WISC-V; Chen et al., 2015).

Kindergarten in Ontario

Kindergarten in Ontario is a two-year play-based program (full-day). Children in the first year of kindergarten, formerly referred to as junior kindergarten, start in September of the year they turn four, and thus range in age from 3:8 to 4:7 (years, months) when they start school. Children in the second year of kindergarten, formerly referred to as senior kindergarten, range from 4:8 to 5:7. Although these two kindergarten levels are now labelled Year 1 and Year 2, I will use the terms junior and senior kindergarten for clarity.

Research supports the benefits of the play-based approach mandated in the Ontario kindergarten program. Throughout the day, children have many opportunities to engage in literacy- and numeracy-rich activities as they play. For example, when children write their names with magnetic letters, they naturally engage in literacy behaviours involving awareness of print concepts, the alphabetic principle, and the fact that letters make sounds and the sounds carry meaning (i.e., phonological awareness). Furthermore, children spontaneously engage in mathematics behaviours during play as they measure, sort, and count (Ginsberg, 2006; Sarama & Clements, 2008; Seo & Ginsberg, 2004; Hunting, 2010).

However, to maximize children's learning, teachers must make intentional and purposeful interactions during play (Baroody et al., 2006; Thomas et al., 2011; Balfanz, 1999; Ginsburg et al., 2008). By being mindful of children's questions and interests, teachers can embed learning experiences in the classroom in a developmentally appropriate way. For example, after a child successfully spells their name with magnetic letters, a teacher can provide the child with the name cards of other students and ask them to spell those names. With regard to the cardinality principle for example, when a child talks about their age (e.g., "I am four now and I am going to be five"), teachers can situate a question that would provoke the child to compare the magnitude of two numbers (e.g., "Are you going to be older or younger on your next birthday?").

As part of the curriculum, children are expected to demonstrate their literacy and mathematics knowledge. Each learning outcome has specific expectations, and educators can assess children's learning through what they *say*, what they *do*, and what they depict or *represent*. For example, children might show that they can draw meaning from texts that they read as beginner readers by saying out loud the words that they sight (e.g., "I know that says *the*"). As beginner writers, children might show that they can communicate with others by writing letters to one another and to family members. They might also show their understanding of number concepts (e.g., counting, quantity and number relationships) through noticing that two groups of objects are the same in number (e.g., both sides of the sorting tray have the same number of objects). Thus, teachers encourage children to engage in literacy and numeracy behaviours that closely resemble real-life situations and assess children's progress by observing their behaviours and listening and asking probing questions.

In summary, the foundational early literacy and numeracy skills that research highlights (e.g., alphabet knowledge, counting) are embedded within the learning objectives of Kindergarten in Ontario. However, these expectations do not indicate where on the spectrum of skills the children should be at by the end of kindergarten, nor is there a boundary between the two levels of kindergarten. Therefore, children's scores on instruments that measure their early literacy and numeracy skills cannot be compared directly to the kindergarten curriculum. It would be helpful for teachers and educators to have access to Ontario norms or district-level norms on a psychometrically sound tool to which they could compare their students' performances.

Age and Schooling

Age and school experience may provide an advantage to children in acquiring literacy and numeracy skills. As children get older, their competencies increase (Fayol et al., 1998; Huang & Invernizzi, 2012; Jordan et al., 2006; Ransdell & Hecht, 2003). However, it is important to separate the effects of schooling from those of age-related development (Burrage et al., 2008). The former is experienced by only one group of children (e.g., senior kindergarteners), whereas the latter is experienced by all children. For example, if a group of senior kindergarteners outperform same-aged junior kindergarteners on an alphabet knowledge test in the fall (who are in a different year of schooling due to the age cut-off criteria in Ontario), this outperformance is in part due to a year of education experienced only by the senior kindergarteners. Indeed, Burrage and colleagues (2008) have found that children in senior kindergarten performed better than their same-age junior kindergarten counterparts on a word reading task both in the fall and spring. The difference in the fall performance may be interpreted as an effect of senior kindergarteners' additional schooling experience of junior kindergarten, which is not experienced by junior kindergarteners. The difference in the spring performance may be interpreted as an effect of the differential schooling experiences (growth from fall to spring) of children in junior and senior kindergarten.

Similar to comparing children's performances across gender, it is important to check that the tasks used are invariant across age or school year (Dimitrov, 2010). For example, junior and senior kindergarteners are likely to perform differently in both fall and spring. However, the differences in performances would only be meaningful if the tasks function similarly across kindergarten level. In addition, to accurately assess growth within each kindergarten level from fall to spring, the tasks must be invariant across time.

Early Identification of Children Who Need Additional Support

Children enter the kindergarten program with different levels of early literacy and numeracy skills. Depending on where they start, the experiences they have, and the instruction received in the classroom, children will achieve varying levels of growth with regards to the learning expectations of the program. Although many children succeed in meeting expectations, some young learners do not acquire well-developed early academic skills (Lonigan et al., 2011). As a result, they will have difficulty in becoming skilled in literacy and numeracy with just the typical instructions delivered in the early elementary grades. Identification of these children, who either enter kindergarten at a disadvantage or do not make sufficient progress, will ensure that they get the help they need to avoid any permanent delay in their academic skills.

Longitudinal studies show that children's reading-related skills in junior kindergarten (i.e., preschool in other countries) have a high degree of continuity in elementary school (e.g., Lonigan et al., 2000; Storch & Whitehurst, 2002); this finding reveals that the developmental precursors that underlie the acquisition of reading appear before children begin formal schooling

(Lonigan et al., 2011). Children who read well also read more, and therefore grow their knowledge in various domains (Mol & Bus, 2011). On the other hand, children who have reading difficulties read less (Allington, 2014), receive less practice in acquiring reading comprehension strategies (Paris, 2005), and develop negative attitudes towards reading (Woolley, 2011). With time, the gap in their reading skills as well as their achievement in other academic areas, becomes greater compared to their literate peers (Chall et al., 2009).

Similar to literacy outcomes, children who perform poorly in early numeracy skills are at risk of having later mathematical learning difficulties (Jordan et al., 2006). In one study, the majority of children who performed below the 10^{th} percentile in mathematics in the fall and spring of senior kindergarten had low levels of performance in the first, third, and/or fifth grades (Morgan et al., 2009). These children had difficulty in memorizing the count sequence, made errors when counting sets of objects, and had difficulty making magnitude comparisons and basic addition and subtraction; these are behaviours that teachers can observe during typical classroom activities (Aunio & Niemivirta, 2010; Desoete et al., 2009; Jordan et al., 2006). Because these differences exist prior to formal schooling (Aunio et al., 2009) and without intervention (Missall et al., 2012), both of which would be expected to produce differences, early assessment in kindergarten is necessary to identify children who need additional support.

Early identification of children at risk for developing later literacy and math challenges is crucial for many reasons. For example, the educational and occupational achievement of children who are having difficulties in these areas are greatly impacted. Indeed, a child who has trouble learning to read will have trouble in other subject areas, which may lead to developing negative attitudes towards both reading and school. Research has shown that children who are poor readers at the end of the first grade are highly likely (roughly 90% probability) to remain poor readers at the end of the fourth grade (Boyer, 1991). Furthermore, the drop-out rate of students with below-average reading skills from high school was reported to be 23 percent, which is the highest compared to students with basic reading skills (9 percent) and proficient readers (4 percent; McGencey, 2011). People with low numeracy were found to have lower incomes and poorer health outcomes, and to be more likely to have been involved with the law (Parsons & Bynner, 2005). Moreover, numerical knowledge at the age of seven is predictive of one's socioeconomic status (SES) at age 42, even after accounting for the students' IQ and the SES of the family that they were born into (Ritchie & Bates, 2013).

Several tools have been developed to assess children's early literacy and numeracy skills. For example, the Get Ready to Read! Revised Screening Tool (GRTR-R; Whitehurst & Lonigan, 2001; Lonigan &Wilson, 2008) is a 25-item task that measures some of the foundational early literacy skills including letter-name and letter-sound knowledge, directionality of print, and phonological awareness. The GRTR-R was found to have good reliability with an alpha value of .88 (Lonigan & Wilson, 2008). It also had adequate latent ability estimates, with a greater precision of measurement for children who have below average abilities, which is in line with its primary purpose of identifying children who are in need of additional assessment (Farrington & Lonigan, 2015). That is, once children with below average abilities are identified using the GRTR-R, it is necessary to follow with a diagnostic assessment to identify any disabilities the children might have and to tailor the intervention to the needs of the children. An example of a recently developed and commonly used early numeracy tool is the Early Numeracy Assessment, which consists of 12 tasks that assess essential early numeracy skills including number identification and cardinality (Purpura & Lonigan, 2015). The reliability of the tasks ranged from an alpha of .69 to .90, and the tool was found to be psychometrically sound.

Although these tools provide valuable resources for researchers, none of these measures are normed for kindergarteners in Ontario, and specifically in Ottawa. In an attempt to fill this gap, a large school board in Ontario, Canada developed the Early Literacy and Numeracy Observation Tool (ELNOT; 2017). The ELNOT is a standardized progress-monitoring tool for teachers and educators to use.

Early Literacy and Numeracy Observation Tool (ELNOT)

History

In 2008-2009, the school board piloted a tool labeled the Early Literacy Observation Tool (ELOT) which included the original or adapted version of three of Clay's (1993) early literacy measures: *Letter Identification*, a shortened version of *Concepts about Print*, and an adapted version of *Hearing and Recording Sounds in Words*. Based on findings from this pilot, they made revisions to the observation protocol and implemented the revised ELOT in the 2009-2010 school year. The tool, however, did not have norms due to several problems (e.g., missing data, inaccurate reporting), nor did it include a numeracy component. In 2014-2015, the school board initiated a project to update the ELOT. To do so, two literacy measures were modified, and a numeracy component was added. The research officer in charge of the project added fundamental numeracy components after consulting with experts in the numeracy field. The goal of the Early Literacy and Numeracy Observation Tool (ELNOT) is to provide educators with a research-based assessment tool that measures early literacy and numeracy skills. In line with that, the final version of ELNOT was established in 2017. It has a total of 10 measures, four literacy and six numeracy components—these are described in the Methods section.

District-Level Norms

The goal of the ELNOT is to give teachers and educators a clear indication of whether a student requires further assessment, which involves discussions with other school staff like the learning support services regarding the observations of the teachers and educators about the student, and intervention. Hence, as a point of reference for evaluating children's performance, the school board developed district-level norms from the administration of the final version of the ELNOT during the 2017-2018 academic year (1,113 kindergarteners; about half in junior kindergarten). With such information available to school staff, they can gauge children's developmental growth throughout the two-year kindergarten program. As was the case with the ELOT, the ELNOT protocol indicates that teachers are to use their judgement to determine whether a student needs to complete any or all of the observation components and when to conduct the measures of the ELNOT during the school year.

Purpose of Study

The purpose of this thesis was to evaluate the psychometric properties of the ELNOT by examining its measurement precision through item analyses and measurement invariance through multiple-group CFA tests. The findings of this thesis provide valuable information on the accuracy and precision of the ELNOT and suggestions for potential revisions.

Research Questions

1. Do the tasks with item-level data provide a good coverage of the latent abilities? Specifically, what are the patterns of difficulty and discrimination for the items on the *Print Concepts*, *Letter Identification*, and *Number Recognition* tasks?

2. Is the ELNOT invariant across gender, kindergarten-level (junior and senior), and time (fall and spring)?
    a. Configural invariance: Do the groups have the same factor structure?
    b. Metric invariance: Do the groups have the same factor loadings?
    c. Scalar invariance: Do the groups have the same item intercepts?
    d. Residual variance invariance: Do the groups have the same item residual variances?

Methods

Participants

Sixty-five schools in the school board that include kindergarten children were randomly selected to participate in the project, from which 1,133 children in kindergarten were randomly selected to participate. Of these children, 569 were in junior kindergarten (50% boys) and 564 were in senior kindergarten (50% girls). For this thesis, only participants who met the following criteria (set by the school board) were included in the analyses: (a) the assessor was one of three primary research officers who used identical ways of assessment to ensure standardization in data collection, and (b) each child completed all components of the ELNOT at both data collection time points conducted in the fall and spring of the school year. As such, the participant sample for this study was 356 children from 32 schools including 173 in junior kindergarten (49% male) and 183 in senior kindergarten (46% male). Participation was voluntary; children could refuse to complete the tasks at any time during the assessments. The data were completely anonymized.

Measure

The ELNOT consists of four literacy and six numeracy measures described below.

Literacy

The four literacy measures, described below, are *Print Concepts*, *Letter Identification*, *Early Spelling*, and *Reading*.

Print Concepts. The *Print Concepts* task measures children's knowledge about the way we print language and whether they have developed important print awareness concepts. It does not require that children know how to read. This measure is an adapted version of Clay's (1993) *Concepts about Print* task which had 24 items. As per the guidelines of the original task, the assessor read one of two books to participating children and asked them questions before or after reading each page. The 22 questions assessed seven categories of concepts as listed in Table 1.

Children received one point for correctly responding to each question with two exceptions. Some items from the concepts of word and concepts of letter categories required children to respond to two questions for them to receive one point (e.g., word concepts – "Show me just one word" and "Now show me two words"). The maximum score was 22. In the present study, children were read books titled *Stones* in the fall and *Follow Me, Moon* in the spring testing session.

Table 1

Seven Categories of Concepts in the Print Concepts Task (Total Items = 22)

Concepts	Example	n items
Book	Title – “Show me the name of this book.”	3
Orientation	Line sequence – “What’s wrong with this?” (reading the bottom line first, then the top line)	3
Reading	Print carries the message as opposed to the picture – “Where do I begin to read?”	2
Directionality	Left to right – “Which way do I go?”	3
Word	Last word – “Use your finger to show me the last word on this page.”	3
Letter concepts	Capital letter – “Use your finger to show me a capital letter.”	4
Punctuation marks	Period – “What is this called/for?”	4

Letter Identification. The *Letter Identification* task (Clay, 1993) measures children’s knowledge of the names and sounds of the alphabet. The 26 letters of the alphabet, both in uppercase and lowercase formats, plus two alternative fonts of lowercase letters a and g (i.e., ɑ, a, ɡ, g) were presented, on a card, in the same random order. To introduce the task, the assessor showed the letter card to the child and asked what they were called. The assessor then pointed at each letter and asked the child to name it (i.e., “What is the name of this letter?”). If the child gave a response, the assessor moved onto the next letter. If the child was unable to respond or did not know the name of the letter, the assessor then asked the child the sound of the letter (i.e., “What sound does it make?”). If the child was unable to respond or did not know the sound of the letter, the assessor asked the child to say a word that starts with the letter (i.e., “Can you tell me a word that starts with this letter?”). Children received a single point for either correctly naming, sounding out, or saying a word that starts a letter. The maximum score was 54.

Early Spelling. The *Early Spelling* task measures children's ability to link sounds with letters. It is adapted from Clay's (1993) *Hearing and Recording Sounds in Words* task in the Observation Survey of Early Literacy Achievement. Children were asked to listen to two sentences and correctly spell each word in the sentences by listening carefully to the sounds. The sentences for the fall session were: I have a big dog at home. Today I am going to take him to school; and those for the spring session were: The boy is riding his bike. He can go very fast on it. The assessor read the sentences first at a neutral speed, then a second time slowly word by word, of which the children were made aware of beforehand, to allow them to listen carefully to each word and attempt to write the sentences accurately. Children received a score for every phoneme (sound) that they spelled correctly. For example, a child wrote "I H A B D A H T I M G T T H T S" in the fall, successfully capturing 16 sounds. The maximum score was 37.

Reading. The *Progress for Meaning Benchmark Reading Assessment Resource* (Nelson, 2003, 2010) is a commercially available basal series of 30 booklets each increasing in difficulty that allow teachers to measure children's level of independent reading level. Depending on the resources available at each of the schools, the assessor used either Kit-1 (Nelson, 2003) or Kit-2 (Nelson, 2010). Each booklet comes with a Reading Record that allowed the recording of reading errors, and a child has to read accurately 95% of the words in the booklet in order to be assessed on comprehension. Each booklet comes with an Assessment Record that includes questions to assess the child's comprehension, with questions increasing in difficulty across the 30 booklets. To pass a reading level (i.e., one point per booklet), the child must correctly answer questions that reflect general to thorough comprehension of the story. The maximum score was 30.

Numeracy

The six numeracy tasks, described below, are *Verbal Counting*, *Number Recognition*, *Number Naming*, *Give-N*, *Compare Sets*, and *Magnitude Comparison*.

Verbal Counting. The *Verbal Counting* task, adapted from Purpura and Lonigan (2015), measures children's verbal counting abilities. The assessor instructed children to start counting from one and go as high as they can. Children received one point for correctly counting to five, and an additional point for correctly counting to each of the following numbers: 10, 15, 20, 25, 40, and 100. The highest count was determined to be the largest number counted before an error. For example, if the child omitted thirteen (e.g., "…11, 12, 14), his/her highest count was recorded as twelve. Or if the child repeated a cycle (e.g., "…28, 29, 30, 21, 22"), his/her highest count was recorded as thirty. The maximum score was seven.

Number Recognition. The *Number Recognition* task, adapted from Olson and colleagues (2009), measures children's ability to map number names to digits. Children were presented with 14 sets of numbers each presented in a row, with four numbers in each row presented in a random order (e.g., 2, 10, 6, 4) that ranged from 0 to 90. There were 56 items in total. The assessor named in a random order each number in the set (e.g., 10, 4, 2, 6; "Show me number 10.") and the child had to point to the number named (e.g., 10, 4, 2, 6). Before beginning the task, children received two practice items, and the assessor corrected them if they were unable to identify the numbers correctly. The random order of administration was consistent across children and test time. Children received one point for each number identified correctly, for a maximum score of 56.

Number Naming. The *Number Naming* task (Lembke & Foegen, 2005) measures children's ability to map digits to number names. This task included two lists of 56 numbers between 1 and 100. Each list was administered once, for a total of 112 items. Numbers in each list were randomly ordered. The assessor pointed to the numbers in order, and children were asked to say the names of as many numbers as they could in one minute (per list). The same lists were used in both assessment periods (i.e., Fall and Spring). Before starting the task, the assessor introduced the task (i.e., "There are numbers in boxes.") and gave four practice items (i.e., 6, 15, 1, and 44 – pointing at every box and asking, "What number is this?"). If children gave an incorrect response, the assessor corrected them. Children received one point for each number that they identified correctly, and their scores on both lists were averaged to assign them a total score for this task. If the child did not respond to an item within 3 seconds, that item was scored as incorrect, and the assessor pointed to the next number. If the child skipped an item or an entire row, each item that they skipped was scored as incorrect. The maximum score was 56.

Give-N. The *Give-N* task, adapted from Wynn (1990), measures children's understanding of the relationship between number words and the quantities that they represent (i.e., the cardinality principle). The assessor asked children to give a certain number of blocks to the finger puppet, starting at three blocks and increasing the number of blocks by one if successful in the previous trial. The task ended when the child gave two correct responses for a number and two incorrect responses for the next number. For example, if unsuccessful at any level, the child moved down by one number (e.g., if unsuccessful at presenting four blocks, asked to present three blocks again); if successful at this stage, the child revisited the level they had failed previously. So, if the child responded incorrectly (i.e., did not present four blocks the second time), they would be identified as a "3-knower" and receive a score of three. However, if the child was successful at presenting four blocks, they continued increasing by one block until they reached a maximum number of eight blocks. If the child accurately reached eight blocks (i.e.., maximum), the assessor asked for seven (and eight) blocks again to ensure that both numbers were revisited. The highest number of blocks that children correctly "gave" the finger puppet was the score that they received on this task. The maximum score was eight.

Compare Sets. This second measure of the Cardinality Principle, adapted from Sarnecka and Wright (2013), requires more sophisticated knowledge on the part of the child than *Give-N* and assesses children's ability to make one-to-one correspondence of two sets with the same number of objects. The task included four items: peaches (5 to 5), cupcakes (5 to 6), apples (6 to 5) and cookies (6 to 6). Children were presented with pictures of these snacks, each set of snacks lined up in same-length rows, with either an equal number of snacks (peaches and cookies) or not (cupcakes and apples). When the two sets of snacks had different amounts, like cupcakes, it was visually apparent that one of the sets had one fewer cupcake. There was also a practice item in the beginning (oranges; 5 to 5) where the assessor corrected any errors in comprehension to clarify the task. The assessor asked the child whether the two sets of snacks that they gave to the two finger puppets were the same or different and asked if a particular set was "five or six." Children had to correctly identify the number of snacks in a set to receive a score. Otherwise, if a child correctly identifies that two sets have the same number of snacks but fails to express the correct amount, the child does not receive any points for that item. Each child was allowed to pick which snack to give next, and therefore the order of the trials was randomized. The maximum score on this task was four.

Magnitude Comparison. The *Magnitude Comparison* task (Lembke & Foegen, 2005) measures children's understanding of the relationship between quantities and digits and their ability to make magnitude comparisons. Randomly selected numbers (ranging from 0 to 20) were paired together, and children were asked to identify the "bigger" number in each set. This task included two lists with a total of 63 different sets (or items) in each, and children answered as many items as they could in one minute (per list). The same lists were used in both assessment periods (i.e., Fall and Spring). Before beginning the task, children were given three practice items, where the assessor repeated the correct response (e.g., "Good. 7 is bigger than 1.") or corrected the child every time if they were unable to identify the bigger number in each of the three sets (e.g., "The number that is bigger is 7."). Children received a score for each item that they identified correctly, and their scores on both lists were averaged to assign them a total score for this task. As was the case with the *Number Naming* task, if the child did not respond to an item within 3 seconds, that item was scored as incorrect, and the assessor moved to the next item. If the child skipped an item or an entire row, each item that they skipped was scored as incorrect. The maximum score was 63.

