The Not So Simple View of Reading in Down Syndrome

By

Alison Hessling

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing and Speech Sciences

May 8, 2020

Nashville, Tennessee

Committee Members:

C. Melanie Schuele, Ph.D.
Jim Bodfish, Ph.D.
Stephen Camarata, Ph.D.
Christopher Lemons, Ph.D.
Tiffany Woynaroski, Ph.D.

To mom and dad

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of a US Department of Education Preparation of Leadership Personnel grant (H325D140087; PI: Schuele), and in part by Vanderbilt CTSA award No. UL1TR000445 from the National Center for Advancing Translational Sciences and a Russell G. Hamilton Graduate Leadership Institute Dissertation Enhancement Grant. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the US Department of Education, the National Center for Advancing Translational Sciences, or the Russel G. Hamilton Graduate Leadership Institute.

First and foremost, I would like to thank my participants and their families for your kindness and willingness to participate, making this project both possible and fun. I enjoyed meeting and working with you and highly value your commitment to advancing the research for individuals with Down syndrome.

I am grateful to my dissertation committee and the Child Language and Literacy Lab for their contributions to this study and their continued support and guidance. I also would like to thank all of my mentors for investing their time and energy to support my career development by both challenging and encouraging me to remain true to my goal of improving outcomes for children with communication disorders. I could not have accomplished any of this without you. Thank you for everything.

Finally, thank you to my family and friends for their unwavering support and encouragement from near and far. I am forever grateful for lessons of hard work and determination that my parents instilled in me from a young age—for teaching me to fish, rather than simply giving me a fish. To my friends, I am forever grateful for the laughter, for the prayers, and for the memories that we shared throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CE: Close-ended

D: Decision

DS: Down syndrome

dB: Decibels

G: Generalizability

ICC: Interclass correlation coefficient

LC: Listening comprehension

MTMM: Multitrait-multimethod matrix

OE: Open-ended

RC: Reading comprehension

SES: Socioeconomic status

TD: Typically developing

CHAPTER I

INTRODUCTION

Proficient reading comprehension serves as a gateway to acquire information, access knowledge, and engage in vocational tasks (Snow, 2002). Many children with primary and secondary language impairment struggle to become proficient readers, placing them at risk for poor outcomes (e.g., Catts et al., 2002; Wei, Blackorby, & Schiller, 2011). Individuals with secondary language impairment, including children with Down syndrome (DS), are of particular interest as there is a recent increased focus on understanding the reading outcomes of children who have historically been underrepresented in reading research (Afacan, Wilerson & Ruppar, 2018). DS, the most common cause of intellectual disability, is characterized by a behavioral phenotype consisting of a pattern of strengths and weaknesses across multiple domains (e.g., cognitive, linguistic, speech-motor, social-emotion; Chapman & Hesketh, 2000; Fidler, 2005; Fidler, Most, & Philofsky, 2008). In relation to reading outcomes, one of the hallmark DS phenotypic characteristics is that individuals often present with language and literacy deficits that are disproportionate to their broader cognitive profiles (Byrne, Buckley, MacDonald, & Bird, 1995).

Despite perpetuated beliefs that children with DS cannot learn to read and comprehend text, an emerging body of evidence challenges this assumption (e.g., Buckley, 2001; Byrne et al., 2002; Lemons et al., 2017). For example, Buckley (2001) found that 60% to 70% of individuals with DS in Australia and the United Kingdom have attained functional levels of literacy. Byrne and colleagues found that some children with DS (ages 4 to 12) demonstrate word-level reading developmental trajectories that are not significantly different compared with development in reading-matched typically developing children (ages 4 to 10). They also found that children with DS presented with word-level reading that is more advanced than their own

cognitive abilities (e.g., Byrne et al., 1995; Byrne et al., 2002). Based on current evidence, many individuals with DS present with a relative strength in word-level reading, however, they often experience persistent difficulties with reading comprehension, the ultimate goal of reading (Catts & Kamhi, 1999). Research for children with DS, though limited, demonstrates that reading comprehension growth tends to progress slowly and achievement rarely reaches levels commensurate to word-level reading skills or oral language abilities (e.g., Byrne, MacDonald, & Buckley, 2002; Groen, Laws, Nation, & Bishop, 2006; Nash & Heath, 2011). Thus, this measurement project evaluated the feasibility and validity of literacy-related measures to lay the foundation for future work in the identification and evaluation of psychometrically-sound assessments and evidence-based interventions for individuals with DS.

Theoretical Models of Reading Comprehension

Across multiple theoretical models of reading that place reading comprehension as the outcome of interest, reading comprehension is viewed as a multidimensional construct (Gough & Tunmer, 1986). Reading comprehension–the construction of meaning from written text and the ultimate goal of reading (Catts & Kamhi, 1999)–requires the coordination of multiple underlying cognitive and linguistic processes (Elleman & Compton, 2017; Fuchs et al, 2018; Kintsch, 1998; Snow, 2002). Proficient word recognition (i.e., decoding) and linguistic comprehension are widely recognized competencies that underlie reading comprehension.

In their simple view of reading, Gough and Tunmer (1986) presented a model wherein reading comprehension equals the product of decoding and linguistic comprehension. Gough and Tunmer (1986) defined decoding as context-free word recognition measured by production of pseudo words (e.g., *stenk*). Linguistic comprehension was defined as the process by which lexical information, sentences, and discourse are interpreted and was measured by retelling a story after hearing it read aloud. Importantly, the simple view presumes that, once printed text is decoded, the reader applies to the text the same mechanisms which s/he would bring to

comprehending its spoken equivalent. Thus, reading comprehension closely parallels linguistic comprehension within the simple view of reading framework. Gough and Tunmer (1986) used listening comprehension as a means to operationalize and measure linguistic comprehension in their seminal publication. Subsequently, researchers have used the terms listening comprehension and language comprehension as synonymous with linguistic comprehension.

In her reading rope model, Scarborough (2001) fleshed out the multidimensionality of these two components that contribute to reading comprehension, labeled in her model as word recognition and language comprehension. Skilled reading, or fluent execution and coordination of word recognition and language comprehension, is achieved as readers become increasingly strategic with language comprehension and increasingly automatic with word recognition. Within the reading rope model, Scarborough defined language comprehension as a multifaceted construct itself, involving background knowledge, language structures, vocabulary, verbal reasoning, and literacy knowledge. Word recognition was similarly defined as a multifaceted construct, involving phonological awareness, decoding, and sight or automatic word recognition.

Due to the multidimensionality of the reading comprehension construct, consideration of the underlying processes—word recognition and linguistic comprehension—is essential to interpreting and understanding reading comprehension outcomes (Snow, 2002; Kamhi & Catts, 2017). Children without language and learning difficulties engage in the task of learning to read with intact linguistic comprehension. Whether they listen to a story or read it themselves, comprehension is comparable. However, children with DS do not engage in the task of learning to read with intact listening comprehension (Cossu et al., 1993; Roch & Levorato, 2009). Because individuals with DS do not have intact listening comprehension, there is a critical need to more closely study listening comprehension if we want to understand reading comprehension in this population.

The Missing Piece: Listening Comprehension

Only a few studies have evaluated the relation between listening comprehension and reading comprehension among individuals with DS (Laws et al., 2016; Roch & Levorato, 2009; Roch et al., 2011). Two limitations emerge from these studies with respect to studying English-speaking children with DS. First, two studies evaluated listening comprehension and reading comprehension in Italian-speaking individuals with DS (Roch & Levorato, 2009; Roch et al., 2011). Unlike English, Italian has a transparent orthography and in transparent languages, listening comprehension is a more powerful predictor of reading comprehension than decoding (Megherbi, Seigneuric, & Ehrlich, 2006). It remains unknown whether a similar relation would emerge among English-speaking individuals with DS. Second, the single study of English-speaking children with DS (Laws et al., 2016) evaluated listening comprehension and reading comprehension at the word, phrase, and sentence levels, but not at the paragraph/passage level.