Procedure

The ELNOT was administered twice, in the fall and spring of the kindergarten year, by three research officers of the school board. The Fall administration was conducted between November and December of 2017, whereas the Spring administration was conducted between May and June of 2018. At each assessment time, the tasks were administered in two sessions, lasting approximately 15 minutes each, because young children's attention span is typically short and that they may lose interest in participation if kept for longer. Moreover, the sessions were scheduled within a maximum of one week from each other because kindergarten children's knowledge can grow quickly. The first session consisted of the following five tasks, in order: *Letter Identification*, *Verbal Counting*, *Number Recognition*, *Print Concepts*, and *Number Naming*. The second session consisted of the following five tasks, in order: *Give-N*, *Early*

Spelling, *Magnitude Comparison*, *Reading*, and *Compare Sets*. To avoid any errors related to scoring, each assessor scored the tasks that they have administered right after each session.

Analysis Overview

The purpose of this thesis was to evaluate the psychometric properties of the ELNOT. Preliminary analyses were performed to assess the suitability of each task for further analyses. Excluded from further analyses were tasks that had more than 15-20% of children scoring at floor or ceiling (McHorney & Tarlov, 1995), or items within tasks that had 100% of children scoring the same (i.e., zero variance; Ellis, 2017). Two series of analyses were conducted on the tasks retained. First, item analyses were conducted on the three tasks for which item-level data were available, namely, *Print Concepts*, *Letter Identification*, and *Number Recognition*. Second, the measurement invariance of the ELNOT was assessed with a series of multiple-group confirmatory factor analyses (CFA). Analyses were conducted separately for the fall and spring data.

Item Analyses

Item Response Theory (IRT) analysis is a model-based method of latent trait measurement that assumes that the probability of every possible response that participants can give to any of the items in a task is closely related to the level of the underlying characteristic that the task assesses (i.e., the latent ability [theta or θ]; Kline, 2005). One of the most frequently used binary-response models (e.g., correct or incorrect) in IRT is a two-parameter logistic model (2-PL; Kline, 2005). The two item-level characteristics, or parameters, estimated in a 2-PL model are the item difficulty and discrimination parameters.

The difficulty parameter (b) of an item represents the latent ability level required for 50% of the participants to correctly respond to that item ($b = \theta$ at 50% probability of a correct response). As such, the difficulty parameter (*b*) provides an estimate of how easy or difficult an item is for the participants. The higher the *b* value, the more difficult is the item. If the difficulty parameters are widely spread from the lowest to the highest levels of the trait, which usually ranges between -3 to +3, the scale measures the entire range of the trait and makes it generalizable to a variety of populations (Kline, 2005).

The discrimination parameter (*a*) gives an estimate of how well an item differentiates between participants of varying latent-ability levels. The larger the *a* value, the more sharply the response categories are defined for participants with different levels of the trait, therefore indicating that an item has more precision. As such, discrimination values from 0.01 to 0.24 are considered very low, 0.25 – 0.63 low, 0.65 – 1.34 moderate, 1.35 – 1.69 high, and greater than 1.7 are considered very high in precision (Baker, 2001).

From the examination of the item parameters, one can determine which items are and are not contributing to the precision of the tasks, that is, items that are too easy or difficult and that are low in discriminating among children with different levels of the latent ability weaken the value of a test or task. The discrimination and difficulty parameters of items work together to affect the probability of children's correct responses. A 2-PL model provides a more realistic picture of achievement-based items and would be suitable for analyzing the tasks used in this thesis.

In addition to the previous analyses, the examination of the test information curves allows one to determine at which level of latent ability the tasks provide the most information and whether each task covers the entire range of latent ability, and as such it provides information on how well a task measures the latent ability along the latent trait continuum. It is ideal to have the curve peak at the average level of the latent trait and spread out on either side, indicating that the test information function is spread across a wide range of thetas. The standard error is the lowest at the theta level for which the most information is provided; the peak point of the test information curve and the trough of the standard error curve occur at the trait level where the information is highest (Kline, 2005).

Unidimensionality

The first step in item analyses is to assess the assumption of unidimensionality, that is, that a task measures only one domain; and local independence, namely, that responses to each item on a task are conditional on the latent ability level assessed and are not conditional on responses to other items (Embretson & Reise, 2000). When a task meets the assumption of unidimensionality, it is assumed to meet the assumption of local independence because both are isomorphic with one another (Lord, 1968). Establishing unidimensionality, via factor analysis, is critical to ascertain that items belong to the same construct (Tay et al., 2015).

To assess the factor structure of each of the three tasks, separate exploratory factor analyses (EFA) were conducted in M*plus* version 7 (Muthén & Muthén, 2012) using data from junior and senior kindergarten together because sample sizes of at least 300 are considered adequate for factor analysis (Comrey & Lee, 1992). To determine whether tasks were unidimensional, I used the “ratio of first to second eigenvalue greater than three” rule (Slocum & Zumbo, 2011). All items that had significant factor loadings at alpha level of .05 were retained in order to remain as close as possible to the initial task. Model fit was assessed with Hu and Bentler’s (1999) recommended indices for a good model: root mean square error of approximation (RMSEA) smaller than .06, comparative fit index (CFI) larger than .95, Tucker-Lewis index (TLI) larger than .95, and standardized root mean square residual (SRMR) smaller than .08.

Differential Item Functioning (DIF) Tests

Assessing the measurement equivalence of items in a task is necessary to ensure that scores can be compared between subgroups, kindergarten level in our case (Vandenberg & Lance, 2000). Analysis of measurement equivalence in the IRT framework is referred to as differential item functioning (DIF) analysis. Using the R mirt package version 1.32 (Chalmers, 2012), IRT likelihood ratio (IRT-LR) tests were conducted on each task to determine whether item-level bias existed between gender or kindergarten level (Thissen et al., 1993). To control for Type-I error in each DIF analyses, I applied the Benjamini and Hochberg (B-H) procedure to identify significant DIF (Benjamini & Hochberg, 1995). Further description of the DIF tests are provided in Appendix A.

Reliability

The classical test theory reliabilities (Cronbach's α) for each of the three tasks were calculated at each time point (if applicable). Reliability coefficients above .70 are considered acceptable, those greater than .80 are considered fair, and those higher than .90 are considered ideal (Nunnally & Bernstein, 1994).

Measurement Invariance Analyses

Given that the ELNOT assesses building blocks predictive of later academic skills and given that previous studies have shown strong correlations between early literacy and numeracy skills (Duncan et al., 2007; Hecht et al., 2001; Matthews et al., 2009; Piasta et al., 2010; Purpura et al., 2011; Singer & Strasser, 2017; Welsh et al., 2010), the model tested was one of a single underlying construct of early academic skills. The psychometric evaluation of the ELNOT required the verification that the structure of the underlying construct was the same across subgroups (e.g., gender) and across time. Age and gender are among the variables that are typically involved in measurement invariance analyses (Tay et al., 2015).

Multiple-group confirmatory factor analyses (CFA), a multistage test within the structural equation modeling framework, were conducted in M*Plus* version 7 (Muthén & Muthén, 2012) to assess the measurement invariance of the ELNOT across groups and across time (Byrne, 2004). Four types of measurement invariance were assessed sequentially, each requiring invariance at the preceding step to move forward. First, configural invariance assessed whether the factors and pattern of factor loadings were similar for all groups (Schmitt & Kuljanin, 2008). Once an invariant model with good data fit is established in this step, this baseline model is then used to compare the restricted models in the next three steps. To do so, the chi-square difference test between the two models is conducted and if it is not statistically significant, then this would suggest that invariance at the subsequent step is in place (Putnick & Bornstein, 2016). The three subsequent models are: (1) metric invariance testing the equivalence of factor loadings across groups by constraining factor loadings to be equal across groups; (2) scalar invariance testing whether the intercepts are similar for all groups; and (3) residual invariance testing, across groups, the equivalence of residuals that are error variances and covariances. At any point along this four-step procedure, the model comparison might not support invariance. The degree of partial invariance in these situations can be determined by modifying the noninvariant model (see Appendix B). Because each step requires that the preceding model be fully or partially supported, any constraints applied in the preceding model are retained.

Results

Preliminary Analyses

Floor or Ceiling Effects

The literacy tasks with item-level data (*Print Concepts* and *Letter Identification*) did not have floor or ceiling effects in the fall or spring testing. This was not the case for the numeracy tasks. On the *Compare Sets* task, 45% and 65% of children, averaged across kindergarten level, were at ceiling in the fall and spring, respectively. On the *Number Recognition,* there was no ceiling or floor effect in the fall, but 37% of children in senior kindergarten were at ceiling in the spring testing. Therefore, *Compare Sets* was eliminated from further analyses, while *Number Recognition* was retained for the fall analyses only.

Of the six tasks without item-level data, one literacy task was not analyzed and four numeracy tasks were not analyzed further due to floor effects on *Reading* and ceiling effects on *Verbal Counting*, *Give-N*, *Compare Sets*, and *Number Recognition (Spring)*. Floor effects were observed for the reading task, with 100% of junior and 88% of senior kindergarten children scoring zero on the *Reading* task in the fall, and 94% of junior and 74% of senior kindergarten children scoring zero in the spring. In contrast, ceiling effects were observed on numeracy tasks. More than half of the children (65%) achieved the highest possible score on the *Give-N* task in the fall, whereas both junior (50%) and senior (57%) kindergarten children were at ceiling when analyzed separately. In the spring, the ceiling effect's severity on the *Give-N* task increased by about 20%, and *Verbal Counting* had a ceiling effect of 20%. However, it should be noted that the ceiling effect on the *Verbal Counting* task was due to the performance of senior kindergarteners alone as no ceiling effects were observed on this task in junior kindergarten.

Specifically, 33% of children in senior kindergarten achieved the maximum score on the *Verbal Counting* task.

Zero Variance Items

The item-level examination of the three tasks retained with item-level data at both assessment times revealed that the *Print Concepts* and *Number Recognition* tasks had a few items with zero variability, whereas *Letter Identification* did not.

Print Concepts. Two items in the fall and one item in the spring had zero variance. Specifically, no child in both junior and senior kindergarten correctly responded to Item 17–"Quotation Mark" in the fall and spring, and no child in junior kindergarten correctly responded to Item 18–"Comma" at both assessment times. These two items were eliminated because they did not contribute to the tasks and were not comparable across groups. Therefore, the new possible maximum score on the *Print Concepts* task was 20.

Number Recognition. All children in senior kindergarten correctly responded to Item 13–"Number 4 in row four." This item was eliminated because it did not make any contributions to the task and was not comparable across groups. Therefore, for the *Number Recognition* task, the new possible maximum score was 55.

Item Analyses

Print Concepts

Factor Structure. The factor loadings for each item at both assessment times are presented in Table 2.

Fall. The first eigenvalue was 8.97, which was substantially larger than the second value, 1.96. The ratio of the two was 4.56—thus, a unidimensional model was acceptable. All items of the *Print Concepts* task significantly loaded onto a single factor at the alpha level of .05. The majority of the items had loadings greater than 0.35, ranging from moderately strong to strong loadings; however, Item 11–"Inversion of Picture" had a particularly weak loading of 0.21. It may be argued that this item relates more distally to the construct; nonetheless, because it managed to surpass the cut-off criteria (significant loading), I did not omit this item from subsequent analyses. The fit measure indices indicated an adequate model fit: RMSEA = .078, CFI = .943, TLI = .936, and SRMR = .145. Therefore, the *Print Concepts* task was judged to be unidimensional.

Spring. The first eigenvalue was 9.17, which was substantially larger than the second value, 1.86. The ratio of the two was 4.93—thus, a unidimensional model was acceptable. All items of the *Print Concepts* task significantly loaded onto a single factor at the alpha level of .05. The majority of the items had loadings greater than 0.35, ranging from moderately strong to strong loadings. Interestingly, the factor loading of Item 11–"Inversion of Picture" that was particularly weak in the fall increased to 0.30. However, this time, Item 3–"Book Title" had a particularly weak loading of 0.22. It may be argued that this item relates more distally to the construct; nonetheless, I kept this item as well. The fit measure indices indicated an adequate model fit: RMSEA = .066, CFI = .947, TLI = .940, and SRMR = .140. Therefore, the *Print Concepts* task was unidimensional at both assessment times, with mostly adequate model fit indices.

Table 2
Exploratory Factor Analyses of the Print Concepts Task as a Function of Testing Time (N = 356)

Item	Concept	Description	Fall Factor Loading 1	Spring Factor Loading 1
3	Book	Title	0.28	0.22
1	Book	Front cover	0.87	0.96
2	Book	Back cover	0.94	0.90
11	Orientation	Inversion of picture	0.21	0.30
12	Orientation	Response to inverted print	0.78	0.78
13	Orientation	Line sequence	0.61	0.43
6	Directionality	Left to right	0.94	0.88
5	Directionality	Beginning of text	0.89	0.89
7	Directionality	Return sweep	0.95	0.88
4	Reading	Print carries the message	0.85	0.88
8	Reading	One-to-one match	0.62	0.59
9	Word	First word	0.70	0.72
21	Word	One word and two words	0.77	0.72
10	Word	Last word	0.66	0.76
15	Letter	Small letter	0.38	0.54
20	Letter	One letter and two letters	0.47	0.57
14	Letter	Capital letter	0.29	0.50
22	Letter	First letter and last letter	0.68	0.74
16	Punctuation	Question mark	0.44	0.45
19	Punctuation	Period	0.67	0.63

Note. Items 17 and 18 were eliminated due to zero variance.

DIF by Gender and K-Level. The anchor set of the *Print concepts* task contained five items (25% of the total number of items; see Table A1 in Appendix A). After applying B-H correction for multiple comparisons, none of the items showed significant DIF by gender in the fall or spring. Therefore, girls and boys were merged in subsequent analyses. However, after applying B-H correction for multiple comparisons, three items in the fall and four items in the spring presented significant DIF by kindergarten level.

Items with significant DIF in the fall were Item 2–"Back cover," Item 10–"Last word," and Item 21–"First letter and last letter." The follow-up DIF test showed that Items 10 and 21 had uniform DIF, whereas Item 2 had nonuniform DIF. While senior kindergarteners had a higher probability of correctly responding to items with uniform DIF across all levels of the latent trait, the direction of the DIF changed at a certain point on Item 2. Junior kindergarteners had a higher probability of giving a correct response until that point, after which the item seemed to favour senior kindergarteners.

Items with significant DIF in the spring were Item 3–"Title," Item 8–"One-to-one match," Item 10–"Last word," and Item 12–"Response to inverted print." The follow-up DIF test showed that Items 3, 8, and 12 had uniform DIF, whereas Item 10 had nonuniform DIF. The senior kindergarteners were at a constant advantage again on the items with uniform DIF. All items with significant DIF were eliminated from the subsequent analyses, resulting in a new total number of 14 items on the *Print Concepts* task.

Item Parameter Estimates. The parameter estimates for each item at both assessment times are presented in Table 3 for junior and senior kindergarten.

Fall.

Junior Kindergarten. The discrimination parameters of the items for junior kindergarten varied greatly, from low to very high. The three most discriminating items were related to directionality concepts. Among these, Item 7–"Return sweep" had a particularly very high discrimination value of 7.96. Overall, the discrimination values ranged from 0.32 (Item 15–"One letter and two letters") to 7.96 (Item 7–"Return sweep"), with a very high average of 1.90.

The difficulty parameters of the items had a wide range, with values between -3.78 (Item 15–"Small letter") and 2.73 (Item 13–"Line sequence"), which covers the entire level of the latent trait. The difficulty values for 5 of the 14 items were below zero, meaning that these items required below average levels of the latent ability for a correct response to be obtained beyond chance. Of the five items with negative difficulty values, three items had difficulty values less than -2, suggesting that children whose level of knowledge about print concepts was higher two standard deviations below the mean had a greater than 50% probability of correctly responding to these items. The remaining nine items had difficulty values above zero. The difficulty parameters of four of these items did not exceed one, meaning that children whose latent ability level was between zero and one standard deviation above the mean had a greater than 50% probability of correctly responding to nine items on this task. On the other hand, three items' difficulty parameters exceeded 1, and two items exceeded 2. Thus, only children whose latent ability level is about three standard deviations above the mean had a greater than 50% chance of correctly responding to all questions about print concepts.

The test information curve (see Figure 1) showed that when all items are considered, the *Print Concepts* task provided the most information for junior kindergarteners with ability levels of about 0.40. This is also where the lowest standard error was observed, indicating good discrimination or high precision. Contrary to what the range of the difficulty parameter values suggested, the task seems to cover a narrow range of the latent ability with more information provided for higher levels of the latent trait. This difference in interpretation could be explained by the jump in magnitude among difficulty parameter values. Although the lower end of the range is at -3.78, the mean of the difficulty values was -0.02, and the median was 0.32. The mean alone suggests that, overall, the *Print Concepts* task was neither too easy nor too hard for junior kindergarteners. However, the median implies that the task is leaned towards being a little challenging for junior kindergarteners. Considering that there are a few difficulty values on the lower end of the range (three values smaller than -2.00), the median presents a more realistic picture of the overall difficulty of this task.

Senior Kindergarten. The discrimination parameters of the items for senior kindergarten varied considerably, from low to very high. Like junior kindergarten, the two items with the highest discrimination values were related to directionality concepts. Overall, the discrimination values ranged from 0.37 (Item 14–"Capital letter") to 5.48 (Item 6–"Left to right"), with an average of 2.00 that is in the very high discrimination range.

The difficulty parameters of the items for senior kindergarten had a relatively narrow range, with values between -2.79 (Item 15–"Small letter") and 1.88 (Item 13–"Line sequence"), about one standard deviation difference from junior kindergarten discrimination range at both ends. The difficulty values for 12 of the 14 items were below zero, meaning that these items required below average levels of the latent ability for a correct response to be obtained beyond chance. In other words, the *Print Concepts* task was much easier for senior kindergarteners than for junior kindergarteners. Despite the differences in item parameter estimates, similar trends were observed between the two groups. For example, the three least difficult items for junior kindergarteners were also the easiest for senior kindergarteners. Also, the two items that had difficulty values above zero (Items 19 and 13) were the most difficult for junior kindergarteners. One item's difficulty parameter did not exceed 1, meaning that children whose latent ability level was between zero and one standard deviation above the mean had a greater than 50% probability of correctly responding to 13 items on this task. On the other hand, the difficulty parameter of one item exceeded 1. Thus, only children whose latent ability level is about two standard deviations above the mean had a greater than 50% chance of correctly responding to all questions about print concepts on this task. The average of the difficulty values was -0.82, and the median was -0.78, which suggests that, overall, the *Print Concepts* task was easy for senior kindergarteners.

The test information curve (see Figure 2) of the *Print Concepts* task for senior kindergarten was bimodal: it first peaked at around θ = -0.60 and then again at around θ = 1.90. However, the task provided the most information at the first peak, which is also where the lowest standard error was observed, indicating good discrimination or high precision. Although the difficulty parameter range of the items for senior kindergarteners was narrower than that of junior kindergarteners, the task provided information on a wider range of the latent ability for senior kindergarteners. However, very little discrimination occurred at higher levels of latent ability. Overall, the test information curve confirmed that this task was easy for senior kindergarteners.

Spring.

Junior Kindergarten. The spring dataset showed a trend similar to the fall dataset. The discrimination parameters of the items for junior kindergarten in the spring varied greatly, from low to very high. Two items related to directionality concept continued to be among the most discriminating items; however, despite being an easy item, Item 4–"Print carries the message" was the second most discriminating item among 14 items. In other words, children who responded incorrectly to this item need to receive additional attention as only low performers or children with low levels of the latent ability do not yet understand that it is the print not the picture that carries the message. Overall, the discrimination values ranged from 0.30 (Item 11–"Inversion of picture") to 3.82 (Item 7–"Return sweep"), with a high average of 1.67.

The difficulty parameters of the items had a similarly wide range, with values between -3.75 (Item 20–"One letter and two letters") and 2.88 (Item 13–"Line sequence"), which covers the entire level of the latent trait. Item 13 continued to be the most difficult item for junior kindergarteners in the spring. Of note, the number of items with difficulty values below zero increased from 5 to 10 out of 14 items. The average of the difficulty values dropped to -0.43 with the median at -0.24, which suggests that, overall, the *Print Concepts* task was easier for junior kindergarteners in the spring than it was in the fall.

The test information curve for the spring dataset (see Figure 3) showed that when all items are considered, the *Print Concepts* task provided the most information for junior kindergarteners with ability levels of about -0.35. This is also where the lowest standard error was observed, indicating good discrimination or high precision. Although a little bit narrower than the range of the difficulty parameters suggested, the distribution supported the task's coverage area of the latent trait. Also, even though the task became easier for junior kindergarteners in the spring, *Print Concepts* covered a wider range of the latent ability in the spring for junior kindergarten compared to the fall. Overall, the test information curve confirmed that while this task was easy for junior kindergarteners in the spring, it could still discriminate children who have not yet attained an average latent ability level.

Senior Kindergarten. The discrimination parameters of the items for senior kindergarten in the spring varied greatly as well, from low to very high. The two items related to the directionality concept that were among the three most discriminating items for junior kindergarteners were also the most discriminating for senior kindergarteners. As was the case in junior kindergarten, these items were easy. Children who incorrectly responded to these items do not yet know where the text begins or that the direction of reading goes from the top line to the bottom line. Considering that these children are at the end of the academic year of senior kindergarten, incorrect responses on these items might indicate a red flag. Overall, the discrimination values ranged from 0.60 (Item 13–"Line sequence") to 4.04 (Item 5–"Beginning of text"), with an average of 1.86 that is in the very high discrimination range.

The four most difficult items for senior kindergarteners were also the hardest for junior kindergarteners. However, the difficulty parameter values of these items were very different across the two groups. Although all four items had difficulty values above zero for junior kindergarteners, only one item (Item 13–"Line sequence") had a positive difficulty value for senior kindergarteners. Interestingly, contrary to the overall trend, this item seems to be more difficult for senior kindergarteners than it was for junior kindergarteners. Also, the most difficult item was also the least discriminating among the 14 items. Overall, the difficulty parameters of the items for senior kindergarten had a relatively narrow range, with all values between -2.78 (Item 1–"Front cover") and -0.36 (Item 19–"Period") except for Item 13–"Line sequence" with a difficulty value of 5.22. Although the average of the difficulty values was the same as in the fall (-0.82), the median was -1.06, suggesting that the *Print Concepts* task was much easier for senior kindergarteners in the spring.

The test information curve for the spring dataset (see Figure 4) showed that when all items are considered, the *Print Concepts* task provided the most information for senior kindergarteners with ability levels of about -1.10. This is also where the lowest standard error was observed, indicating good discrimination or high precision. The distribution supported that the task covered only the lower levels of the latent ability. In fact, the information provided for latent ability levels above average was close to none. Overall, the *Print Concepts* task was too easy for senior kindergarteners in the spring.

Table 3

Discrimination (a) and Difficulty (b) Estimates for Print Concepts for Junior (n = 173) and Senior (n = 183) Kindergarten in Fall and Spring

			Fall				Spring			
			Junior Kindergarten		Senior Kindergarten		Junior Kindergarten		Senior Kindergarten	
Item	Concept	Description	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	Book	Front cover	0.36	-3.46	1.49	-2.04	1.44	-1.78	1.79	-2.78
11	Orientation	Inversion of picture	0.38	-0.73	0.37	-1.14	0.30	-1.58	0.80	-0.78
13	Orientation	Line sequence	2.22	2.73	3.95	1.88	1.59	2.88	0.60	5.22
6	Directionality	Left to right	3.88	0.29	5.48	-0.69	3.19	-0.42	3.02	-1.29
5	Directionality	Beginning of text	4.02	0.34	2.66	-0.59	2.97	-0.24	4.04	-1.03
7	Directionality	Return sweep	7.96	0.45	5.30	-0.45	3.82	-0.15	3.12	-1.08
4	Reading	Print carries the message	2.46	-0.01	2.73	-0.99	3.59	-0.70	2.68	-1.63
9	Word	First word	1.39	1.13	1.23	-0.27	1.58	0.33	1.29	-0.68
15	Letter	Small letter	0.32	-3.78	0.82	-2.79	1.31	-2.15	1.38	-2.47
20	Letter	One letter and two letters	0.45	-2.21	1.15	-1.88	0.47	-3.75	2.18	-1.90
14	Letter	Capital letter	0.35	0.03	0.37	-1.72	0.77	0.43	0.89	-0.51
22	Letter	First letter and last letter	1.09	1.08	1.01	-0.44	1.17	-0.24	2.27	-1.02
16	Punctuation	Question mark	0.39	1.68	0.59	-0.88	0.38	-0.09	0.95	-1.22
19	Punctuation	Period	1.30	2.15	0.84	0.53	0.83	1.40	1.10	-0.36

Note. Items 17 and 18 were removed due to zero variance, and items 2, 3, 8, 10, 12, and 21 were removed due to DIF. The final total number of items was 14.

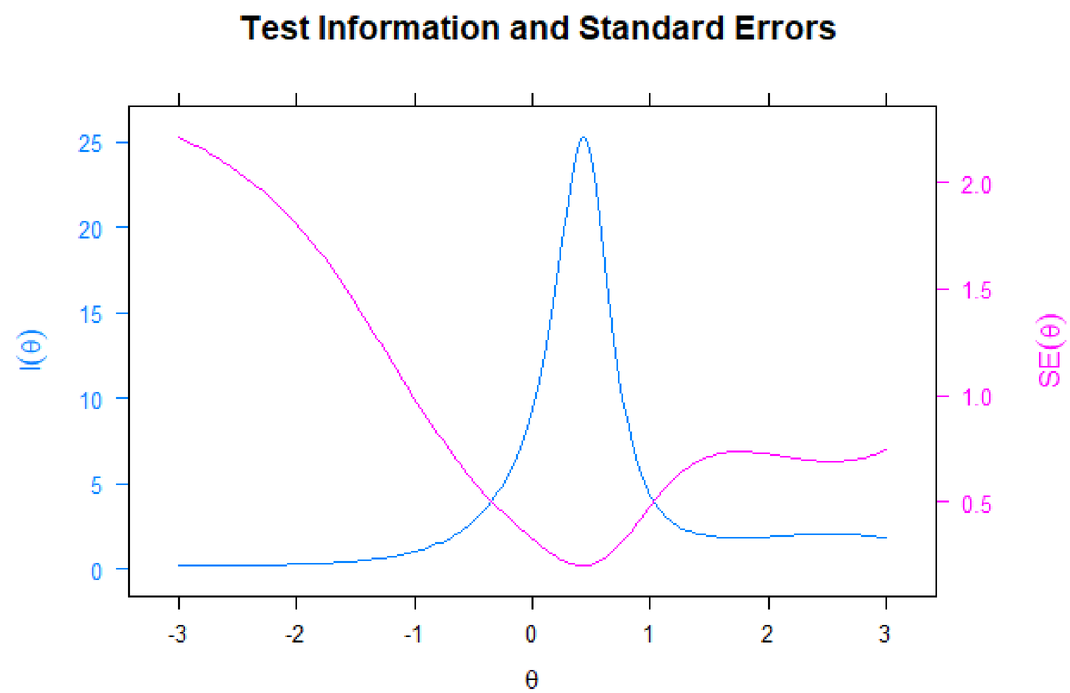

Figure 1
Test Information Curve of the Print Concepts Task for Junior Kindergarten (N = 173) in the Fall

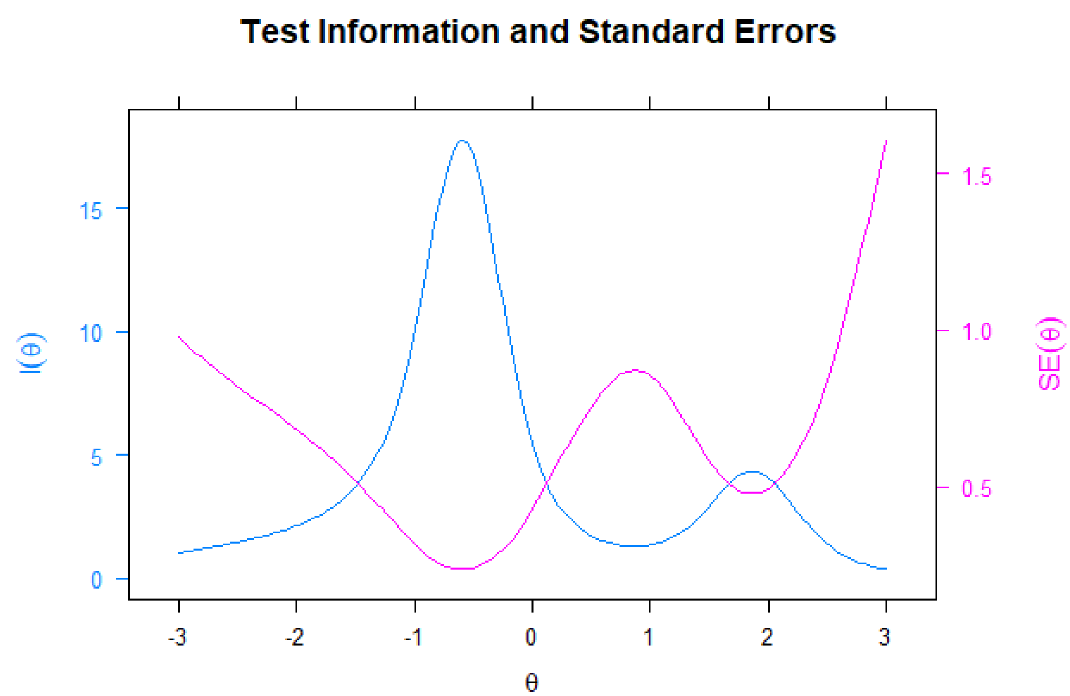

Figure 2
Test Information Curve of the Print Concepts Task for Senior Kindergarten (N = 183) in the Fall

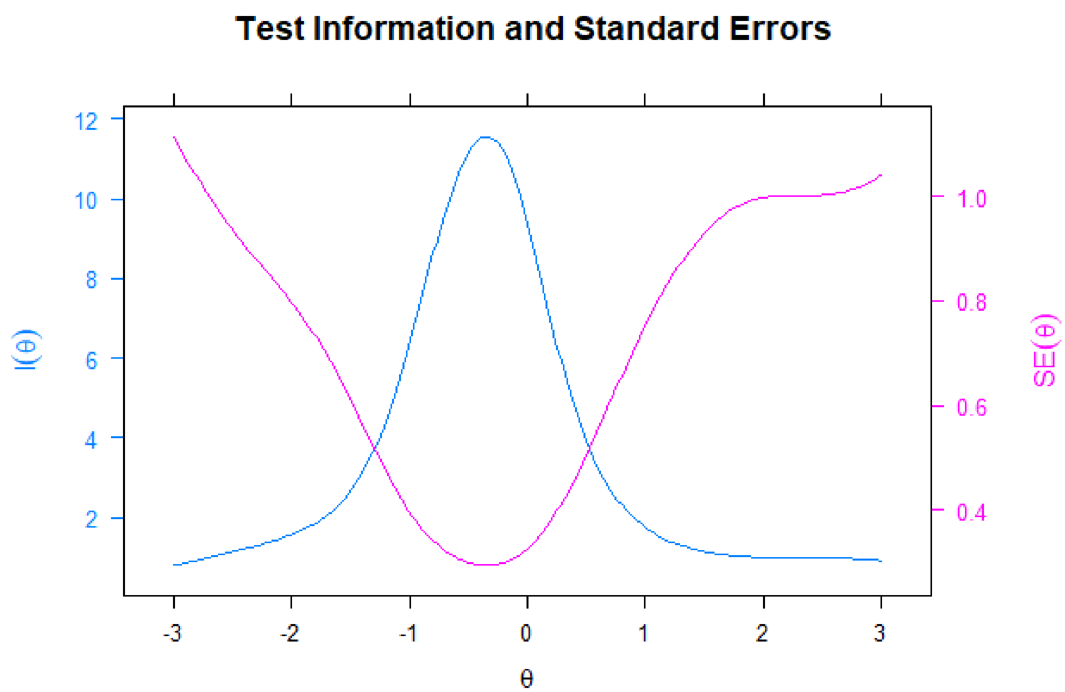

Figure 3
Test Information Curve of the Print Concepts Task for Junior Kindergarten (N = 173) in the Spring

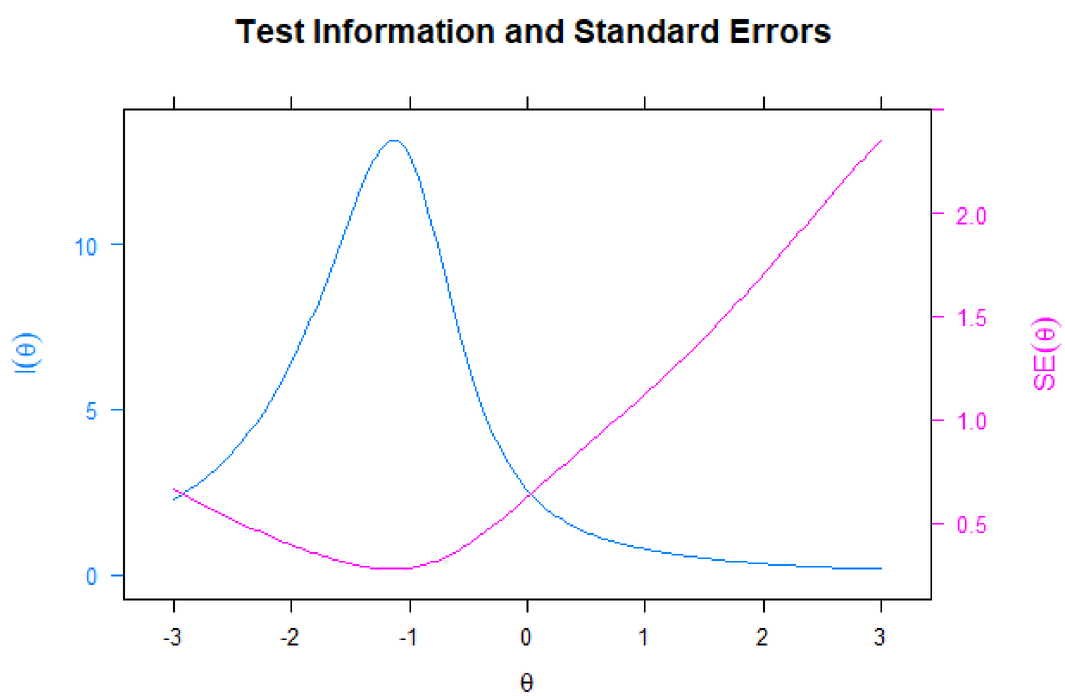

Figure 4
Test Information Curve of the Print Concepts Task for Senior Kindergarten (N = 183) in the Spring

Reliability. After eliminating items with zero variance and those with a significant DIF, the *Print Concepts* task had acceptable reliabilities of α = .79 and .80 in the fall and spring, respectively.

Letter Identification

Factor Structure. At both assessment times, the *Letter Identification* task was unidimensional, with mostly good model fit indices. The factor loadings for each letter at both assessment times are presented in Table 4.

Fall. The first eigenvalue was 40.34, which was significantly larger than the second value, 2.12. The ratio of the two was 19.04—thus, a unidimensional model was acceptable. All items of the *Letter Identification* task significantly loaded onto a single factor at the alpha level of .05. In fact, all items had strong loadings, with the smallest loading at 0.64. It would be surprising to have found otherwise, because the underlying construct of this task is alphabet knowledge, and each item on the task is a letter from the alphabet directly linked to the construct. The fit measure indices indicated a good model fit: RMSEA = .049, CFI = .984, TLI = .984, and SRMR = .082. Therefore, the *Letter Identification* task was unidimensional.

Spring. The first eigenvalue was 36.50, which was significantly larger than the second value, 2.81. The ratio of the two was 12.98—thus, a unidimensional model was acceptable. All items of the *Letter Identification* task significantly loaded onto a single factor at the alpha level of .05. In fact, all items had strong loadings, with the smallest loading at 0.52. The fit measure indices indicated a good model fit: RMSEA = .042, CFI = .974, TLI = .973, and SRMR = .094. Therefore, the *Letter Identification* task was unidimensional.

Table 4
Exploratory Factor Analyses for the Letter Identification Task as a Function of Testing Time (N = 356)

	Fall				Spring		
	Factor Loading		Factor Loading		Factor Loading		Factor Loading
Letter	1	Letter	1	Letter	1	Letter	1
A	0.86	ɑ	0.89	A	0.83	ɑ	0.91
		a	0.89			a	0.87
B	0.85	b	0.72	B	0.84	b	0.63
C	0.93	c	0.92	C	0.84	c	0.81
D	0.94	d	0.78	D	0.87	d	0.70
E	0.88	e	0.90	E	0.87	e	0.87
F	0.92	f	0.97	F	0.94	f	0.91
G	0.92	ɡ	0.86	G	0.88	ɡ	0.87
		g	0.79			g	0.72
H	0.88	h	0.87	H	0.87	h	0.86
I	0.89	i	0.84	I	0.80	i	0.78
J	0.85	j	0.84	J	0.78	j	0.79
K	0.89	k	0.91	K	0.81	k	0.86
L	0.87	l	0.71	L	0.81	l	0.66
M	0.89	m	0.87	M	0.84	m	0.85
N	0.93	n	0.88	N	0.90	n	0.90
O	0.90	o	0.92	O	0.85	o	0.86
P	0.89	p	0.89	P	0.94	p	0.88
Q	0.92	q	0.63	Q	0.88	q	0.52
R	0.90	r	0.90	R	0.90	r	0.88
S	0.91	s	0.87	S	0.82	s	0.73
T	0.91	t	0.81	T	0.88	t	0.82
U	0.93	u	0.92	U	0.95	u	0.90
V	0.92	v	0.95	V	0.87	v	0.89
W	0.89	w	0.90	W	0.88	w	0.82
X	0.87	x	0.90	X	0.76	x	0.76
Y	0.90	y	0.90	Y	0.89	y	0.88
Z	0.90	z	0.90	Z	0.90	z	0.91

Note. Items are presented in alphabetical order in this table; however, they were presented to the children in a random order.

DIF by Gender and K-Level. The *Letter Identification* task's anchor set contained 14 items (25% of the total number of items; see Table A2 in Appendix A). After applying B-H correction for multiple comparisons, none of the items had significant DIF by gender. Therefore, girls and boys were merged in subsequent analyses. Also, none of the items had significant DIF by kindergarten level. Thus, none of the letters were removed from the task.

Item Parameter Estimates. The item parameters for junior and senior kindergarten at both assessment times are presented in Table 5.

Fall.

Junior Kindergarten. The discrimination parameters of the letters varied substantially, from moderate to very high. Two of the letters had particularly high discrimination values: 4.35 (letter f) and 4.66 (letter o). These letters had the highest degree of precision among the 54 letters in discriminating children's latent ability, that is, between children with high and low levels of alphabet knowledge.

Looking closely at the discrimination parameters, uppercase letters had discrimination values ranging from 1.50 (letter X) to 3.74 (letter N), while lowercase letters had values ranging from 1.18 (letter l) to 4.66 (letter o). Uppercase letters on average (2.57) were not significantly different in discriminating between low and high performers than were lowercase letters (2.58), both of which are classified as very highly discriminating.

The range of all letters' difficulty parameters was between -1.15 (letter O) and 2.83 (letter q), which is wide enough to cover the entire range of the latent trait with an average of 0.19. The difficulty values for 21 of the 54 letters (62% uppercase) were below zero, meaning that a below-average level of alphabet knowledge was sufficient to identify these letters correctly at an above-chance probability. Of the 21 letters with negative difficulty values, four letters (x, o, X, and O) had difficulty values less than -1, suggesting that children whose level of alphabet knowledge was more than one standard deviation below the mean had a greater than 50% probability of correctly identifying these letters. The remaining 33 letters had difficulty values above zero. The difficulty parameters of 28 of these letters did not exceed one, meaning that children whose latent ability level were between zero and one standard deviation above the mean had a greater than 50% probability of correctly identifying 49 letters. On the other hand, the difficulty parameters of three letters (t, d, and g) exceeded 1, and two letters (l and q) exceeded 2. Thus, only children whose alphabet knowledge level is about three standard deviations above the mean had a greater than 50% chance of correctly identifying all letters of the alphabet.

Examining the letters' difficulty parameters more specifically, the difficulty values for the uppercase letters ranged from -1.15 (letter O) to 0.57 (letter V). On the other hand, the lowercase letters' difficulty values ranged from -1.02 (letter o) to 2.83 (letter q), which is a much wider range than that of the uppercase letters with a substantially higher upper bound. Thus, lowercase letters on average (0.48) were considerably more difficult than uppercase letters (-0.11).

The test information curve (see Figure 5) showed that when all items are considered, the *Letter Identification* task provided the most information for junior kindergarteners with ability levels of about 0.15. This is also where the lowest standard error is observed, indicating good discrimination or high precision. The distribution presented a strong picture, with a peak near the average level of the latent trait and tails spread out on either side. Thus, the test information curve confirmed that this task covered a wide range of latent ability and was neither too easy nor too difficult for junior kindergarteners.

Senior Kindergarten. Like junior kindergarten, the discrimination values of the letters for senior kindergarten varied substantially, from moderate to very high. Uppercase letters had discrimination values ranging from 1.20 (letter X) to 5.69 (letter D), while lowercase letters had values ranging from 0.73 (letter x) to 5.50 (letter v). Uppercase letters on average (2.89) were not significantly different in discriminating between children with different latent ability levels than were lowercase letters (2.50), both of which are classified as very highly discriminating.

The difficulty parameter range for all letters was between -4.08 (letter x) and 1.66 (letter q), with a much higher lower bound. The difficulty values for 50 of the 54 letters (52% uppercase) were below zero, meaning that a below-average level of alphabet knowledge was sufficient to identify these letters correctly at an above-chance probability. Of the 50 letters with negative difficulty values, 26 letters (62% uppercase) had difficulty values less than -1, suggesting that children whose level of knowledge of the alphabet were higher one standard deviation below the mean had a greater than 50% probability of correctly identifying these letters. The remaining four letters had difficulty values above zero. The difficulty parameters of three of these letters (d, g, and l) did not exceed one, meaning that children whose latent ability level were between zero and one standard deviation above the mean had a greater than 50% probability of correctly identifying 53 letters. However, because one letter's (q) difficulty parameter exceeded 1, only children whose level of alphabet knowledge is between one and two standard deviations above the mean had a greater than 50% chance of correctly identifying all letters of the alphabet in both uppercase and lowercase formats.

The uppercase letters' difficulty values ranged from -2.92 (letter X) to -0.58 (letter V), which suggests that uppercase letters were very easy for senior kindergarteners. On the other hand, lowercase letters' difficulty values ranged from -4.08 (letter x) to 1.66 (letter q), which is again a much wider range than that of the uppercase letters with a substantially higher upper bound. Thus, lowercase letters on average (-0.79) were considerably more difficult than uppercase letters (-1.29). Letters q and l had the highest difficulty values among the 54 letters for senior kindergarten as well, with letter d as the fourth and letter b as the seventh most difficult letter.

The test information curve (see Figure 6) showed that when all items are considered, the *Letter Identification* task provided the most information for senior kindergarteners with ability levels of about -0.80. This is also where the standard error was lowest, indicating good discrimination or high precision. As the difficulty parameter range suggested, the task covered latent ability levels below one standard deviation above the mean. There was less discrimination at higher levels of the latent ability compared to junior kindergarten. Hence, the test information curve confirmed that this task was easy for senior kindergarteners and covered a narrower range of latent ability relative to junior kindergarten.

Spring.

Junior Kindergarten. The spring dataset showed a trend similar to that of the fall dataset. The item discrimination parameters for junior kindergarten in the spring varied greatly, from moderate to very high. Both the uppercase and lowercase letters showed a similar range of discrimination parameters as in the fall. Overall, the discrimination values ranged from 1.00 (letter q) to 4.64 (letter U), with a very high average of 2.55.