Roch and colleagues (2009, 2011) evaluated the simple view of reading with Italian-speaking adolescents with DS. Parallel measures were used in which participants listened to or read a story and answered multiple-choice questions to assess listening comprehension and reading comprehension, respectively. Compared to a TD group matched on reading comprehension, the DS group had weaker listening comprehension and in both groups, listening comprehension (r = .41, p < .01 in DS group; r = .27, p > .05 in TD group) more strongly correlated with reading comprehension compared with word recognition (r = .52, p < .01 in DS group; r = .27, p > .05 in TD group; Roch & Levorato, 2009). Further, listening comprehension, but not word recognition, predicted reading comprehension (Roch et al., 2011). In a longitudinal study of a heterogeneous sample of Dutch-speaking children with intellectual disabilities, van Wingerden, Segers, van Blakom, and Verhoeven (2018) found that reading comprehension was predicted by not only decoding and listening comprehension but that foundational literacy skills and nonverbal reasoning contributed to reading comprehension as

well. In the discussion of their findings, the authors acknowledged that further research must consider various task formats when evaluating the simple view of reading—the relation between decoding, listening comprehension, and reading comprehension.

Laws et al. (2016) compared listening comprehension and reading comprehension in children with DS (6- to-13-year-olds) and TD children matched on word-level reading. Listening comprehension was assessed with the Clinical Evaluation of Language Fundamentals–Third Edition UK (CELF-3UK; Semel, Wiig, & Secord, 2000) Sentence Structure subtest in which the examiner reads a short sentence and the child points to the picture (from a field of four) that illustrates the sentence. Reading comprehension was assessed using an experimenter-developed parallel version in which the child reads a short sentence and selects the picture that illustrates it. The CELF-3UK performance was more closely associated with reading comprehension for the DS group (r = .56, p < .05) as compared to the TD group (r = -.04, p > .05). Listening comprehension also predicted later reading comprehension, although different measures were used to examine the longitudinal relation between listening comprehension and reading comprehension (i.e., Reynell Developmental Language Scales [RDLS; Edwards et al., 1997], an omnibus measure of expressive and receptive language, and Reading Understanding subtest of the Kaufman Assessment Battery for Children [KABC; Kaufman & Kaufman, 1983]; children act out written directions, e.g., ‘stand’, respectively).

Given that listening comprehension predicts reading comprehension and is more strongly correlated with reading comprehension in children with DS than TD peers, listening comprehension is hypothesized as the main barrier to reading comprehension for children with DS. If this is the case, then evaluating listening comprehension using psychometrically sound measures is an important consideration in studies related to reading outcomes for individuals with DS. However, it is challenging to obtain valid estimates of these skills in individuals with DS, as there is limited information specific to DS about the psychometrics of common measures of listening comprehension and reading comprehension.

Challenges to Valid Assessment

Three challenges to valid assessment of listening comprehension and reading comprehension include (a) challenges with the constructs, (b) challenges with measures commonly used to assess the constructs of interest, and (c) challenges specific to the DS phenotype.

Challenges with the constructs. First, given that reading comprehension and listening comprehension are multidimensional constructs, the degree to which measures tap various underlying cognitive and linguistic processes differs based on how listening comprehension or reading comprehension is operationalized.

Reading comprehension. To illustrate the challenges that emerge, Cutting and Scarborough (2006) found that the relative contributions of word reading (R^2s = 6.1–11.9%) and oral language (R^2s = 9–15%) to reading comprehension varied across three reading comprehension measures: Wechsler Individual Achievement Test (WIAT; Wechsler, 1992), Gates-MacGinitie Reading Test (MacGinitie, MacGinitie, Maria, & Dreyer, 2000), and Gray Oral Reading Test (GORT; Wiederholt & Bryant, 1992). Additionally, Keenan, Betjemann, and Olson (2008) found only modest intercorrelations (*r*s = .31–.70) among four commonly used reading comprehension measures: GORT (Wiederholt & Bryant, 1992), Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2001), Woodcock-Johnson Test of Achievement Passage Comprehension subtest (Woodcock, McGrew, & Mather, 2001), and Peabody Individual Achievement Test Reading Comprehension subtest (Dunn & Markwardt, 1970). Thus, based on these findings, various reading comprehension measures do not seem to be converging on the same construct, which instead suggests that these tests are measuring different aspects of reading comprehension. Researchers have identified several additional reader characteristics that contribute to comprehending written text, some of which may account for the lack of