The item difficulty parameters had a wider range compared to that in the fall, with values between -1.92 (letter X) and 4.55 (letter q), which covers the entire level of the latent trait. The three least difficult letters in the fall (letters X, O, x) continued to be the easiest ones in the spring. Likewise, the six most difficult letters in the fall (q, l, g, d, t, and ɡ) continued to be the hardest ones in the spring. Of note, the number of letters with difficulty values below zero increased from 21 to 46 out of 54 items (57% uppercase). Interestingly, the difficulty value of all letters except for letter q decreased over time. The difficulty values' average dropped to -0.52, which is more than half a standard deviation difference. While the uppercase letters only had difficulty values below zero, lowercase letters had a wide enough range of item difficulty parameters with both negative and positive values (-1.89 to 4.55). Overall, the *Letter Identification* task was easier for junior kindergarteners in the spring than it was in the fall.

The test information curve for the spring dataset (see Figure 7) showed that when all items are considered, the *Letter Identification* task provided the most information for junior kindergarteners with ability levels of about -0.60. This is also where the lowest standard error was observed, indicating good discrimination or high precision. Although there was less information for higher latent ability levels, the task continued to have a large coverage area of the latent trait. Overall, the test information curve confirmed that this task was easy for junior kindergarteners in the spring but could still discriminate well between children who have and have not yet attained an average level of the latent ability.

Senior Kindergarten. The letters' discrimination parameters in the spring varied greatly for senior kindergarten as well, from low to very high. Three letters had an exceptionally high discrimination value of 26.38: A, O, and o. All three of these letters also had a difficulty value of -2.50; super easy yet key items in identifying children who have yet to develop alphabet knowledge. Overall, the discrimination values ranged from 0.57 (letter x) to 26.38 (letter A), with an average of 4.09 that is in the very high discrimination range.

The five most difficult letters for junior kindergarteners, t, d, g, l, and q, were also the hardest for senior kindergarteners. However, these items' difficulty parameter values were very different across the two groups: there was at least one standard deviation difference, with each letter being easier for senior kindergarteners. Overall, the difficulty parameters of the letters for senior kindergarten had a relatively narrow range, with all values between -6.92 (letter x) and 0.12 (letter g) except for letter q with a difficulty value of 1.51. The upper bound of the latent trait was not covered well, and the uppercase letters had a narrower range of difficulty values (all below zero) compared to lowercase letters. The difficulty values' average reduced to -1.63, suggesting that the *Letter Identification* task was much easier for senior kindergarteners in the spring.

The test information curve (see Figure 8) of the *Letter Identification* task for senior kindergarten in the spring was bimodal: it first peaked at around $\theta = -2.50$ and then again at around $\theta = -1.40$. However, the task provided the most information (unusually high) at the first peak, which is also where the lowest standard error was observed, indicating good discrimination or high precision. As the range of the item difficulty parameters indicated, the task provided information on a narrower range of the latent ability for senior kindergarten compared to junior kindergarten. In fact, almost no information was present for latent ability levels above average. Overall, the test information curve confirmed that this task was too easy for senior kindergarteners in the spring.

Table 5

Discrimination (a) and Difficulty (b) Estimates for Letter Identification for Junior (n = 173) and Senior (n = 183) Kindergarten in Fall and Spring

	Fall				Spring			
	Junior Kindergarten		Senior Kindergarten		Junior Kindergarten		Senior Kindergarten	
Letter	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
A	2.35	-0.97	3.62	-2.34	2.15	-1.79	26.38	-2.50
B	2.04	-0.67	2.94	-1.73	2.85	-1.26	2.29	-2.49
C	2.67	-0.37	2.79	-1.36	2.31	-1.29	2.51	-1.79
D	2.84	0.17	5.69	-0.73	3.01	-0.65	3.19	-1.61
E	2.06	-0.22	3.13	-1.19	3.02	-0.98	3.65	-1.42
F	2.95	-0.01	3.23	-1.02	3.93	-0.72	6.12	-1.46
G	2.67	0.36	4.07	-0.75	2.21	-0.42	5.78	-1.23
H	2.15	-0.08	2.93	-1.24	2.55	-0.82	3.79	-1.56
I	2.33	0.41	2.86	-0.79	2.18	-0.46	1.55	-1.74
J	2.02	0.33	2.41	-0.80	1.77	-0.50	2.20	-0.95
K	3.15	-0.22	2.54	-1.25	2.30	-0.88	2.78	-1.54
L	2.26	0.05	2.43	-1.08	2.69	-0.77	1.91	-1.88
M	3.25	-0.25	1.70	-1.85	2.25	-0.84	3.12	-1.79
N	3.74	0.09	3.02	-0.93	3.38	-0.65	3.96	-1.40
O	3.06	-1.15	2.18	-2.78	3.08	-1.85	26.38	-2.50
P	2.19	0.01	2.59	-1.00	3.65	-0.83	4.65	-1.74
Q	2.94	0.43	3.31	-0.85	3.01	-0.22	2.78	-1.36
R	2.83	0.01	2.52	-1.26	3.01	-0.94	5.19	-1.51
S	2.41	-0.57	2.58	-1.88	1.91	-1.33	2.44	-2.70
T	2.80	0.12	3.62	-0.83	2.46	-0.50	4.15	-1.33
U	3.22	0.29	2.93	-0.76	4.64	-0.30	4.21	-1.20
V	2.78	0.57	3.22	-0.58	3.16	-0.40	3.23	-1.19
W	2.10	-0.03	1.77	-1.44	3.02	-0.70	1.46	-2.14
X	1.50	-1.12	1.20	-2.92	1.66	-1.92	0.98	-4.36
Y	2.23	0.18	2.77	-0.80	2.95	-0.44	1.73	-1.65
Z	2.24	-0.27	3.13	-1.40	2.70	-1.23	1.94	-2.55

(Continued)

Table 5
Continued

	Fall				Spring			
	Junior Kindergarten		Senior Kindergarten		Junior Kindergarten		Senior Kindergarten	
Letter	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
ɑ	2.59	-0.12	2.94	-1.14	4.03	-0.79	2.93	-1.93
a	3.02	0.62	2.25	-0.56	2.79	-0.17	2.79	-1.36
b	1.33	0.70	1.95	-0.25	1.00	0.11	1.65	-0.73
c	2.95	-0.29	2.48	-1.20	2.09	-1.08	1.98	-1.90
d	2.23	1.30	1.48	0.07	1.13	0.52	1.69	-0.64
e	2.44	0.16	3.02	-1.04	2.63	-0.73	3.53	-1.43
f	4.35	0.31	4.96	-0.75	2.92	-0.42	4.58	-1.22
ɡ	2.85	0.96	2.08	-0.15	2.60	0.18	2.69	-0.80
g	2.40	1.73	2.02	0.52	1.71	1.44	1.36	0.12
h	2.39	0.75	2.21	-0.76	2.75	-0.17	2.43	-1.26
i	1.85	0.57	2.26	-0.62	1.50	0.03	2.79	-1.02
j	2.33	0.55	2.16	-0.45	1.81	-0.05	2.34	-0.71
k	2.96	-0.13	3.15	-1.19	3.22	-0.79	2.61	-1.64
l	1.18	2.50	1.30	0.57	1.04	1.55	1.44	0.03
m	3.22	-0.03	1.47	-1.62	2.75	-0.68	2.24	-1.72
n	2.50	0.66	3.03	-0.46	3.38	-0.08	3.00	-1.21
o	4.66	-1.02	2.05	-2.58	3.64	-1.84	26.38	-2.50
p	2.12	0.11	2.79	-0.80	2.28	-0.73	2.99	-1.55
q	2.01	2.83	2.60	1.66	1.00	4.55	1.12	1.51
r	2.60	0.36	2.89	-0.89	2.62	-0.69	4.81	-1.41
s	2.31	-0.62	1.87	-1.82	1.72	-1.39	1.04	-3.59
t	1.73	1.13	3.06	-0.06	1.70	0.51	2.76	-0.69
u	3.11	0.67	3.02	-0.48	2.76	-0.16	3.20	-0.91
v	3.02	0.56	5.50	-0.45	3.29	-0.36	4.26	-1.14
w	2.26	0.01	1.78	-1.37	2.28	-0.83	1.05	-2.32
x	2.49	-1.00	0.73	-4.08	1.80	-1.89	0.57	-6.92
y	2.56	0.36	2.45	-0.65	2.91	-0.30	1.76	-1.42
z	2.75	-0.28	2.42	-1.41	2.65	-1.18	2.76	-2.09

Note. Items are presented in alphabetical order in this table; however, they were presented to the children in a random order.

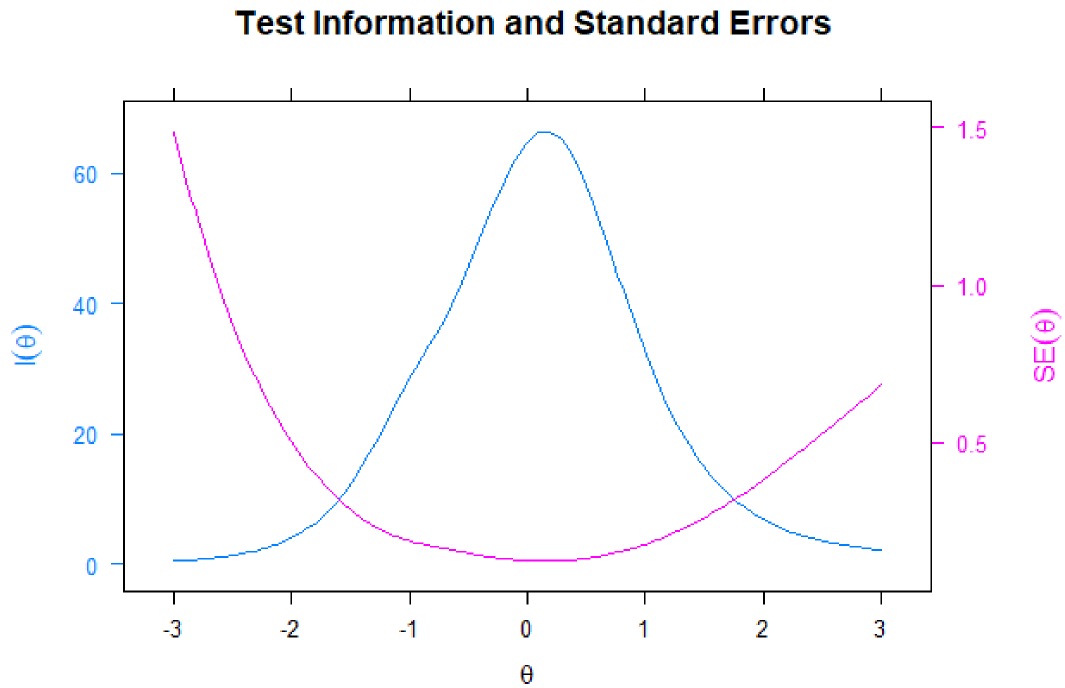

Figure 5
Test Information Curve of the Letter Identification Task for Junior Kindergarten (N = 173) in the Fall

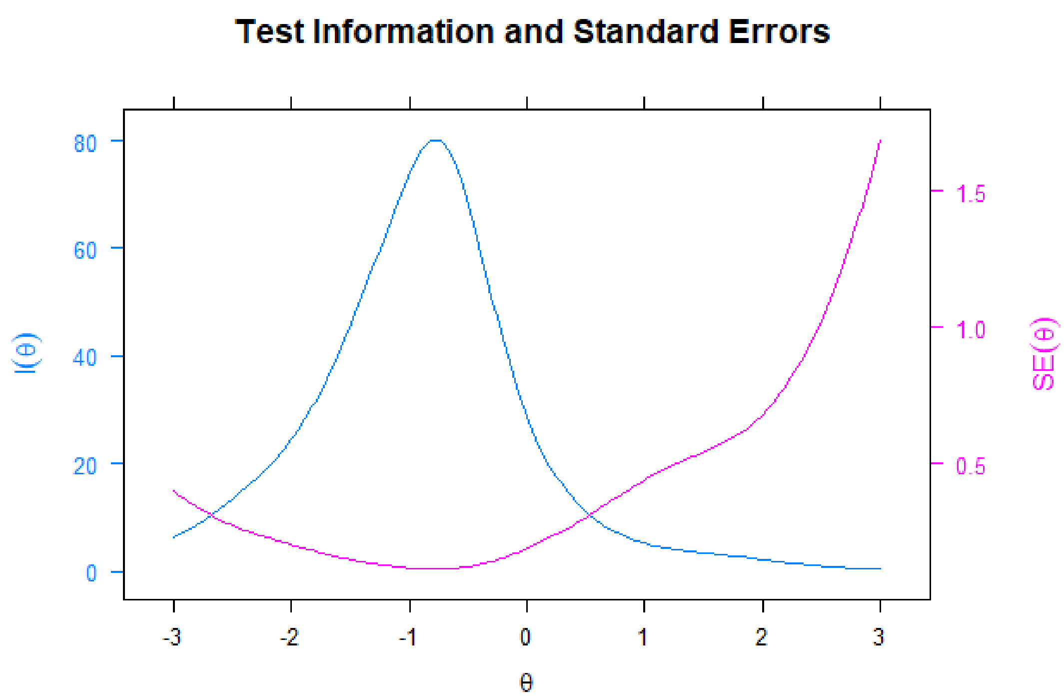

Figure 6
Test Information Curve of the Letter Identification Task for Senior Kindergarten (N = 183) in the Fall

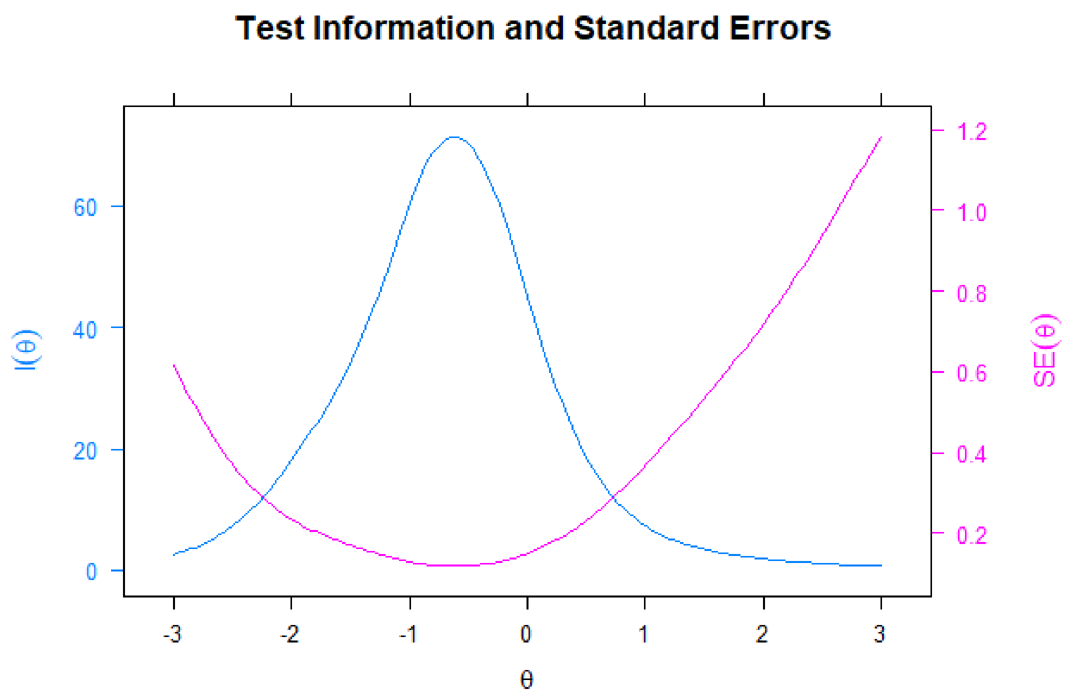

Figure 7
Test Information Curve of the Letter Identification Task for Junior Kindergarten (N = 173) in the Spring

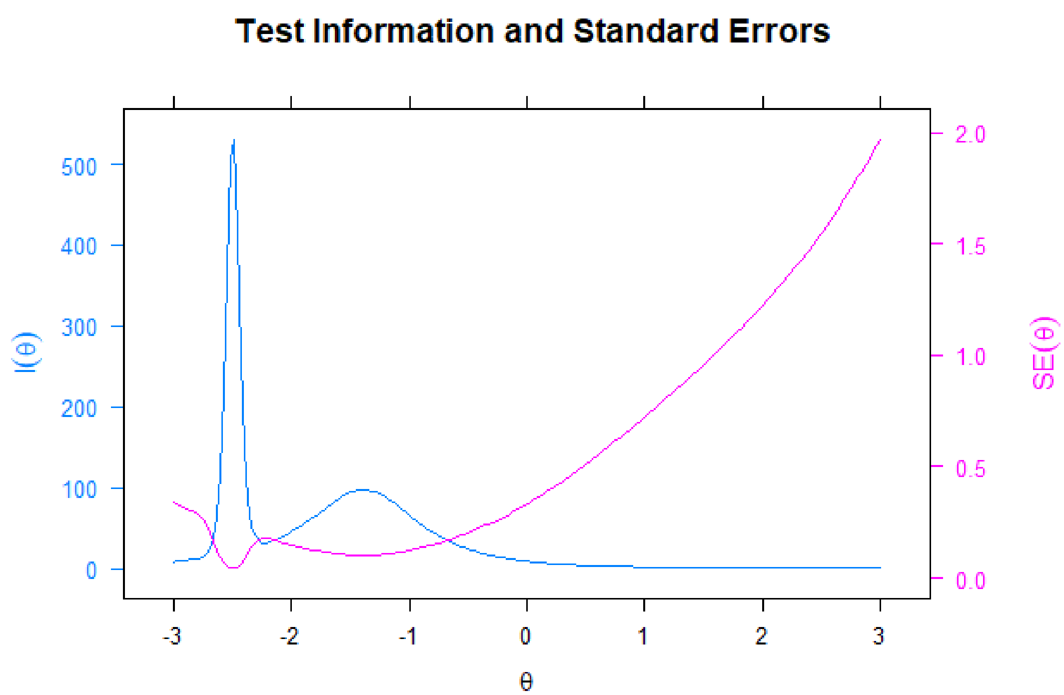

Figure 8
Test Information Curve of the Letter Identification Task for Senior Kindergarten (N = 183) in the Spring

Reliability. The *Letter Identification* task's reliability was α = .98 and .97 in the fall and spring, respectively. When the uppercase and lowercase letters were analyzed separately, the task continued to have excellent reliability coefficients. The uppercase letters' coefficients were .96 and .94 in the fall and spring, whereas the lowercase letters were .97 and .95 in the fall and spring, respectively.

Number Recognition

The item analyses for *Number Recognition* was conducted only for the fall because, as was found through the preliminary analyses, it showed ceiling effects in the spring.

Factor Structure. The first eigenvalue was 31.95, which was noticeably larger than the second value, 5.36. The ratio of the two was 5.96—thus, a unidimensional model was acceptable. The factor loadings are presented in Table 6. All except two of the items, Number 64 and Number 14 (row 14), on the *Number Recognition* task significantly loaded onto a single factor at the alpha level of .05. Other than these two items, Number 82 had a particularly weak loading of 0.17. Although this item met the cut-off criteria (significant loading), I decided to remove this item a posteriori to item parameter estimate analyses where I observed this item to be problematic: its discrimination parameter value was negative. A negative discrimination parameter is a clear indication of a problematic item as discrimination values must be positive.

A second EFA was performed without the following three items: Number 64, Number 82, and Number 14 (row 14). The resulting factor loadings are presented in Table 7. The one-factor model slightly improved after removing these three items, with the first eigenvalue ratio to the second increasing to 7.40. All items significantly loaded onto a single factor at the alpha level of .05. Two of the 52 items had weak loadings of 0.21 and 0.24, while all remaining items had moderately strong to strong loadings with the lowest loading at 0. 0.41. The fit measure indices indicated a good model fit: RMSEA = 0.052, CFI = 0.967, TLI = 0.965, and SRMR = 0.126.

Therefore, the *Number Recognition* task was unidimensional.

Table 6

Exploratory Factor Analysis for the Number Recognition Task for Fall as a Function of Stimuli Size (N = 356)

			Factor Loading				Factor Loading
Stimuli	Row	Item	1	Stimuli	Row	Item	1
0	12	48	0.85	18	3	9	0.87
1	2	5	0.92	18	8	29	0.82
1	8	31	0.90	18	11	44	0.92
2	1	3	0.92	19	4	15	0.93
2	10	40	0.93	19	9	35	0.84
3	9	34	0.79	20	2	8	0.91
4	1	2	0.89	20	9	33	0.45
5	4	16	0.95	20	10	37	0.87
5	10	39	0.93	24	11	41	0.60
6	1	4	0.85	27	11	43	0.61
6	3	10	0.76	29	14	54	0.22
6	8	30	0.85	30	6	22	0.76
6	11	42	0.93	30	14	53	0.81
7	6	24	0.93	33	7	27	0.72
8	2	7	0.91	34	9	36	0.88
8	3	11	0.90	36	6	23	0.41
9	3	12	0.56	38	13	50	0.75
9	12	47	0.83	43	5	19	0.89
10	1	1	0.86	45	14	56	0.74
10	5	17	0.90	46	7	26	0.61
11	5	20	0.90	48	4	14	0.60
12	13	51	0.74	55	6	21	0.85
13	8	32	0.79	64	7	28	0.11
14	7	25	0.25	69	10	38	0.87
14	14	55	0.10	70	5	18	0.93
15	12	46	0.95	82	13	52	0.16
16	12	45	0.88	90	13	49	0.51
17	2	6	0.89				

Note. Item 13 was removed due to zero variance.

Table 7
Second Exploratory Factor Analysis for the Number Recognition Task for Fall as a Function of Stimuli Size (N = 356)

			Factor Loading				Factor Loading
Stimuli	Row	Item	1	Stimuli	Row	Item	1
0	12	48	0.85	17	2	6	0.89
1	2	5	0.92	18	3	9	0.86
1	8	31	0.90	18	8	29	0.82
2	1	3	0.92	18	11	44	0.92
2	10	40	0.93	19	4	15	0.93
3	9	34	0.79	19	9	35	0.84
4	1	2	0.89	20	2	8	0.91
5	4	16	0.95	20	9	33	0.44
5	10	39	0.93	20	10	37	0.88
6	1	4	0.85	24	11	41	0.60
6	3	10	0.76	27	11	43	0.61
6	8	30	0.85	29	14	54	0.21
6	11	42	0.93	30	6	22	0.76
7	6	24	0.93	30	14	53	0.82
8	2	7	0.91	33	7	27	0.72
8	3	11	0.90	34	9	36	0.88
9	3	12	0.56	36	6	23	0.41
9	12	47	0.83	38	13	50	0.75
10	1	1	0.86	43	5	19	0.89
10	5	17	0.91	45	14	56	0.75
11	5	20	0.90	46	7	26	0.61
12	13	51	0.74	48	4	14	0.60
13	8	32	0.79	55	6	21	0.85
14	7	25	0.24	69	10	38	0.87
15	12	46	0.95	70	5	18	0.93
16	12	45	0.88	90	13	49	0.50

Note. Item 13 was removed due to zero variance, and Items 28, 52 and 55 were removed due to insufficient loadings.

DIF by Gender and K-Level. The anchor set of the *Number Recognition* task contained 13 items (25% of the total number of items; see Table A3 in Appendix A). After applying B-H correction for multiple comparisons, none of the items showed significant DIF by gender in the fall. Therefore, girls and boys were merged in subsequent analyses. However, after applying B-H correction for multiple comparisons, four items had significant DIF by kindergarten level in the fall: Numbers 30 and 36 each in row six, Numbers 20 and 69 each in row 10. Hence, a follow-up DIF test was performed to determine whether the differences were uniformed or not. Of the items that showed DIF, all except Number 36 (row six) showed uniform DIF, where the senior kindergarteners had outperformed junior kindergarteners along all points on the latent trait continuum. However, Number 36 (row six) seemed to favour junior kindergarteners up to one standard deviation below the mean latent trait ($\theta = -1$), after which senior kindergarteners had a higher probability of correctly recognizing this number. In sum, these four items were removed from subsequent analyses, resulting in a new total number of 48 items on the *Number Recognition* task.

Item Parameter Estimates. Each item's difficulty and discrimination parameters in the fall for junior and kindergarten are presented in Table 8.

Junior Kindergarten. The discrimination parameters of the items for junior kindergarten varied from very low to very high. Three of the items had especially high discrimination values: 7.36, 7.66, and 10.01 for numbers 2 (row one), 5 (row ten), and 1 (row two), respectively. Overall, the discrimination values ranged from 0.13 (Number 14, row seven) to 10.01 (Number 1, row two), with an average of 3.10. The ten most discriminating numbers were 1, 2, 4, 5, 7, 8, and 19 – with numbers 1, 2 and 5 appearing twice in different rows.

The range of difficulty parameters of all except two items was between -1.91 (Number 90) and 0.65 (Number 30), which is not wide enough to cover the entire level of the latent trait. The two least discriminating items were also notably easy: Number 14 (row seven) had a difficulty value of -9.05, and Number 29 had a difficulty value of -3.53. The difficulty values of

35 items were below zero. Ten items had difficulty values of less than -1, and all items had difficulty values below 1. Thus, children whose latent ability level is about one standard deviations above the mean had a greater than 50% chance of correctly recognizing all numbers on this task. The average of the difficulty values was -0.70, which suggests that, overall, the *Number Recognition* task was easy for junior kindergarteners. Interestingly, the smallest numbers were not necessarily the easiest to recognize for children. In fact, the largest number on the task, Number 90, was the third least difficult number with a difficulty value of -1.91.

The test information curve (see Figure 9) showed that when all items are considered, the *Number Recognition* task provided the most information for junior kindergarteners with ability levels of about -0.85. This is also where the lowest standard error is observed, indicating good discrimination or high precision. Although not wide enough to cover the entire range of the latent trait, the task covered at least one standard deviation above and below the average level of the latent ability. However, there was less information provided for higher levels of latent ability. In sum, the test information curve confirmed that the *Number Recognition* task was easy for junior kindergarteners.

Senior Kindergarten. The discrimination parameters of the numbers for senior kindergarten varied from moderate to very high. Three items had extremely high discrimination values, namely, numbers 1 (rows two and eight) and number 2 (row one) at 38.21, 25,11, and 24,25, respectively. Overall, the discrimination values ranged from 0.71 (Number 48) to 38.21 (Number 1, row two), with an average of 4.59. The ten most discriminating numbers were similar to that of junior kindergarten, except a digit greater than 20 made it to the list: 1, 2, 4, 5, 6, 7, 10, and 70 – numbers 1 and 2 appearing twice.

The ease of this task for senior kindergarteners is reflected by difficulty parameters that were all below zero. This means that the task did not cover the average or higher latent abilities. In fact, 42 items had difficulty values that were less than -1; however, only children whose latent ability levels were between zero and one standard deviation below the mean had a greater than 50% probability of correctly recognizing all numbers on this task. Of note, as was the case with junior kindergarten, smaller numbers were not necessarily the easiest items, and neither were the largest numbers the most difficult. For example, Number 90 was the fifth least difficult item on the task (b = -2.67). On the other hand, Number 45 was the most difficult among all items (b = -0.41). These two numbers were among the largest numbers on the task, yet they had very different parameter values.

The test information curve (see Figure 10) of the *Number Recognition* task for senior kindergarten provided the most information at about θ = -2.10, which also where the lowest standard error was observed, indicating good discrimination or high precision. As the range of the item difficulty parameters indicated, the task provided almost no information for an average level of latent ability or higher. Overall, the test information curve demonstrated that this task was extremely easy for senior kindergarteners.

Table 8

Discrimination (a) and Difficulty (b) Estimates for Number Recognition for Junior (n = 173) and Senior (n = 183) Kindergarten in Fall and Spring

			Junior Kindergarten		Senior Kindergarten	
Stimuli	Row	Item	*a*	*b*	*a*	*b*
0	12	48	3.55	-0.82	4.06	-2.18
1	2	5	10.01	-1.15	38.21	-2.12
1	8	31	6.65	-1.07	25.11	-2.05
2	1	3	7.36	-0.96	24.25	-2.05
2	10	40	6.76	-0.81	9.70	-1.94
3	9	34	2.57	-0.94	2.40	-2.35
4	1	2	5.87	-0.86	6.58	-2.19
5	4	16	6.51	-0.68	2.93	-2.13
5	10	39	7.66	-0.72	4.86	-1.88
6	1	4	3.46	-0.81	5.38	-2.08
6	3	10	1.99	-0.64	2.23	-1.76
6	8	30	3.53	-0.68	2.11	-2.48
6	11	42	4.89	-0.40	3.15	-1.52
7	6	24	5.89	-0.46	4.67	-1.69
8	2	7	5.35	-0.53	4.02	-1.72
8	3	11	1.41	0.61	1.91	-0.63
9	3	12	1.07	-1.06	1.62	-1.96
9	12	47	2.09	-0.43	3.36	-1.53
10	1	1	4.28	-0.87	3.98	-2.18
10	5	17	3.03	-0.16	5.39	-1.35
11	5	20	3.57	-0.20	3.98	-1.26
12	13	51	1.07	0.65	1.92	-0.89
13	8	32	1.95	-0.10	2.06	-1.53
14	7	25	0.13	-9.05	0.84	-3.05
15	12	46	3.61	0.10	4.17	-1.11
16	12	45	1.97	0.03	3.45	-1.07
17	2	6	1.92	0.03	2.47	-1.26
18	3	9	1.29	0.46	1.54	-0.82
18	8	29	1.96	0.05	2.32	-1.39
18	11	44	3.91	-0.32	4.36	-1.40
19	4	15	5.12	-0.54	2.06	-2.03
19	9	35	1.97	-0.18	2.49	-0.95

(Continued)

Table 8
Continued

Stimuli	Row	Item	Junior Kindergarten *a*	Junior Kindergarten *b*	Senior Kindergarten *a*	Senior Kindergarten *b*
20	2	8	2.13	0.09	3.18	-1.15
20	9	33	0.94	-1.48	1.15	-2.16
24	11	41	1.32	-1.57	1.22	-3.23
27	11	43	1.48	-1.31	1.76	-2.07
29	14	54	0.42	-3.53	0.77	-2.88
30	14	53	1.37	0.65	2.15	-0.69
33	7	27	1.59	-0.38	1.18	-2.45
34	9	36	2.31	-0.17	2.87	-1.06
38	13	50	1.45	0.05	1.86	-1.11
43	5	19	2.91	-0.20	3.54	-1.43
45	14	56	1.27	0.55	1.66	-0.44
46	7	26	1.01	0.30	0.85	-1.30
48	4	14	1.63	-1.84	0.71	-6.78
55	6	21	2.50	-0.34	3.73	-1.37
70	5	18	2.93	0.15	4.93	-1.14
90	13	49	1.00	-1.91	1.40	-2.67

Note. Item 13 was removed due to zero variance. Items 28, 52 and 55 were removed due to insufficient loadings. Items 22, 23, 67 and 38 were removed due to significant DIF by kindergarten level.
The final total number of items was 48.

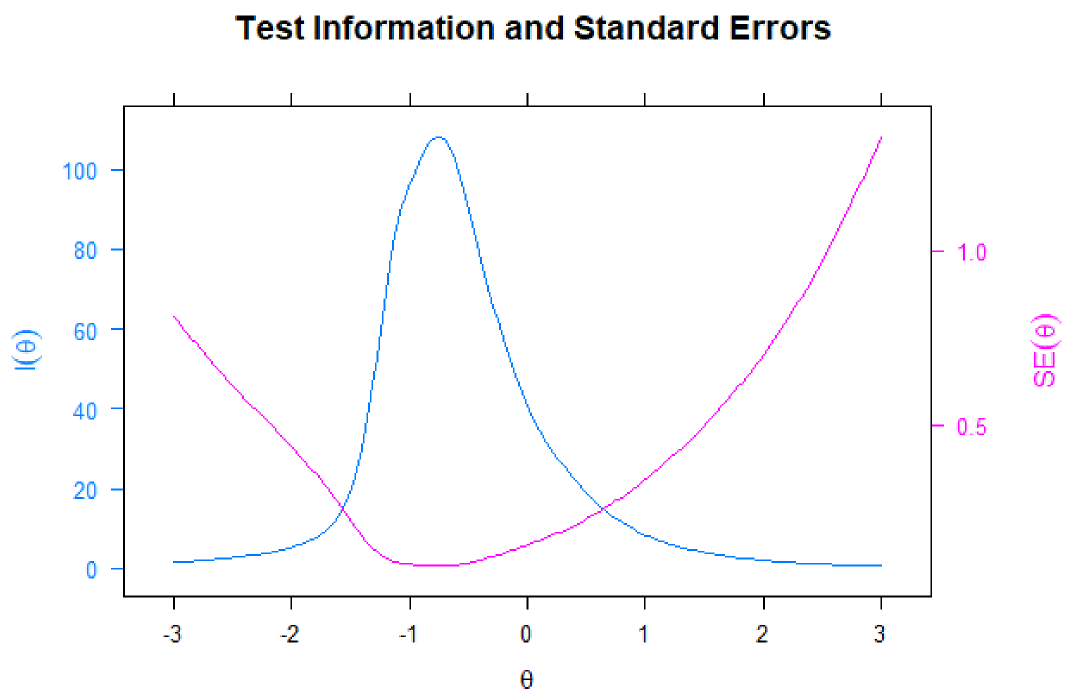

Figure 9
Test Information Curve of the Number Recognition Task for Junior Kindergarten (N = 173) in the Fall

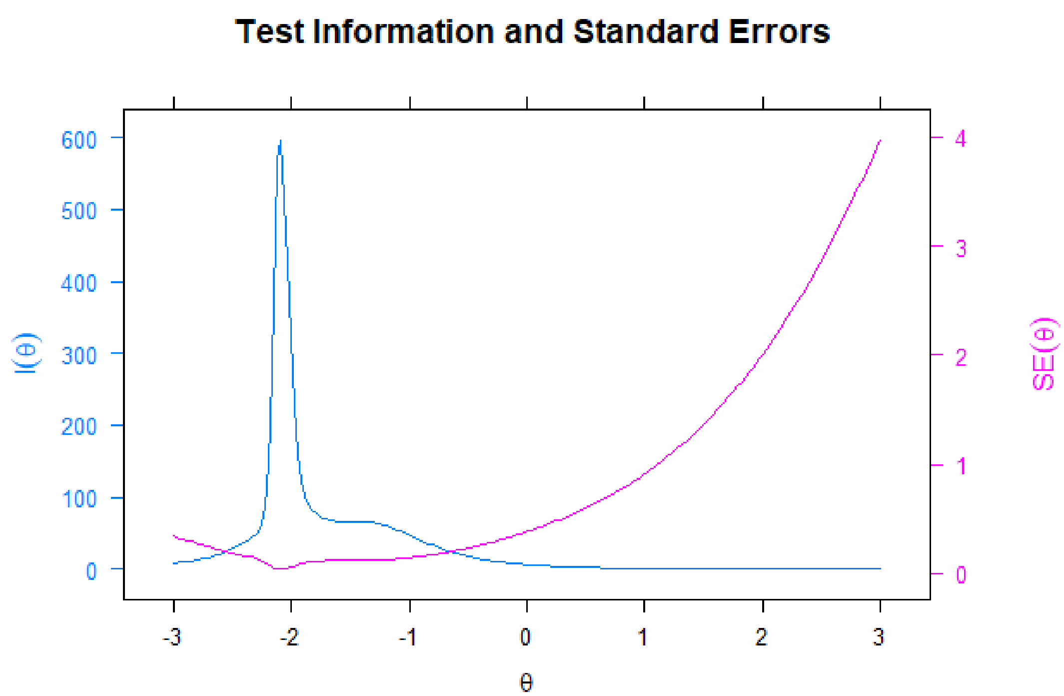

Figure 10
Test Information Curve of the Number Recognition Task for Senior Kindergarten (N = 183) in the Fall

Reliability. After eliminating items with zero variance, insufficient loadings, and significant DIF, the *Number Recognition* task had very good reliability of $\alpha = .96$ in the fall. Recall that the spring reliability could not be assessed because the task showed ceiling effects.

Measurement Invariance Results

Configural invariance was examined for all three variables: gender, kindergarten level, and time. Other levels of invariance (e.g.., scalar invariance) could not be tested.

Preliminary Analyses

Normality. I examined the normality of the data set by inspecting histograms and normal probability plots (P-P plots) and evaluating kurtosis and skew. The statistics for each kindergarten level at both assessment times are reported in Table 9; however, junior and senior kindergarten were merged during normality analyses as the measurement invariance analyses were conducted on the joint sample ($N = 356$), except for two additional invariance analyses across time, where each group alone had sufficient data points (N of junior kindergarten = 346, N of senior kindergarten = 366).

Fall. Visual inspection of the five tasks' histograms as well as P-P plots using both junior and senior kindergarten data showed that none of the tasks had a normal distribution. *Print Concepts* was too flat (kurtosis = -1.03, skewness = -0.15); *Letter Identification* was too flat and negatively skewed (kurtosis = -1.23, skewness = -0.53); and *Early Spelling* (kurtosis = 1.83, skewness = -1.31), *Number Naming* (kurtosis = 4.67, skewness = 1.65) and *Magnitude Comparison* (kurtosis = 0.35, skewness = 0.75) were too peaked and positively skewed. The assumption of multivariate normality was not met when each kindergarten level was analyzed separately, either.

Spring. Like In the spring, *Print Concepts* was negatively skewed (kurtosis = -0.08, skewness = -0.80); *Letter Identification* was too peaked and negatively skewed (kurtosis = 0.56, skewness = -1.25); *Early Spelling* was too flat and positively skewed (kurtosis = -0.52, skewness = 0.75); *Number Naming* was too peaked and positively skewed (kurtosis = 1.54, skewness = 1.30); and *Magnitude Comparison* was slightly positively skewed (kurtosis = -0.45, skewness = 0.38). The assumption of multivariate normality was not met when each kindergarten level was analyzed separately, either.

Table 9
Descriptive Statistics of the Retained Tasks of the ELNOT for Junior (n = 173) and Senior (n = 183) Kindergarten for Fall and Spring

	Fall							
	Junior Kindergarten				Senior Kindergarten			
Task	*M*	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis
Print Concepts	5.92	2.83	0.29	-0.65	9.01	3.00	-0.76*	-0.19
Letter Identification	24.71	16.45	0.12	-1.39*	41.38	13.78	-1.43*	0.80*
Early Spelling	4.00	4.39	1.20*	1.66*	12.10	8.19	0.98*	0.77*
Number Naming	7.64	5.12	1.74*	7.06*	15.04	8.20	1.58*	4.16*
Magnitude Comparison	12.14	4.31	1.00*	1.59*	19.56	6.30	0.45*	0.31
	Spring							
	Junior Kindergarten				Senior Kindergarten			
	M	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis
Print Concepts	7.95	3.04	-0.39	-0.52	10.32	2.72	-1.48*	2.16*
Letter Identification	35.92	14.26	-0.79*	-0.47	46.23	10.08	-2.03*	3.48*
Early Spelling	7.13	6.95	1.42*	2.05*	17.79	10.47	0.23	-1.14*
Number Naming	11.80	7.01	1.64*	5.27*	21.46	11.62	0.92*	0.24
Magnitude Comparison	16.48	5.48	0.70*	0.04	25.27	7.09	-0.05	0.07

Note. Reading, *Verbal Counting*, *Number Recognition*, *Give-N* and *Compare Sets* were eliminated due to floor or ceiling effects.
* $p < .05$

Descriptive Statistics and Correlations. The means and standard deviations of the five tasks for each kindergarten level at both assessment times are presented in Table 9. Senior kindergarteners outperformed junior kindergarteners on both literacy and numeracy tasks at both assessment times. The correlations for junior and senior kindergarten in the fall and spring are reported in Tables 10 and 11. At both times, all retained tasks were significantly correlated with each other within each kindergarten level as well as when both groups were analyzed together.

Table 10
Correlations Among Tasks Retained for Junior (n = 173; below the diagonal) and Senior (n = 183; above the diagonal) Kindergarten in the Fall

	1	2	3	4	5
1. Print Concepts		.54	.55	.43	.54
2. Letter Identification	.43		.64	.59	.56
3. Early Spelling	.43	.67		.65	.55
4. Number Naming	.53	.69	.56		.64
5. Magnitude Comparison	.40	.36	.38	.53	

Note. The *Reading*, *Verbal Counting*, *Number Recognition*, *Give-N*, and *Compare Sets* tasks were eliminated due floor or ceiling.
All correlations were significant at $p < .01$

Table 11
Correlations Among Tasks Retained for Junior (n = 173; below the diagonal) and Senior (n = 183; above the diagonal) Kindergarten in the Spring

	1	2	3	4	5
1. Print Concepts		.59	.56	.51	.55
2. Letter Identification	.57		.62	.52	.51
3. Early Spelling	.51	.62		.62	.46
4. Number Naming	.51	.64	.61		.66
5. Magnitude Comparison	.53	.51	.55	.63	

Note. The *Reading*, *Verbal Counting*, *Number Recognition*, *Give-N*, and *Compare Sets* tasks were eliminated due floor or ceiling.
All correlations were significant at $p < .01$

Invariance Across Gender

The CFA model of an early academic skills factor was tested for invariance across boys and girls across kindergarten level as a function of testing time ($N = 356$). The standardized factor loadings from the baseline model for each group (boys and girls) at both assessment times are presented in Table 12.

Table 12
Standardized Factor Loadings and Standard Errors for Boys (n = 170) and Girls (n = 186) in Fall and Spring

	Fall		Spring	
Early Academic Skills	Boys	Girls	Boys	Girls
Print Concepts	.69 (.05)	.76 (.04)	.75 (.04)	.73 (.04)
Letter Identification	.78 (.04)	.83 (.03)	.73 (.04)	.80 (.03)
Early Spelling	.83 (.03)	.86 (.02)	.83 (.03)	.83 (.03)
Number Naming	.89 (.02)	.86 (.02)	.89 (.02)	.84 (.03)
Magnitude Comparison	.81 (.03)	.78 (.03)	.82 (.03)	.81 (.03)

Note. The *Reading*, *Verbal Counting*, *Number Recognition*, *Give-N*, and *Compare Sets* tasks were eliminated due floor or ceiling.

Fall. The goodness-of-fit indices indicated a good model fit: RMSEA = .076, CFI = .991, TLI = .981, and SRMR = .019. The 5-task ELNOT had a single underlying construct that both the early literacy and numeracy tasks loaded onto, and factor loadings associated with both boys' and girls' responses had the same pattern. Hence, the CFA model involving the five tasks of the ELNOT had configural invariance across gender.

After establishing configural invariance, I tested for metric invariance. Because the data did not have multivariate normality, I computed the Satorra-Bentler scale chi-square difference ($\Delta SBS\chi^2$) under the MLM estimation instead of using the usual chi-square values (Satorra & Bentler, 1999). As presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2(4) = 18.81$, $p < .001$, which means that there is not complete invariance of the factor loadings between boys and girls. The CFI difference between the models further supported this finding: ΔCFI = -.02, which was less than the -.01 mark that signals a lack of invariance in the respective comparison of nested models (Cheung & Rensvold, 2002). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Spring. Similar results were obtained with the spring data to that of the fall. The model had an adequate fit: RMSEA = .139, CFI = .968, TLI = .936, and SRMR = .029. Thus, the CFA model involving the five tasks of the ELNOT had configural invariance across gender at both assessment times.

After establishing configural invariance, I tested for metric invariance. As presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2$ (4) = 17.11, p = .002, which means that there is not complete invariance of the factor loadings between boys and girls. The CFI difference between the models further supported this finding: ΔCFI = -.013, which was less than the -.01 mark that signals a lack of invariance in the respective comparison of nested models (Cheung & Rensvold, 2002). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Invariance Across Kindergarten-Level

The CFA model of a single factor (early academic skills) was tested for invariance across kindergarten level (N = 356). The standardized factor loadings from the baseline model for each group (junior and senior kindergarten) at both assessment times are presented in Table 13.