association across reading comprehension measures (Miller, Cutting, & McCardle, 2013). These characteristics include reading fluency, working memory, verbal reasoning, background knowledge, motivation and engagement, and executive functioning (e.g., Cutting & Scarborough, 2006; Kintsch & Kintsch, 2005; Perfetti, Marron, & Foltz, 1996; Snow, 2002).

Listening comprehension. It is not surprising that there is similarly a lack of consensus among researchers on how to operationalize listening comprehension and whether listening comprehension and oral language are distinct constructs. Some researchers propose that oral language contributes to listening comprehension, or the opposite, that listening comprehension is part of a broader construct of oral language, and yet others suggest that oral language and listening comprehension are separate constructs (e.g., Gray, Catts, Logan, & Pentimonti, 2017; Hogan, Adolf, & Alonzo, 2014; Catts, Herrera, Nielson, & Bridges, 2015; Kim & Phillips, 2014). The Language and Reading Research Consortium (2017) evaluated the dimensionality of oral language and listening comprehension based on confirmatory factor analysis of data from a population-based sample of preschool through third grade children (n = 1,869). Evidence of oral language and listening comprehension operating as a single construct was stronger in the preschool and kindergarten data as compared with the first through third grade data. Although the best fitting model at all grade levels included two separate factors for oral language (i.e., expressive and receptive vocabulary and grammar) and listening comprehension, oral language and listening comprehension were highly correlated (r = .87-.91). The authors concluded that oral language and listening comprehension were best characterized as a single oral language construct, and thus measures of oral language and listening comprehension appear to assess the same underlying construct. Based on this conclusion, measures of oral language and measures of listening comprehension can presumably be used interchangeably, as they all would yield an estimate of "listening comprehension."

For the purpose of the current study, **reading comprehension** was operationalized as constructing meaning from written text and **listening comprehension** was operationalized as constructing meaning from read-aloud written text. As such, listening comprehension operationalized in this manner is distinct from listening comprehension in conversation or as operationalized in some oral language measures. Unlike listening comprehension as defined above, listening comprehension in the context of conversation includes a certain level of redundancy, additional nonverbal cues, and the opportunity to repair any lapses in comprehension. Further, listening comprehension operationalized as such is distinct from other oral language measures (e.g., vocabulary, grammar comprehension) that do not necessarily require text-level processing (Catts & Kamhi, 1986) and instead often involve comprehension of language at the single word or phrase level. According to the simple view of reading, parallel measures of listening comprehension and reading comprehension are essential to adequately capturing the relation between these two constructs (Gough & Tunmer, 1986). Therefore, common measures or oral language (e.g., grammar comprehension, vocabulary) do not necessarily have parallel formats with measures of reading comprehension. In contrast, the measures included in this study reflect parallel measures of listening comprehension and reading comprehension that align with the operational definitions above.

Challenges with commonly-used measures. Next, listening comprehension and reading comprehension assessment is complicated by substantial variation across measurement methods. To illustrate, Francis et al. (2005) reported a stronger association between decoding and reading comprehension when comprehension was assessed with a cloze-procedure measurement method compared with a multiple-choice question method. Commonly used measures vary in the (a) text format that is presented at the single word, phrase, sentence, or paragraph/passage level and (b) response format that requires the test

taker, for example, to point to a picture to identify the referent or to verbally answer multiple-choice, close-ended, or open-ended questions.