Table 13

Standardized Factor Loadings and Standard Errors for Junior (n = 173) and Senior (n = 183) Kindergarten in Fall and Spring

Early Academic Skills	Fall		Spring	
	Junior Kindergarten	Senior Kindergarten	Junior Kindergarten	Senior Kindergarten
Print Concepts	.60 (.06)	.66 (.05)	.68 (.05)	.73 (.04)
Letter Identification	.81 (.04)	.77 (.04)	.78 (.04)	.74 (.04)
Early Spelling	.73 (.05)	.80 (.03)	.76 (.04)	.76 (.04)
Number Naming	.84 (.04)	.78 (.04)	.81 (.04)	.78 (.04)
Magnitude Comparison	.55 (.06)	.75 (.04)	.73 (.04)	.72 (.04)

Note. The *Reading*, *Verbal Counting*, *Number Recognition*, *Give-N*, and *Compare Sets* tasks were eliminated due floor or ceiling.

Fall. The goodness-of-fit indices indicated a good model fit: RMSEA = .131, CFI = .959, TLI = .918, and SRMR = .034. The ELNOT with the five retained tasks had a single underlying construct, and factor loadings associated with both junior and senior kindergarteners' responses had the same pattern. Hence, the CFA model involving the five tasks of the ELNOT had configural invariance across kindergarten level.

After establishing configural invariance, I tested for metric invariance. Because the data did not have multivariate normality, I computed the Satorra-Bentler scale chi-square difference ($\Delta SBS\chi^2$) under the MLM estimation instead of using the usual chi-square values (Satorra & Bentler, 1999). As presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2(4) = 45.579$, $p < .001$, which means that there is not complete invariance of the factor loadings between boys and girls. The goodness-of-fit indices of Model 1 also indicated poor fit: RMSEA was .178, CFI = .894, TLI = .894, and SRMR = .114. The decrease in the CFI value was too great ($\Delta CFI = -.065$), signalling a lack of invariance in the respective comparison of nested models (Cheung & Rensvold, 2002). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Spring. Similar results were obtained with the spring data to that of the fall. The baseline model had good fit indices: RMSEA = .137, CFI = .958, TLI = .915, and SRMR = .033. Thus, configural invariance was established across junior and senior kindergarten. However, as presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2$ (4) = 51.383, p < .001, which means that there is not complete invariance of the factor loadings between junior and senior kindergarten. The goodness-of-fit indices of Model 1 also indicated poor fit: RMSEA = .170, CFI = .909, TLI = .870, and SRMR = .120. The decrease in the CFI value was too much (ΔCFI = -.049). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Invariance Across Time

The CFA model of a single factor (early academic skills) was tested for invariance across time using junior and senior kindergarten together (N = 712), only junior kindergarten (N = 346), and only senior kindergarten (N = 366). The standardized factor loadings from the baseline model for each assessment time using the samples mentioned above are presented in Table 14.

Table 14
Standardized Factor Loadings and Standard Errors by Time for Junior (n = 173) and Senior (n = 183) Kindergarten

	Junior and Senior Kindergarten		Junior Kindergarten		Senior Kindergarten	
Early Academic Skills	Fall	Spring	Fall	Spring	Fall	Spring
Print Concepts	.73 (.03)	.75 (.03)	.60 (.06)	.68 (.05)	.66 (.05)	.73 (.04)
Letter Identification	.81 (.02)	.76 (.03)	.81 (.04)	.78 (.04)	.77 (.04)	.74 (.04)
Early Spelling	.84 (.02)	.82 (.02)	.73 (.05)	.76 (.04)	.80 (.03)	.76 (.04)
Number Naming	.85 (.02)	.84 (.02)	.84 (.04)	.81 (.04)	.78 (.04)	.78 (.04)
Magnitude Comparison	.80 (.02)	.82 (.02)	.55 (.06)	.73 (.04)	.75 (.04)	.72 (.04)

Note. The *Reading*, *Verbal Counting*, *Number Recognition*, *Give-N*, and *Compare Sets* tasks were eliminated due floor or ceiling.

Junior and Senior Kindergarten. The goodness-of-fit indices indicated a good model fit: RMSEA = .113, CFI = .978, TLI = .956, and SRMR = .022. The ELNOT with the five retained tasks had a single underlying construct, and factor loadings associated with junior and senior kindergarteners' performances in both fall and spring had the same pattern. Hence, the CFA model involving the five tasks of the ELNOT had configural invariance across time.

After establishing configural invariance, I tested for metric invariance. Because the data did not have multivariate normality, I computed the Satorra-Bentler scale chi-square difference ($\Delta SBS\chi^2$) under the MLM estimation instead of using the usual chi-square values (Satorra & Bentler, 1999). Although the fit of Model 1 indicated an acceptable fit (RMSEA = .146, CFI = .949, TLI = .927, and SRMR = .093), as presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2\,(4) = 66.962$, $p < .001$, which means that there is not complete invariance of the factor loadings between fall and spring. The CFI value decrease also indicated a lack of invariance between the models compared ($\Delta CFI = -.029$). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Junior Kindergarten. A sample of just junior kindergarteners ($N = 346$) produced similar results to that of the joint sample of both kindergarten levels. The baseline model had good fit indices: RMSEA = .118, CFI = .966, TLI = .932, and SRMR = .032. Thus, configural invariance was established across time for junior kindergarten alone as well. However, as presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2\,(4) = 20.154$, $p < .001$, which means that there is not complete invariance of the factor loadings between fall and spring. The fit of Model 1 was also poorer: RMSEA = .141, CFI = .933, TLI = .904, and SRMR = .085. The decrease in the CFI value was indicative of a lack of invariance as well (ΔCFI = -.033). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Senior Kindergarten. Like junior kindergarten, a sample of just senior kindergarteners (N = 366) produced a baseline model with good fit: RMSEA = .148, CFI = .966, TLI = .932, and SRMR = .035. Hence, configural invariance was established across time for senior kindergarten alone as well. However, as presented in Table 15, the chi-square difference between Model 1 and Model 0 was statistically significant, $\Delta SBS\chi^2(4) = 24.834$, $p < .001$, which means that there is not complete invariance of the factor loadings between fall and spring. The fit of Model 1 was very poor: RMSEA = .154, CFI = .926, TLI = .894, and SRMR = .101. The decrease in the CFI value was too much (ΔCFI = -.040). Because metric invariance was not established, the next stages in multiple-group CFA were not tested.

Reliability

The five tasks of the ELNOT produced acceptable reliability coefficients of $\alpha = .82$ in the fall and $\alpha = .86$ in the spring.

Table 15

Measurement Invariance Tests Across Gender, Kindergarten Level, and Time (N = 356)

Testing Across	Sample	Model	χ^2	SBS χ^2	*df*	CFI	SRMR	Model Comparison	ΔSBS χ^2	Δdf	ΔCFI [a]	Decision
Gender	Fall	M0 Configural invariance	20.249	20.641	10	.991	.019					
		M1 Metric invariance	45.893	42.094	14	.971	.069	M1-M0	18.809**	4	-.020	Reject
	Spring	M0 Configural invariance	44.224	46.79	10	.968	.029					
		M1 Metric invariance	62.452	63.757	14	.955	.071	M1-M0	17.108*	4	-.013	Reject
K-Level	Fall	M0 Configural invariance	40.487	37.953	10	.959	.034					
		M1 Metric invariance	92.985	85.224	14	.894	.114	M1-M0	45.579**	4	-.065	Reject
	Spring	M0 Configural invariance	43.523	40.674	10	.968	.029					
		M1 Metric invariance	85.686	85.792	14	.955	.071	M1-M0	51.383**	4	-.049	Reject
Time	JK and SK	M0 Configural invariance	55.567	53.055	10	.978	.022					
		M1 Metric invariance	120.649	117.615	14	.949	.093	M1-M0	66.962**	4	-.029	Reject
	JK	M0 Configural invariance	34.181	30.751	10	.966	.032					
		M1 Metric invariance	62.071	52.189	14	.933	.085	M1-M0	20.154**	4	-.033	Reject
	SK	M0 Configural invariance	49.829	48.601	10	.966	.035					
		M1 Metric invariance	74.743	73.352	14	.926	.101	M1-M0	24.834**	4	-.040	Reject

Note. χ^2 = conventional chi-square fit statistic; SBS χ^2 = Satorra-Bentler scale chi-square test (under maximum-likelihood estimation);
ΔSBS χ^2 = Satorra-Bentler scales chi-square difference; CFI = comparative fit index; SRMR = standardized root mean square residual;
M0 = baseline model (no invariance imposed); M1 = invariant factor loadings;
K-Level = kindergarten level; JK = junior kindergarten (n = 173); SK = senior kindergarten (n = 183)
[a] ΔCFI ≤ -0.01 signals lack of invariance in the respective comparison of nested models.
* $p < .01$
** $p < .001$

Discussion

The purpose of this thesis was to evaluate the psychometric properties of the ELNOT through two series of analyses. First, item analyses were performed on three tasks that had item-level data (and were eligible) to determine the tasks' degree of precision in measuring kindergarten children's early academic skills. These three tasks are: *Print Concepts, Letter Identification,* and *Number Recognition*. Items on these tasks did not function significantly differently across gender. However, *Print Concepts* and *Number Recognition* had a few items with significant DIF across kindergarten level, which were removed from subsequent analyses. The three tasks were more suitable to be administered in the fall and for children in junior kindergarten, since the tasks covered a better range of abilities for junior kindergarteners compared to senior kindergarteners. The tasks were easy, if not too easy, for senior kindergarteners, even in the fall. Overall, *Print Concepts, Letter Identification,* and *Number Recognition* demonstrated sufficient measurement precision for a majority of children in junior kindergarten and for some children in senior kindergarten.

In the second series of analyses, measurement invariance tests were performed on five tasks that were retained after preliminary analyses using the multiple-group CFA method. These tasks were *Print Concepts*, *Letter Identification*, *Early Spelling*, *Number Naming*, and *Magnitude Comparison*. The ELNOT with these five tasks presented a single factor model with adequate data fit. The model structure was invariant across all comparisons; that is, the ELNOT had configural invariance across gender, kindergarten level, and time. However, metric invariance, and therefore, other higher forms of invariances, were not established.

Item Analyses

The *Print Concepts* and *Letter Identification* tasks showed similar trends with regards to item difficulty parameter across kindergarten level at both assessment timepoints. As would be expected, the tasks were easier for senior kindergarteners; in addition, the overall difficulty decreased over time for the two groups because children improved their abilities in between the testing times. Both tasks were neither too easy nor too difficult for junior kindergarteners in the fall but became easy in the spring. The two tasks were easy for senior kindergarteners in the fall and became even easier in the spring. The *Number Recognition* task was easy for junior kindergarteners, but extremely easy for senior kindergarteners in the fall.

In general, tasks are less precise in discriminating at very low or very high levels of the latent ability. However, it is important to consider both item difficulty and discrimination parameters together because a low discriminating item is not necessarily too easy or too difficult. The tasks that were tested for item analyses had items with low discrimination values that were also easy, average, or difficult. For example, in *Print Concepts* (junior kindergarten; fall), Item 15–"Small letter" was very easy (b = -3.78), Item 14–"Capital letter" was average (b = 0.03), and Item 16–"Question mark" was very difficult (b = 1.68). All these items had discrimination values below 0.40. The opposite is also true; although the test information curve generally peaks around the average level of latent ability, a highly discriminating item does not necessarily have an average level of difficulty. For example, in *Letter Identification* (junior kindergarten; fall), letter O was very easy (b = -1.15), letter R had an average level of difficulty (b = 0.01), and letter q was very difficult (b = 2.83). All these letters had discrimination values above 2.00. In *Number Recognition*, of the items with high discrimination values ($a \geq 1.35$), 76% and 100% had negative difficulty values for junior and senior kindergarten, respectively. Hence, there was no specific pattern between the discrimination and difficulty parameters of the items.

Print Concepts

For junior kindergarten in the fall, the *Print Concepts* task showed better precision for children with average or above average latent abilities than it did for children with below average abilities. In other words, the task is unlikely to be accurate in identifying junior kindergarteners who may need additional support because children who have sufficient knowledge of print concepts for their age may also get certain items incorrect. However, in the spring, the task had better precision for average or below average latent abilities than it did for children with above average abilities. This means that the task is likely to be accurate in identifying junior kindergarteners who have had insufficient progress within the academic year and may need additional support. So, depending on the purpose of administration (e.g., identifying children who are potentially at-risk or gifted), the appropriate time of administration of this task changes. If aiming to identify potentially gifted children, it is better to administer *Print Concepts* in the fall, whereas if aiming to identify children who are having difficulties in early literacy, it is better to administer it in the spring.

For senior kindergarten in the fall, *Print Concepts* had better precision for children with below average latent abilities than it did for children with average abilities or higher. Almost all items on *Print Concepts* had negative difficulty values. Thus, the task can identify senior kindergarteners with low latent ability levels. In the spring, the latent abilities for which *Print Concepts* showed precision shifted to lower levels. The amount of information that *Print Concepts* provided for below average abilities was reduced in the spring. In other words, the task may only be able to identify children at low levels of the latent ability. However, given the reduction in precision, there is very little value in administering this task to senior kindergarteners in the spring.

Letter Identification

For junior kindergarten in the fall, the *Letter Identification* task covered a good range of the latent ability, thus good precision across all levels. However, in the spring, it showed better precision for children with average or below average abilities than it did for children with above average abilities. Hence, regardless of the purpose of administration, it is appropriate to administer this task in the fall to junior kindergarteners; however, in the spring, it is appropriate to administer it only if the purpose is to identify children who are at risk.

For senior kindergarteners even in the fall, *Letter Identification* showed better precision for children with below average latent abilities compared to children with average abilities or higher. Almost all letters had negative difficulty values, including all uppercase letters. Thus, there is no value in administering the uppercase letters in the *Letter Identification* task to children in senior kindergarten unless the purpose of administration would be to identify those children who are at risk. In the spring, the amount of information that *Letter Identification* provided at lower levels of the latent trait peaked significantly. In other words, even though the *Letter Identification* task was too easy for senior kindergarteners in the spring, the task could discriminate children with very low latent abilities. For example, a senior kindergartener who cannot yet correctly identify the letter A (a = 26.38, b = -2.50) in the spring should be considered for targeted intervention.

During the administration process, assessors had observed that five lowercase letters were highly confusing for children because of their visual similarities: letters that are mirror images of one another (i.e., b and d, p and q) and letter l that looks like number 1. For both junior and senior kindergarten at both assessment times, these letters were among the least discriminating letters. In addition, these visually confusing letters tended to be more difficult compared to other letters. For example, letters q and l had the highest difficulty values among the 54 letters for both kindergarten levels. Hence, teachers and educators can focus more on these letters when teaching them.

Consistent with Bowles and colleagues' (2014) findings, lowercase letters were more difficult than uppercase letters. Moreover, lowercase letters alone provided an equally precise assessment of children's alphabet knowledge compared to all 54 letters included. Hence, unless the target is to measure children's knowledge of specific letters in specific forms (e.g., uppercase), teachers and educators can test children's alphabet knowledge through lowercase letters just as precisely as including both forms of letters.

Alternatively, letters with the same difficulty values could be compared, and the one with the lower discrimination value could be dropped (see Tables C1 and C2 in Appendix C). There were six letters that could be dropped in the fall for each kindergarten level, as well as seven letters for junior kindergarten and six for senior kindergarten that could be dropped in the spring. If all letter comparisons are considered together (across kindergarten level and time), 22 letters (55% uppercase) would be dropped from the task, leaving a total of 32 letters. This procedure serves more than merely the shortening of a task. Multiple items with the same difficulty parameter value might reduce the precision of a task because of double counting of information on the same latent ability level. Thus, some children's total scores on the task could be inflated (Purpura & Lonigan, 2015). Therefore, the letters recommended to be dropped can be found in Tables C1 and C2.

Another adjustment that could be made would be to administer different letters for fall and spring assessment, especially for children who have correctly identified the letters in the fall–they are not likely to forget what they have learned (Confrey et al., 2014). Rather, children build on their existing knowledge by learning new letters during the academic year. All letters became easier in the spring, so shortening the task by selecting the most difficult letters among these now easier ones would increase the efficiency of the testing. This improvement would be especially the case if the dropped items in the spring were already tested in the fall. Of course, if the children could not correctly identify certain letters in the fall, a follow-up may be beneficial to monitor children's growth.

Number Recognition

The *Number Recognition* task covered a satisfactory range of latent ability for junior kindergarten; however, most information was provided for children with below average abilities. Despite being an easy task for junior kindergarteners even in the fall, depending on the purpose of its administration (e.g., identification of at-risk children), *Number Recognition* could still contribute to the precision of the ELNOT. For senior kindergarten in the fall, the *Number Recognition* task showed precision only for children with below average latent abilities. All items had difficulty values below zero. Hence, unless it is to identify children who need additional support, there is no value in administering this task to senior kindergarteners.

Several numbers had identical difficulty parameter values. Items recommended to be dropped from the task are presented in Table C3 (see Appendix C). There were six items that could be dropped for junior kindergarten and five for senior kindergarten. If all item comparisons are considered together (across kindergarten level), nine items would be dropped from the task, leaving a total of 39 items.

Suggestions for Revision

Children are expected to eventually reach perfect scoring on some of the tasks of the ELNOT. Although tasks like *Give-N* and *Compare Sets* that showed severe ceiling effects might as well be removed from the ELNOT, or administered to only children who are experiencing difficulties in early numeracy, it may be possible to increase the precision of easy tasks for children with higher latent abilities by adding items with higher difficulty levels. However, it is not necessary to do so if the primary purpose of administration is to identify children who are most in need of help. Also, easy tasks without ceiling effects may still be informative for children with higher latent abilities. For example, on the *Letter Identification* task, a score close to 54 (perfect score) would inform the teachers and educators of how close the child is to knowing all the letters, and which few letters they must work on.

Although some showed improvements between the assessment times, a majority of children in both junior and senior kindergarten found the *Reading* task extremely difficult, which was evident by extreme floor effects at all time points. Hence, unless the goal is to identify children who may be gifted, there is no value in administering this task to children in kindergarten level at any time.

A specific recommendation towards the revision of the ELNOT pertains to the *Letter Identification* task. Given the importance of letter-sound knowledge (Hulme et al., 2012; McArthur et al., 2018; Perfetti, 2007; Schatschneider et al., 2004), it would be beneficial to assess letter-sound knowledge of each letter separately along with letter-name knowledge. The *Letter Identification* task as was administered gives information on children's ability to identify a letter in some way, but not precise information about children's letter-name or letter-sound knowledge alone, which together make up alphabet knowledge–a core early literacy skill (Puranik et al., 2011).

Limitations

A number of limitations need be taken into consideration when interpreting the findings of this thesis. First, the sample size needs to be considered. Compared to other studies that conduct psychometric evaluations of assessment tools, the sample size used in this thesis was small, which limits the study's power. It would be more appropriate to test measurement invariance across gender within each kindergarten level with large samples so that any non-invariances present across kindergarten level do not affect the results. The measurement invariance analyses in particular must be re-run with a larger sample.

The cut-off criterion to retain an item on a task was set to minimal (significant loading at an alpha of .05) in order to stay as close as possible to the standard forms of the tasks when performing item analyses, whereas in test construction studies, good items are identified as those with factor loadings of at least 0.35 (e.g., Clark & Watson, 1995). However, one must be cautious when dropping items based on a cut-off criterion, for theory is an important aspect of test construction. The researcher might choose to keep items with weak loadings if these items were predicted a priori to be strong indicators of a certain factor (Simms, 2008). Hence, had this been a test construction study, or the goal of this thesis been to revise the ELNOT, the recommended cut-off of 0.35 would have been used.

The results of the DIF tests need to be interpreted with caution. The methods employed during these tests (see Appendix A) may be prone to errors. It would be better to perform item analyses, especially the DIF tests, on a software like IRTPRO that does not require known anchors to run the DIF tests (Tay et al., 2015).

Conclusion

As its name indicates, ELNOT is an observation tool aimed at guiding the teaching strategies of teachers and educators according to children's responses on the tasks, through which they can identify children's competencies and confusions (e.g., letters p and q), as well as their strengths and weaknesses. As Clay (2013) notes, effective teaching may benefit more from observational assessments than standardized tests, and provides the analogy of a football game: the quality of the team's play cannot be improved by looking at the final score; instead, the coach must support the players in developing strategic moves to improve their overall performance during the game, which will then lead to a better final score.

In conclusion, ELNOT is a tool with potential. It is most appropriate for assessing children who are having difficulties in early literacy or numeracy skills, for it showed better measurement precision at lower levels of the latent abilities. Depending on the goal of administration (assessing at-risk or gifted children, or monitoring growth), easy or difficult tasks may be skipped for one kindergarten level or the other, or both.

References

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/s15324818ame1403_2

Allington, R. L. (2014). Reading moves: What not to do. *Educational Leadership, 72*(2), 16-21.

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. *Factor Analysis and Measurement in Sociological Research: A Multidimensional Perspective*, 249-280.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14,* 255–259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Aubrey, C., & Godfrey, R. (2003). The development of children's early numeracy through Key Stage 1. *British Educational Research Journal, 29*(6), 821–840. https://doi.org/10.1080%2F0141192032000137321

Aubrey, C., Godfrey, R., & Dahl, S. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal*, 18, 27–46. https://doi.org/10.1007/BF03217428

Aunio, P., Niemivirta, M., Hautamäki, J., Van Luit, J., Shi, J., & Zhang, M. (2006). Young children's number sense in China and Finland. *Scandinavian Journal of Educational Research, 50*(5), 483−502. https://doi.org/10.1080/00313830600953576

Aunio, P., Aubrey, C., Godfrey, R., Pan, Y., & Liu, Y. (2008). Children's early numeracy in England, Finland and People's Republic of China. *International Journal of Early Years Education, 16*(3), 203-221. https://doi.org/10.1080/09669760802343881

Aunio, P., Hautamäki, J., Sajaniemi, N., & Van Luit, J. E. (2009). Early numeracy in low-performing young children. *British Educational Research Journal*, *35*(1), 25-46. https://doi.org/10.1080/01411920802041822

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*(5), 427-435. https://doi.org/10.1016/j.lindif.2010.06.003

Aunio, P., & Räsänen, P. (2016). Core numerical skills for learning mathematics in children aged five to eight years–a working model for educators. *European Early Childhood Education Research Journal*, *24*(5), 684-704. https://doi.org/10.1007/978-3-319-43974-7

Aunola, K., Leskinen, E., Lerkkanen, M. -K., & Nurmi, J. -E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96*(4), 699−713. https://doi.org/10.1037/0022-0663.96.4.699

Baker, F. (2001). *The basics of item response theory.* ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Balfanz, R. (1999). *Why do we teach young children so little mathematics? Some historical considerations*. In J. Copley (Ed.), *Mathematics in the early years* (p. 3–10). NCTM.

Baroody, A. J., Lai, M. L., & Mix, K. S. (2006). The development of young children’s early number and operation sense and its implications for early childhood education. *Handbook of Research on the Education of Young Children*, *2*, 187-221.

Bartelet, D., Vaessen, A., Blomert, L., & Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *Experimental Child Psychology*, *117*, 12-28. https://doi-org/10.1016/j.jecp.2013.08.010

Beach, S. A., & Robinson, R. J. (1992). Gender and grade level differences in the development of concepts about print. *Reading Psychology*, *13*(4), 309–328. https://doi.org/10.1080/027027192130403

Bear, D., & Templeton, S. (1998). Explorations in developmental spelling: Foundations for learning and teaching phonics, spelling, and vocabulary. *The Reading Teacher, 52*, 222–242.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, 67-92.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300. https://doi-org/10.1111/j.2517-6161.1995.tb02031.x

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, *46*(2), 545–551. https://doi.org/10.1037/a0017887

Bialystok, E. (1992). Attentional control in children's metalinguistic performance and measures of field independence. *Developmental Psychology, 28*, 654–664. https://doi-org/10.1037/0012-1649.28.4.654

Blachman, B. A. (2000). *Phonological awareness.* In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research, Vol. 3* (p. 483–502). Lawrence Erlbaum Associates Publishers.

Boardman, M. (2006). The impact of age and gender on prep children's academic achievements. *Australian Journal of Early Childhood, 31*(4), 1−6.

Bowden, S. (2013). Theoretical convergence in assessment of cognition. *Journal of Psychoeducational Assessment*, *31*(2), 148–156. https://doi.org/10.1177/0734282913478035

Bowles, R. P., Pentimonti, J. M., Gerde, H. K., & Montroy, J. J. (2014). Item response analysis of uppercase and lowercase letter name knowledge. *Journal of Psychoeducational Assessment*, *32*(2), 146-156. https://doi.org/10.1177/0734282913490266

Boyer, E. L. (1991). *Ready to learn: A mandate for the nation*. Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Brainerd, C. J. (1979). *The origins of the number concept.* Praeger.

Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods*, *49*(4), 1361-1373. https://doi.org/10.3758/s13428-016-0792-3

Briars, D., & Siegler, R. S. (1984). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology, 20*, 607–618.

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology, 70*, 117–141.

Burrage, M. S., Ponitz, C. C., McCready, E. A., Shah, P., Sims, B. C., Jewkes, A. M., & Morrison, F. J. (2008). Age-and schooling-related effects on executive functions in young children: A natural experiment. *Child Neuropsychology*, *14*(6), 510-524. https://doi.org/10.1080/09297040701756917

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(2), 272-300. https://doi.org/10.1207/s15328007sem1102_8

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software].* Scientific Software International.

Carr, M., & Jessup, D. (1997). Gender differences in first grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology, 89*, 318−328.

Carroll, J. M., Snowling, M. J., Stevenson, J., & Hulme, C. (2003). The development of phonological awareness in preschool children. *Developmental Psychology, 39*(5), 913–923. https://doi-org/10.1037/0012-1649.39.5.913

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., . . . Keating, D. (1996). The Role of Central Conceptual Structures in the Development of Children's Thought. *Monographs of the Society for Research in Child Development, 61*(1-2), 1-295. https://doi.org/10.2307/1166077

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276.

Catts, H. W., McIlraith, A., Bridges, M. S., & Nielsen, D. C. (2017). Viewing a phonological deficit within a multifactorial model of dyslexia. *Reading and Writing*, *30*(3), 613–629. https://doi.org/10.1007/s11145-016-9692-2

Chall, J. S. (1983). *Stages of Reading Development*. McGraw-Hill.

Chall, J. S., Jacobs, V. A., Baldwin, L. E., & Chall, J. S. (2009). *The reading crisis: Why poor children fall behind*. Harvard University Press.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114-140. https://doi-org/10.1177/0013164415584576

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(2), 3-14. https://doi-org/10.1177/073724770503000202

Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler Intelligence Scale for Children–Fifth Edition. *Personality and Individual Differences*, *86*, 1–5. https://doi.org/10.1016/j.paid.2015.05.020

Chen, F., & Chalhoub-Deville, M. (2016). Differential and long-term language impact on math. *Language Testing*, *33*(4), 577-605.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cirino, P. T. (2011). The interrelationships of mathematical precursors in kindergarten. *Journal of Experimental Child Psychology, 108*(4), 713–733. https://doi.org/10.1016/j.jecp.2010.11.004

Clark, M. (1976). *Young fluent readers: What can they teach us?* London: Heinemann.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. https://doi-org/10.1037/1040-3590.7.3.309

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33,* 234–248.

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Heinemann.

Clay, M. M. (1993). *An observation survey of early literacy achievement*. Heinemann.

Clay, M. M. (2013). *An observation survey of early literacy achievement* (3rd ed.). Heinemann.

Comrey, A. L., and Lee, H. B. (1992). *A First Course in Factor Analysis* (2nd ed.). Lawrence Erlbaum Associates.

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), 22–38. https://doi-org/10.1037/0096-3445.137.1.22

Confrey, J., Maloney, A., & Corley, A. (2014). Learning trajectories: a framework for connecting standards with curriculum. *ZDM*, *46*(5), 719–733. https://doi.org/10.1007/s11858-014-0598-7

Crone, D. A., & Whitehurst, G. J. (1999). Age and schooling effects on emergent literacy and early reading skills. *Journal of Educational Psychology, 91*(4), 604-614. http://dx.doi.org/10.1037/0022-0663.91.4.604

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934–945. https://doi-org/10.1037/0012-1649.33.6.934

Dehaene, S. (1997). *The number sense. How the mind creates mathematics.* Penguin Books.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Demie, F. (2001). Ethnic and gender differences in educational achievement and implications for school improvement strategies. *Educational Research, 43*(1), 91−106. https://doi-org/10.1080/00131880110040968

De Smedt, B., Noėl, M. P., Gilmore, C., & Ansari, D. (2013). The relationship between symbolic and non-symbolic numerical magnitude processing skills and the typical and atypical development of mathematics: A review of evidence from brain and behavior. *Trends in Neuroscience and Education, 2*, 48-55. https://doi-org/10.1016/j.tine.2013.06.001

Desoete, A., & Grégoire, J. (2006). Numerical competence in young children and in children with mathematics learning disabilities. *Learning and individual differences*, *16*(4), 351-367. https://doi-org/10.1016/j.lindif.2006.12.006

Desoete, A., Stock, P., Schepens, A., Baeyens, D., & Roeyers, H. (2009). Classification, seriation, and counting in grades 1, 2, and 3 as two-year longitudinal predictors for low achieving in numerical facility and arithmetical achievement? *Journal of Psychoeducational Assessment, 27*, 252-264. https://doi-org/10.1177/0734282908330588

Dickinson, D. K., & Snow, C. E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly, 2*, 1–25. https://doi-org/10.1016/0885-2006(87)90010-X

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149. https://doi-org/10.1177/0748175610373459

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428.

Durkin, D. (1966). *Children who read early.* Teachers College Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists (Multivariate Applications Books Series).* Lawrence Erlbaum Associates Publishers.

Farrington, A. L., & Lonigan, C. J. (2015). Examining the measurement precision and invariance of the revised get ready to read! *Journal of learning disabilities*, *48*(3), 227-238. https://doi-org/10.1177/0022219413495568

Fayol, M., Barrouillet, P., & Marinthe, C. (1998). Predicting arithmetical achievement from neuro-psychological performance: A longitudinal study. *Cognition*, *68*(2), B63-B70. https://doi.org/10.1016/S0010-0277(98)00046-8

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 6−11. https://doi-org/10.3102/0013189X027005006

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3–17. https://doi-org/10.1037/0022-0663.88.1.3

Frye, D., Braisby, N., Lowe, J., Maroudas, C., Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child Development, 60*(5), 1158-1171. https://doi-org/10.2307/1130790

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., Bryant, J. D., & Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental psychology,* 46(6), 1731-1746. https://doi-org/10.1037/a0020662

Fuson, K. C. (1988). *Children's counting and concepts of number*. Springer-Verlag.

Gallagher, A., Frith, U., & Snowling, M.J. (2000). Precursors of literacy delay among children at genetic risk of dyslexia. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 41*(2), 203–213. https://doi.org/10.1017/S0021963099005284

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15. https://doi.org/10.1177/00222194040370010201

Gelman R, & Gallistel C. R. (1978). *The child's concept of number*. Harvard University Press.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Genty, J. R., & Gilet, J. W. (1993). *Teaching kids to spell*. Heinemann.

Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education, 33*(1), 18-28.

Ginsberg, H. (2006). Mathematical play and playful mathematics: A guide for early education. In D. Singer, R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Play=learning: How play motivates and enhances children's cognitive and social-emotional growth* (p. 145–165). Oxford University Press.

Ginsburg, H. P., and Baroody, A. J. (2003). *TEMA: Test of Early Mathematics Ability* (3rd ed.) PRO-ED.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics Education for Young Children: What It Is and How to Promote It. *Social Policy Report* by *Society for Research in Child Development, 22*(1), 3-11.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). University of Oregon, Institute for the Development of Educational Achievement.

Gorard, S., Rees, G., & Salisbury, J. (2001). Investigating the patterns of differential attainment of boys and girls at school. *British Educational Research Journal, 27*(2), 125−139. https://doi-org/10.1080/01411920120037090

Göbel, S. M., Watson, S. E., Lervåg, A., & Hulme, C. (2014). Children's arithmetic development: It is number knowledge, not the approximate number sense, that counts. *Psychological science*, *25*(3), 789-798. http://dx.doi.org/10.1177/0956797613516471

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology, 94*(4), 669–686. https://doi-org/10.1037/0022-0663.94.4.669

Haden, C. A., Reese, E., & Fivush, R. (1996). Mothers' extratextual comments during storybook reading: Stylistic differences over time and across texts. *Discourse Processes, 21*(2), 135–169. https://doi-org/10.1080/01638539609544953

Hammill, D.D. (2004). What we know about correlates of reading. *Exceptional Children, 70*(4), 453–468.

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology*, *79*(2), 192-227. https://doi.org/10.1006/jecp.2000.2586

Hecht, S. A., & Close, L. (2002). Emergent literacy skills and training time uniquely predict variability in responses to phonemic awareness training in disadvantaged kindergartners. *Journal of Experimental Child Psychology*, *82*(2), 93–115. https://doi.org/10.1016/S0022-0965(02)00001-2

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365-388. https://doi-org/10.1177/0013164414548576

Holland, P. W., & Wainer, H. (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Erlbaum.

Hornburg, C. B., Schmitt, S. A., & Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *Journal of Experimental Child Psychology*, *176*, 84–100. https://doi.org/10.1016/j.jecp.2018.07.005

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, F. L., & Invernizzi, M. A. (2012). The Association of Kindergarten Entry Age with Early Literacy Outcomes. *The Journal of Educational Research*, *105*(6), 431–441. https://doi.org/10.1080/00220671.2012.658456

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read. *Psychological Science, 23*(6), 572-577. https://doi.org/10.1177/0956797611435921

Hunting, R. (2010). Little people, big play, and big mathematical ideas. In *Shaping the future of mathematics education: Proceedings of the 33rd annual conference of the Mathematics Education Research Group of Australasia* (p. 727-730).

Hurst, M., Anderson, U., & Cordes, S. (2017). Mapping among number words, numerals, and nonsymbolic quantities in preschoolers. *Journal of Cognition and Development*, *18*(1), 41–62. https://doi.org/10.1080/15248372.2016.1228653

Jiménez Lira, C., Carver, M., Douglas, H., & LeFevre, J.-A. (2017). The integration of symbolic and non-symbolic representations of exact quantity in preschool children. *Cognition*, *166*, 382–397. https://doi.org/10.1016/j.cognition.2017.05.033

Jordan, N. C., Kaplan, D., Olah, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*, 153–175. https://doi-org/10.1111/j.1467-8624.2006.00862.x

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36–46. https://doi.org/10.1111/j.1540-5826.2007.00229.x

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), 850–867. https://doi.org/10.1037/a0014939

Kavkler, M., Tancig, S., & Magajna, L. (2003). Follow-up study of children with very low and very high mathematical competence in preschool years. *Paper presentation in EARLI conference,* Padova Italy.

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, *19*(6), 513-526. https://doi.org/10.1016/j.learninstruc.2008.10.002

Klibanoff, R. S., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. V. (2006). Preschool children's mathematical knowledge: The effect of teacher" math talk.". *Developmental Psychology*, *42*(1), 59. https://doi-org/10.1037/0012-1649.42.1.59

Kline, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781483385693

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22-56. https://doi-org/10.1177/0013164414529792

Koponen, T., Aunola., K., Ahonen, T., & Nurmi., J.-E. (2007). Cognitive predictors of single-digit and procedural calculation and their covariation with reading skill. *Journal of Experimental Child Psychology, 97*, 220 –241. https://doi-org/10.1016/j.jecp.2007.03.001

Kurdek, L. A., & Sinclair, R. J. (2001). Predicting reading and mathematics achievement in fourth-grade children from kindergarten readiness scores. *Journal of Educational Psychology, 93*(3), 451−455. https://doi-org/10.1037/0022-0663.93.3.451

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. https://doi-org/10.1016/j.cogpsych.2005.07.002

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. https://doi-org/10.1016/j.cognition.2006.10.005

LeFevre, J. A., Smith-Chant B. L., & Fast, L. (2006). What counts as knowing? The development of conceptual and procedural knowledge of counting from kindergarten through Grade 2. *Journal of Experimental Child Psychology, 93*(4)*,* 285-303. https://doi.org/10.1016/j.jecp.2005.11.002

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. https://doi.org/10.1111/j.1467-8624.2010.01508.x

Lembke, E., & Foegen, A. (2005). *Identifying indicators of early mathematics proficiency in Kindergarten and Grade 1*. Technical Report 6). Minneapolis, MN: University of Minnesota, College of Education and Human Development, Research Institute on Progress Monitoring.

Lembke, E., & Foegen, A. (2009). Identifying Early Numeracy Indicators for Kindergarten and First-Grade Students. *Learning Disabilities Research & Practice*, *24*(1), 12–20. https://doi.org/10.1111/j.1540-5826.2008.01273.x

Limbrick, L., Wheldall, K., & Madelaine, A. (2011). Why do more boys than girls have a reading disability? A review of the evidence. *Australasian Journal of Special Education, 35*(1), 1-24. https://doi.org/10.1375/ajse.35.1.1

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*(5), 596–613. https://doi-org/10.1037/0012-1649.36.5.596

Lonigan, C. J., & Wilson, S. B. (2008). Report on the revised Get Ready to Read! screening tool: Psychometrics and normative information (Tech. Rep.). *National Center on Learning Disabilities*. New York, NY.

Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. *Developing early literacy: Report of the national early literacy panel*, 55-106.

Lonigan, C. J., Allan, N. P., & Lerner, M. D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. *Psychology in the Schools*, *48*(5), 488-501. https://doi-org/10.1002/pits.20569

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*(4), 989-1020. https://doi-org/10.1177/001316446802800401

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model. *Psychometrika*, *35*(1), 43-50. https://doi-org/10.1177/001316447303300201

Lourenco, S. F., & Bonny, J. W. (2017). Representations of numerical and non-numerical magnitude both contribute to mathematical competence in children. *Developmental science*, *20*(4), 10.1111/desc.12418. https://doi-org/10.1111/desc.12418

Lundberg, I., Larsman, P., & Strid, A. (2012). Development of phonological awareness during the preschool year: The influence of gender and socio-economic status. *Reading and writing*, *25*(2), 305-320. https://doi-org/10.1007/s11145-010-9269-4

Maller, S. J. (2001). Differential Item Functioning in the Wisc-III: Item Parameters for Boys and Girls in the National Standardization Sample. *Educational and Psychological Measurement*, *61*(5), 793–817. https://doi.org/10.1177/00131640121971527

Matejko, A. A., & Ansari, D. (2016). Trajectories of symbolic and nonsymbolic magnitude processing in the first year of formal schooling. PLoS ONE, 11(3), e0149863. http://dx.doi.org/10.1371/journal.pone.0149863.

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of educational psychology*, *101*(3), 689. https://doi-org/10.1037/a0014240

McArthur, T., Lam-McArthur, J., & Fontaine, L. (2018). *Oxford companion to the English language*. Oxford University Press.

McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quarterly, 45*(2), 285–308.

McGencey, S. (2011). The time is now: Creating opportunities for young children to succeed. *National Civic Review*, *100*(4), 56-58. https://doi.org/10.1002/ncr.20088

McGuinness, D., McGuinness, C., & Donohue, J. (1995). Phonological training and the alphabet principle: Evidence for reciprocal causality. *Reading Research Quarterly*, *30*(4), 830-852. https://doi.org/10.2307/748200

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of life research*, *4*(4), 293-307. https://doi-org/10.1007/bf01593882

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016–1031. https://doi-org/10.1037/a0027934

Merkley, R., & Ansari, D. (2016). Why numerical symbols count in the development of mathematical skills: Evidence from brain and behavior. *Current Opinion in Behavioral Sciences, 10*, 14–20. https://doi.org/10.1016/j.cobeha.2016.04.006

Missall, K. N., Mercer, S. H., Martínez, R. S., & Casebeer, D. (2012). Concurrent and longitudinal patterns and trends in performance on early numeracy curriculum-based measures in kindergarten through third grade. *Assessment for Effective Intervention*, *37*(2), 95-106. https://doi-org/10.1177/1534508411430322

Mix, K. S. (2002). The construction of number concepts. *Cognitive Development*, *17*(3-4), 1345–1363. https://doi.org/10.1016/S0885-2014(02)00123-5

Mix, K. S. (2009). How Spencer made number: First uses of the number words. *Journal of Experimental Child Psychology*, *102*(4), 427–444. https://doi.org/10.1016/j.jecp.2008.11.003

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267-296. http://dx.doi.org/10.1037/a0021890

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. https://doi.org/10.1016/j.jecp.2016.05.003

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Kindergarten predictors of recurring externalizing and internalizing psychopathology in the third and fifth grades. *Journal of Emotional and Behavioral Disorders*, *17*(2), 67-79. https://doi-org/10.1177/1063426608324724

Mou, Y., Li, Y., Hoard, M. K., Nugent, L. D., Chu, F. W., Rouder, J. N., & Geary, D. C. (2016). Developmental foundations of children's fraction magnitude knowledge. *Cognitive development*, *39*, 141-153. https://doi.org/10.1016/j.cogdev.2016.05.002

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520. https://doi-org/10.1038/2151519a0

Muldoon, K., Towse, J., Simms, V., Perra, O., & Menzies, V. (2013). A longitudinal analysis of estimation, counting skills, and mathematical ability across the first school year. *Developmental psychology*, *49*(2), 250. https://doi-org/10.1037/a0028240

Muthén, L. K., & Muthén, B. O. (2012). Mplus Version 7 user's guide. *Los Angeles, CA: Muthén & Muthén*.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. National Institute for Literacy.

Neumann, M. M. (2018). Using tablets and apps to enhance emergent literacy skills in young children. *Early Childhood Research Quarterly, 42*(1), 239-246. https://doi.org/10.1016/j.ecresq.2017.10.006

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and nonsymbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PLoS ONE*, *8*(7), e67918. https://doi.org/10.1371/journal.pone.0067918

Nunes, T. & Bryant, P. (1996). *Children doing mathematics*. Blackwell.

Nunnally, J. C., & Bernstein, I. H., (1994). *Psychometric theory*. McGraw-Hill.

O'Connor, R. E., & Jenkins, J. R. (1999). Improving the generalization of sound/symbol knowledge: Teaching spelling to kindergarten children with disabilities. *The Journal of Special Education, 29*(3), 255–275. https://doi-org/10.1177/002246699502900301

Olson, L. A., Evans, J. R., & Keckler, W. T. (2006). Precocious readers: Past, present, and future. *Journal for the Education of the Gifted*, *30*(2), 205-235. https://doi-org/10.4219/jeg-2006-260

Olson, J., Foegen, A., & Singamaneni, S. (2009). *Iowa Early Numeracy Indicator screening data: 2007-2008*. Technical Report 23). University of Minnesota, College of Education and Human Development, Research Institute on Progress Monitoring.

Ontario (2008). *The Ontario curriculum grades 1-8: Language*. Toronto: The Ministry.