Further, many commonly used standardized measures have psychometric weaknesses, often failing to meet minimal criteria for test reliability and validity (Paris & Stahl, 2005). Petersen and Stoddard (2018) argued that because emphasis has been placed on test reliability, many reading comprehension measures with weak validity have emerged. In particular, content validity, or how well test items adequately represent the entirety of the measured construct, comes into question. Due to weaknesses in content validity, any conclusions about listening comprehension and reading comprehension must be considered in the context of the specific measure used. For any particular measure of comprehension, it is important to evaluate how the construct is operationalized (e.g., recalling facts, constructing inferences), presentation of the test stimuli (e.g., visual or oral), the response format (e.g., oral or written; multiple-choice or open-ended), and the test format (e.g., timed or untimed; individual or group administration; Fuchs et al., 2018).

Challenges specific to the DS phenotype. Finally, listening comprehension and reading comprehension assessment for individuals with DS warrants careful consideration because most measures were not developed with sufficient attention to the myriad characteristics of children with disabilities. Given phenotypic characteristics of DS (e.g., cognitive and linguistic deficits), standardized assessments may not yield valid measurement for this population, despite the demonstration of validity for other populations. The DS behavioral phenotype consists of patterns of strengths and challenges across not only cognitive and linguistic domains, but also speech-motor and social-emotional domains. Two challenges characteristic of the DS phenotype, but perhaps not of other groups of children with intellectual disabilities, may contribute to underestimation of skills. First, the speech of individuals with DS is characterized by persistent, atypical phonological error patterns that have a negative effect on

intelligibility (Stoel-Gammon, 1997). Reduced speech intelligibility may be a confounding factor for reading comprehension measures requiring a verbal response. Second, when faced with cognitive challenges, children with DS are more likely than TD peers to engage in positive and negative behaviors to avoid tasks (Wishart, 1996). This behavior reflects overall poor task persistence and higher levels of off-task social behaviors, especially when cognitive processes are strained, for example, in reading comprehension assessment (Fidler, 2006; Wishart, 1996).

Historically, researchers have not considered behavioral phenotypes in selecting or developing assessment measures to address these challenges (Lemons et al., 2017). Thus, the purpose of this study was to evaluate the feasibility and validity of listening comprehension and reading comprehension measures for individuals with DS. Further, we aimed to ascertain which methods of listening comprehension and reading comprehension assessment are optimal for individuals with DS and their peers with TD. We considered optimal to be feasible to collect and valid measures of the construct of interest.

Purpose and Research Questions

In this study we evaluated the feasibility and construct validity, the degree to which a test measures what it claims to be measuring, for four parallel measures of listening comprehension and reading comprehension. The Multitrait-Multimethod matrix (MTMM; Campbell & Fisk, 1959) is an approach using a matrix of correlations to facilitate the assessment and interpretation of the construct validity of measures across various methods. Within the MTMM, convergent validity and discriminant validity is assessed. Convergent validity refers to the degree to which there is empirical evidence that a measure correlates with other measures of the same construct which are assumed to relate based on theory. Discriminant validity refers to the degree to which there is empirical evidence that constructs can be meaningfully differentiated (i.e., not highly correlated) from other theoretically distinct constructs (Campbell & Fisk, 1959). Several traits and several methods are measured and evaluated within the MTMM. In this study,

we evaluated two traits—listening comprehension and reading comprehension—and four methods (nonverbal response, cloze-procedure, passage-level with close-ended questions, and passage-level with open-ended questions), resulting in an 8 x 8 matrix.

As evidenced by the National Institute of Health's recent commitment to expanding DS research by appropriating funds specific to this clinical population (NIH Research Plan on Down Syndrome, 2014), this study aimed to address the critical need to promote equity in reading research for persons with DS. To obtain valid estimates of listening comprehension and reading comprehension in individuals with DS, an important first step is to evaluate the feasibility of administering and the construct validity of commonly used measures. The following research questions were addressed.

Study 1: Children with TD.

1) What percentage of children complete all listening comprehension and reading comprehension measures and yield a score above the floor level (i.e., raw score > 0)?
2) Across the participant group, what is the mean, standard deviation, and range for administration and scoring time for each listening comprehension measure and reading comprehension measure?
3) Are measures of the same construct that use different methods (monotrait-heteromethod) more strongly correlated than (a) measures of different constructs that use the same method (heterotrait-monomethod) and (b) measures of different constructs that use different methods (heterotrait-heteromethod)?