Ortiz, M., Folsom, J. S., Al Otaiba, S., Greulich, L., Thomas-Tate, S., & Connor, C. M. (2012). The componential model of reading: Predicting first grade reading performance of culturally diverse students from ecological, psychological, and cognitive factors assessed at kindergarten entry. *Journal of Learning Disabilities*, *45*(5), 406-417. https://doi-org/10.1177/0022219411431242

Ouellette, G., & Sénéchal, M. (2008). Pathways to literacy: A study of invented spelling and its role in learning to read. *Child Development, 79*(4), 899–913. https://doi.org/10.1111/j.1467-8624.2008.01166.x

Ouellette, G., & Sénéchal, M. (2016). Invented spelling in kindergarten as a predictor of reading and spelling in grade 1: A new pathway to literacy, or just the same road, less known? *Developmental Psychology, 53*(1), 77-88. https://doi-org/10.1037/dev0000179

Ozernov-Palchik, O., & Gaab, N. (2016). Tackling the 'dyslexia paradox': reading brain and behavior for early markers of developmental dyslexia. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(2), 156–176. https://doi-org/10.1002/wcs.1383

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading research quarterly*, *40*(2), 184-202. https://doi-org/10.1598/RRQ.40.2.3

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* London: National Research and Development Centre for Adult Literacy and Numeracy.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly, 45*(1), https://doi-org/10.1598/RRQ.45.1.2

Piasta, S. B., Purpura, D. J., & Wagner, R. K. (2010). Fostering alphabet knowledge development: A comparison of two instructional approaches. *Reading and Writing*, *23*(6), 607-626. https://dx-doi-org/10.1007%2Fs11145-009-9174-x

Pinto, G., Bigozzi, L., Tarchi, C., Vezzani, C., & Accorti Gamannossi, B. (2016). Predicting reading, spelling, and mathematical skills: A longitudinal study from kindergarten through first grade. *Psychological reports*, *118*(2), 413-440. https://doi-org/10.1177/0033294116633357

Puranik, C. S., Lonigan, C. J., & Kim, Y. S. (2011). Contributions of emergent literacy skills to name writing, letter writing, and spelling in preschool children. *Early Childhood Research Quarterly*, *26*(4), 465–474. https://doi.org/10.1016/j.ecresq.2011.03.002

Purcell-Gates, V. (1996). Stories, coupons, and the TV guide: Relationships between home literacy and emergent literacy knowledge. *Reading Research Quarterly, 31*, 406– 428. https://doi-org/10.1598/RRQ.31.4.4

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, *110*(4), 647–658. https://doi.org/10.1016/j.jecp.2011.07.004

Purpura, D. J., & Lonigan, C. J. (2015). Early Numeracy Assessment: The Development of the Preschool Early Numeracy Scales. *Early Education and Development*, *26*(2), 286–313. https://doi.org/10.1080/10409289.2015.991084

Purpura, D. J., & Napoli, A. R. (2015). Early Numeracy and Literacy: Untangling the Relation Between Specific Components. *Mathematical Thinking and Learning*, *17*(2–3), 197–218. https://doi.org/10.1080/10986065.2015.1016817

Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, *36*, 259-268. https://doi-org/10.1016/j.ecresq.2015.12.020

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Ransdell, S., & Hecht, S. (2003). Time and resource limits on working memory: Cross-age consistency in counting span performance. *Journal of Experimental Child Psychology*, *86*(4), 303-313. https://doi.org/10.1016/j.jecp.2003.08.002

Read, C. (1971). Preschool children's knowledge of English phonology. *Harvard Educational Review, 41*(1), 1-34. https://doi.org/10.17763/haer.41.1.91367v0h80051573

Resnick, L. B. (1989). Developing mathematical knowledge. *American Psychologist, 44*(2), 162–169. https://doi-org/10.1037/0003-066X.44.2.162

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science*, *24*(7), 1301-1308. https://doi-org/10.1177/0956797612466268

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* Routledge. https://doi.org/10.4324/9780203883785

Sarnecka, B. W., & Gelman, S. A. (2004). Six does not just mean a lot: preschoolers see number words as specific. *Cognition*, *92*(3), 329–352. https://doi-org/10.1016/j.cognition.2003.10.001

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. https://doi-org/10.1016/j.cogpsych.2006.09.001

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674. https://doi.org/10.1016/j.cognition.2008.05.007

Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge during early childhood. *Journal of Experimental Child Psychology*, *103*(3), 325-337. https://doi.org/10.1016/j.jecp.2009.02.007

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. https://doi.org/10.1111/cogs.12043

Satorra, A., & Bentler, P. M. (1999). *A scaled difference chi-square test statistic for moment structure analysis* (UCLA Statistics Series No. 260). Retrieved from http://preprints.stat.ucla.edu

Scarborough, H.S. (1998). *Early identification of children at risk for reading disabilities.* In B.K. Shapiro, P.J. Accardo, & A.J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (p. 75–120). York Press.

Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology, 6*, 357–379. https://doi.org/10.1016/0010-0285(74)90017-6

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten Prediction of Reading Skills: A Longitudinal Comparative Analysis. *Journal of Educational Psychology*, *96*(2), 265–282. https://doi.org/10.1037/0022-0663.96.2.265

Schmitt, N. & Kuljanin,G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schneider, W., Ennemoser, M., Roth, E., & Küspert, P. (1999). Kindergarten prevention of dyslexia: Does training in phonological awareness work for everybody? *Journal of Learning Disabilities*, *32*(5), 429–436. https://doi-org/10.1177/002221949903200508

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, *20*(3), e12372. https://doi.org/10.1111/desc.12372

Sénéchal, M., LeFevre, J. A., Smith-Chant, B. L., & Colton, K. V. (2001). On refining theoretical models of emergent literacy: The role of empirical evidence. *Journal of school psychology*, *39*(5), 439-460. https://doi.org/10.1016/S0022-4405(01)00081-4

Sénéchal, M., Ouellette, G., Pagan, S., & Lever, R. (2012). The role of invented spelling on learning to read in low-phoneme awareness kindergartners: A randomized-control-trial study. *Reading and Writing: An Interdisciplinary Journal, 25*(4), 917–934. https://doi-org/10.1007/s11145-011-9310-2

Sénéchal, M. (2017). Testing a nested skills model of the relations among invented spelling, accurate spelling, and word reading, from kindergarten to grade 1. *Early Child Development and Care, 187*(3-4), 358-370. https://doi-org/10.1080/03004430.2016.1205044

Shusterman, A., Gibson, D., & Finder, B. (2010). *Acquiring first number words: The developmental trajectory of children's meanings for "Two".* In Franich K, Iserman KM, Keil LL (Eds), *Proceedings of the 34th Annual Boston University Conference on Language Development [BUCLD 34]* (p. 375–384). Cascadilla Press.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2*(1), 414–433. https://doi-org/10.1111/j.1751-9004.2007.00044.x

Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly*, *32*(4), 435–448. https://doi.org/10.1037/spq0000197

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443-461. https://doi.org/10.1007/s11205-010-9682-8

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. https://doi.org/10.1016/j.cognition.2013.05.011

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.). (1998). *Preventing Reading Difficulties in Young Children*. National Academy Press.

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*(3), 353-363. https://doi-org/10.1037/0022-0663.84.3.353

Stainthorp, R., & Hughes, D. (2004). An illustrative case study of precocious reading ability. *Gifted Child Quarterly, 48*, 107-120. https://doi.org/10.1177/001698620404800204

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental psychology*, *38*(6), 934-947. https://doi-org/10.1037/0012-1649.38.6.934

Strand, S. (1997). Pupil progress during Key Mage 1: A value-added analysis or school effects. *British Educational Research Journal, 23*(4), 471-487. https://doi-org/10.1080/0141192970230406

Strand, S. (1999). Ethnic group, sex, and economic disadvantage: associations with pupils' educational progress from Baseline to the end of Key Stage 1. *British Educational Research Journal, 25(*2), 179-202. https://doi-org/10.1080/0141192990250204

Sulzby, E., Barnhart, J., & Hieshima, J. (1989). *Forms of writing and rereading from writing: A preliminary report*. In J. Mason (Ed.), *Reading and writing connections*. Needham Heights, MA: Allyn and Bacon.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67–113). Lawrence Erlbaum Associates, Inc.

Thomas, L., Warren, E., & DeVries, E. (2011). Play-based learning and intentional teaching in early childhood contexts. *Australasian Journal of Early Childhood*, *36*(4), 69-75. https://doi-org/10.1177/183693911103600410

Tizard, B., Blatchford, P., Burke, J., Farquhar, C., & Plewis, I. (1988). *Young children at school in the inner city*. Hove, UK: Lawrence Erlbaum.

Torgesen, J. K., & Burgess, S. R. (1998). *Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies.* In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (p. 161–188). Lawrence Erlbaum Associates Publishers.

Torppa, M., Poikkeus, A.-M., Laakso, M.-L., Eklund, K., & Lyytinen, H. (2006). Predicting delayed letter knowledge development and its relation to Grade 1 reading achievement among children with and without familial risk for dyslexia. *Developmental Psychology,* *42*(6), 1128–1142. https://doi-org/10.1037/0012-1649.42.6.1128

Treiman, R. (1998). *Why spelling? The benefits of incorporating spelling into beginning reading instruction.* In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (p. 289–313). Lawrence Erlbaum Associates Publishers.

Treiman, R., & Broderick, V. (1998). What's in a name: Children's knowledge about the letters in their own name. *Journal of Experimental Child Psychology, 70*(2), 97–116. https://doi-org/10.1006/jecp.1998.2448

Tunmer, W. E., Herriman, M. L., & Nesdale, A. R. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23*(2), 134–158. https://doi-org/10.2307/747799

Vanbinst, K., Ceulemans, E., Peters, L., Ghesquière, P., & De Smedt, B. (2018). Developmental trajectories of children's symbolic numerical magnitude processing skills and associated cognitive competencies. *Journal of Experimental Child Psychology, 166*, 232-250. https://doi.org/10.1016/j.jecp.2017.08.008

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70. https://doi-org/10.1177/109442810031002

van Marle, K., Chu, F. W., Li, Y., & Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Developmental Science*, *17*(4), 492–505. https://doi.org/10.1111/desc.12143

Wagner, S., & Walters, J. A. (1982). *A longitudinal analysis of early number concepts: From numbers to number*. In G. Forman (Ed.), *Action and thought* (p. 137-161). New York: Academic Press.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30*(1), 73–87. https://doi-org/10.1037/0012-1649.30.1.73

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of educational psychology*, *102*(1), 43-53. https://doi-org/10.1037/a0016738

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition*. San Antonio, TX: Psychological Corporation.

Whitehurst, G. J., & Lonigan, C. J. (2001). Get Ready to Read! screening tool. *New York, NY: National Center for Learning Disabilities*.

Woolley, G. (2011). *Reading Comprehension*. Springer Netherlands. https://doi.org/10.1007/978-94-007-1174-7

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. https://doi.org/10.1016/0010-0285(92)90008-P

Xu, C. (2018). *Knowledge of Number Integration* [Text, Carleton University]. https://curve.carleton.ca/943592de-85cb-403b-a957-8dbda79ec3d1

Xenidou-Dervou, I., Molenaar, D., Ansari, D., van der Schoot, M., & van Lieshout, E. (2017). Nonsymbolic and symbolic magnitude comparison skills as longitudinal predictors of mathematical achievement. *Learning and Instruction, 50*, 1-13. https://doi.org/10.1016/j.learninstruc.2016.11.001

APPENDICES

Appendix A: Further Explanation of the Differential Item Functioning (DIF) Tests

Item Analyses

Differential Item Functioning (DIF) Tests

A significant DIF indicates that an item-level bias exists within a certain subgroup comparison (Holland & Wainer, 1993). That is, the estimated item parameters in the focal group (e.g., junior kindergarten) significantly differ from that of the reference group (e.g., senior kindergarten), even after accounting for the overall differences in the construct being measured. As is typically done in this type of analysis, the reference group is designated to be the group with a larger sample, namely females and senior kindergarteners in this thesis, and the focal group is the group with a smaller sample (Tay et al., 2015). Importantly, the assignment of reference and focal groups does not affect the computation of DIF.

The IRT-LR test uses a chi-square distribution to compare two nested IRT models: a full model and a restricted model (Desjardins & Bulut, 2018). In the full model, all item parameters are set free, whereas, in the restricted model, the discrimination and difficulty parameters are constrained between subgroups. Items with statistically significant likelihood ratio statistic exhibit DIF, and therefore require follow-up tests to identify the type of DIF. If one subgroup has a constant advantage over the other on an item across the latent trait continuum, the item has uniform DIF (Desjardins & Bulut, 2018). On the other hand, if the magnitude and/or the direction of this advantage changes along the latent trait continuum, the item has nonuniform DIF. Therefore, items that have a different meaning for children in any of the subgroups would impede the accuracy of the tasks and confound the comparisons between these subgroups (Tay et al., 2015). To identify the type of DIF that an item exhibits, I conducted follow-up DIF tests by constraining only the discrimination parameter (a) between subgroups instead of both item parameters (Desjardins & Bulut, 2018). Items with significantly different discrimination parameters have nonuniform DIF.

To control for Type-I error in each DIF analyses, I applied the Benjamini and Hochberg (B-H) procedure to identify significant DIF (Benjamini & Hochberg, 1995). In this procedure, the p values of all items are ordered from smallest to largest, and the smallest p value is compared to the conventional alpha level of .05 divided by the total number of items (n). Then, in a linear fashion, the next smallest p value is compared to .05/n-1, the next to .05/n-2 until the item with the largest p value is compared to the conventional alpha level of .05.

Anchor Selection. A requirement for carrying out the IRT-LR test for DIF detection is to equate the tasks between groups (Kolen & Brennan, 2004). That is, differences in groups' latent means and variances may produce unwarranted DIF, and this can be prevented by selecting a set of anchor items that are assumed to have invariant item parameters across groups. After setting the parameters of anchor items to be constrained across groups, the mean and variance of the latent trait can then be freely estimated in all but the reference group (senior kindergarten). These two estimation properties help fix the metric of the groups (Chalmers et al., 2016).

While there are many different anchor classes and selection strategies (Kopf et al., 2015), the IRT-LR test uses known anchors – a limitation in the case of this thesis and many studies as "it is seldom known which items are invariant prior to DIF testing" (Tay et al., 2015). As an alternative to the default known anchor approach, I used the iterative forward anchor class method with the all-other anchor selection strategy. This technique prescribes choosing the items with the smallest test statistic value, or equivalently, the largest p values as the anchor items (Kopf et al., 2015). Also, it is recommended to select 25% of total number of items as anchors for maximum power of correctly detecting items with true DIF (Meade & Wright, 2012). Due to the limitation of the IRT-LR test and the mirt package, following the exact steps described in Kopf et al. (2015) to employ the iterative forward anchor class method with the all-other anchor selection strategy was not possible. Hence, I used a strategy that is very close to it. First, I ran a multiple-group DIF test without constraining any items. Next, I selected the item with the largest *p* value to be the first anchor in the anchor set and fixed its parameters to be equal across groups. This anchor selection then allowed to also command the test to freely estimate latent means and variances in the focal group. I continued this iterative forward technique until 25% of the total number of items were selected as anchors in each of the tasks.

Table A1

Anchor Set of the Print Concepts Task Used for Differential Item Functioning by Kindergarten Level Test (Fall)

	Fall			Spring		
Anchor	Item	Concept	Description	Item	Concept	Description
1	11	Orientation	Inversion of picture	11	Orientation	Inversion of picture
2	13	Orientation	Line sequence	15	Letter	Small letter
3	3	Book	Title	13	Orientation	Line sequence
4	6	Directionality	Left to right	14	Letter	Capital letter
5	4	Reading	Print carries the message	9	Word	First word

Note. The total number of items on the *Print Concepts* task was 20. Hence, five items were selected as anchors.

Table A2
Anchor Set of the Letter Identification Task Used for Differential Item Functioning by Kindergarten Level Tests

	Fall	Spring
Anchor	Letter	Letter
1	q	C
2	C	E
3	y	J
4	j	c
5	F	j
6	U	K
7	k	u
8	a	V
9	f	N
10	P	e
11	z	v
12	L	k
13	u	F
14	s	z

Note. The total number of items on the *Letter Identification* task was 54. Hence, 14 items were selected as anchors.

Table A3
Anchor Set of the Number Recognition Task Used for Differential Item Functioning by Kindergarten Level Test

Anchor	Item	Row	Stimuli
1	54	14	29
2	33	9	20
3	43	11	27
4	56	14	45
5	3	1	2
6	20	5	11
7	44	11	18
8	10	3	6
9	40	10	2
10	7	2	8
11	42	11	6
12	49	13	90
13	1	1	10

Note. The total number of items on the *Number Recognition* task was 52. Hence, 13 items were selected as anchors.

Appendix B: Further Explanation of the Measurement Invariance Analyses Measurement Invariance Analyses

I tested the measurement invariance of the ELNOT using multiple-group CFA in M*Plus* (Muthén & Muthén, 2012) using the codes published by Dimitrov (2010). The significance of the chi-square tests of model fit were evaluated based on the experiment-wise error rate of .007. As was mentioned under the item analysis section, it is a general rule of thumb to have at least 300 participants when performing factor analysis (Comrey & Lee, 1992). Therefore, the multiple-group CFAs testing for each of the variables (gender, kindergarten level and time) were conducted with junior and senior kindergarteners together.

Partial invariance can be tested by freeing the intercept, for example, of some of the items (tasks in this thesis) based on their modification index (MI). The MIs are provided in the M*Plus* output. The MI indicates that the "expected drop in the model's chi-square value if this parameter is freely estimated," and is statistically significant if greater than 3.84 (Dimitrov, 2010). It is recommended to free one item parameter (e.g., intercept) at a time when there are several parameters with statistically significant MIs, starting from the one with the largest MI. However, freeing up to 20% of the parameters is considered acceptable for a partial invariance. In sum, complete or partial invariance needs to be established at the end of each step to proceed to the next. Otherwise, the invariance tests must be terminated.

Preliminary Analyses

Before proceeding with the measurement invariance tests, the data need to be assessed for the assumption of multivariate normality, that is, the distributions of each variable (task) and all linear combinations of variables are normal (Tabachnick & Fidell, 2007). This assumption is violated if the individual variables are not normally distributed, without needing to test the second part of the assumption. The tasks for which the only available data were total scores were checked for floor or ceiling effects, defined as 15-20% of children perform at the lowest or highest end, respectively (McHorney & Tarlov, 1995).

The result of the normality assessment impacts the choice of estimation technique used for the multiple-group CFAs. The default estimator in M*Plus* is maximum likelihood (ML) for datasets where the multivariate assumption is met. If the assumption is violated, the maximum likelihood estimation technique with robust standard errors and a mean-adjusted chi-square test (MLM) can be used. The MLM estimator is also referred to as the Satorra-Bentler scale chi-square test (SBSχ^2; Satorra & Bentler, 1999).

Appendix C: Comparison of Items in *Letter Identification* and *Number Recognition* with Overlapping Information

Table C1
Recommendation to Drop Overlapping Items from the Letter Identification Task Based on Item Parameter Estimates in the Fall for Junior and Senior Kindergarten

Comparison Within	Letter	*a*	*b*	Recommended to Retain?
Junior Kindergarten	W	2.10	-0.03	No
	m	3.22	-0.03	Yes
	P	2.19	0.01	No
	w	2.26	0.01	No
	R	2.83	0.01	Yes
	y	2.56	0.36	No
	r	2.60	0.36	No
	G	2.67	0.36	Yes
	i	1.85	0.57	No
	V	2.78	0.57	Yes
Senior Kindergarten	E	3.13	-1.19	No
	k	3.15	-1.19	Yes
	J	2.41	-0.80	No
	Y	2.77	-0.80	No
	p	2.79	-0.80	Yes
	h	2.21	-0.76	No
	U	2.93	-0.76	Yes
	G	4.07	-0.75	No
	f	4.96	-0.75	Yes
	j	2.16	-0.45	No
	v	5.50	-0.45	Yes

Table C2
Recommendation to Drop Overlapping Items from the Letter Identification Task Based on Item Parameter Estimates in the Spring for Junior and Senior Kindergarten

Comparison Within	Letter	*a*	*b*	Recommended to Retain?
Junior Kindergarten	w	2.28	-0.83	No
	P	3.65	-0.83	Yes
	k	3.22	-0.79	No
	a	4.03	-0.79	Yes
	p	2.28	-0.73	No
	e	2.63	-0.73	Yes
	D	3.01	-0.65	No
	N	3.38	-0.65	Yes
	J	1.77	-0.50	No
	T	2.46	-0.50	Yes
	f	2.92	-0.42	No
	G	2.21	-0.42	Yes
	y	2.91	-0.30	No
	U	4.64	-0.30	Yes
Senior Kindergarten	O	26.38	-2.50	No
	A	26.38	-2.50	No
	o	26.38	-2.50	Yes
	C	2.51	-1.79	No
	M	3.12	-1.79	Yes
	I	1.55	-1.74	No
	P	4.65	-1.74	Yes
	y	1.76	-1.42	No
	E	3.65	-1.42	Yes
	Q	2.78	-1.36	No
	a	2.79	-1.36	Yes

Table C3

Recommendation to Drop Overlapping Items from the Number Recognition Task Based on Item Parameter Estimates for Junior and Senior Kindergarten

Comparison Within	Stimuli	Row	Item	*a*	*b*	Recommended to Retain?
Junior Kindergarten	6	1	4	3.46	-0.81	No
	2	10	40	6.76	-0.81	Yes
	6	8	30	3.53	-0.68	No
	5	4	16	6.51	-0.68	Yes
	43	5	19	2.91	-0.20	No
	11	5	20	3.57	-0.20	Yes
	17	2	6	1.92	0.03	No
	16	12	45	1.97	0.03	Yes
	38	13	50	1.45	0.05	No
	18	8	29	1.96	0.05	Yes
	12	13	51	1.07	0.65	No
	30	14	53	1.37	0.65	Yes
Senior Kindergarten	10	1	1	3.98	-2.18	No
	0	12	48	4.06	-2.18	Yes
	2	1	3	24.25	-2.05	No
	1	8	31	25.11	-2.05	Yes
	13	8	32	2.06	-1.53	No
	9	12	47	3.36	-1.53	Yes
	17	2	6	2.47	-1.26	No
	11	5	20	3.98	-1.26	Yes
	38	13	50	1.86	-1.11	No
	15	12	46	4.17	-1.11	Yes