Study 2: Individuals with DS.

4) What percentage of individuals complete all listening comprehension and reading comprehension measures and yield a score above the floor level (i.e., raw score > 0)?

5) Across the participant group, what is the mean, standard deviation, and range for administration and scoring time for each listening comprehension measure and reading comprehension measure?
6) Are measures of the same construct that use different methods (monotrait-heteromethod) more strongly correlated than (a) measures of different constructs that use the same method (heterotrait-monomethod) and (b) measures of different constructs that use different methods (heterotrait-heteromethod)?

Group Comparisons.

7) Does the proportion of participants completing each measure and yielding a score above the floor level differ between the DS group and the TD group?
8) Is evidence of construct validity moderated by group?

Evaluating the construct validity of commonly used measures that employ different methods will enable researchers and clinicians to ascertain optimal methods for assessing listening comprehension and reading comprehension. The first two research questions in study 1 and study 2 address feasibility of administering measures and thus will provide information for each group regarding the effectiveness and efficiency of test administration for each measure. Research questions three and six address construct validity. For both groups, we hypothesized strong (i.e., $r > .5$) correlations between measures of the two traits that use different measurement methods, weaker correlations between measures of different constructs that use the same method (heterotrait-monomethod), and the weakest correlations between measures of each trait using different measurement methods (heterotrait-heteromethod). In comparing the evidence of construct validity across groups, we hypothesized that at least some of the associations within the MTMM would be moderated by group.

CHAPTER II

METHOD

The study procedures were approved by the Vanderbilt University Institutional Review Board.

Participants

To compare the patterns of construct validity, two groups of participants who were matched on word-level reading were included: (a) individuals with DS and (b) children with TD (control group). To form the TD control group, each participant with DS was matched to one TD participant (i.e., a TD participant could only be paired with a single DS participant) based on word-level reading and sex when possible. Using the Word Identification subtest of the Woodcock Reading Mastery Test-III (WRMT-3; Woodcock, 2011), participants were considered eligible matches if their word-level reading raw score was within three points of the participant with DS' raw score (e.g., a DS participant with a Word Identification raw score of 10 could be matched with a TD participant with a raw score between 7 and 13). Participants were not matched on socioeconomic status (SES), although SES-matching should be considered in future studies. Participant recruitment occurred in three rounds. In the first round, participants with TD were recruited from the Nashville metropolitan area by distributing study flyers (a) on research listservs (e.g., ResearchMatch, Vanderbilt Research Distribution and Notification List), (b) to families whose children had participated in previous research studies in the lab, and (c) at community organizations (e.g., community center, public library). In the second round of recruitment, participants with DS were recruited by distributing study flyers (a) at private schools in Nashville and Dallas/Fort Worth, (b) with DS community organizations (e.g., Down Syndrome Association of Middle Tennessee, University of Alabama Intellectual Disabilities Participant

Registry), (c) on research listservs, and (d) to families whose children had participated in previous research studies in the lab. In the final round, additional children with TD were recruited to create DS-TD matches that were not accomplished in the first two recruitment rounds. During the final round of recruitment, teachers in two Nashville elementary schools sent recruitment packets home with first and second-grade students who were reading on grade-level. Each consented child completed the eligibility session to determine if he/she was a word-level reading match for one of the DS participants. If so, the TD child was considered eligible to complete the remaining study activities. Four additional TD children completed all the study activities in the final round of recruitment.

In total, 43 children with TD completed the eligibility session, of which five did not meet eligibility criteria. Of the TD children who were not eligible to participate, four children were excluded due to limited word-level reading abilities (i.e., reading fewer than 80% of words correctly on the screening), and one child was not a monolingual English speaker. Eight of the TD children were not included as participants (i.e., did not complete the second assessment session) because they were recruited in the third round and deemed not to be a match for one of the DS participants already enrolled in the study. Of the remaining 30 TD children, 19 TD children were matched to one DS participant. Thus, 11 typically developing children who completed the study procedures were not included in the analyses because they were not the "best" match to a DS participant (i.e., their word-level reading score was not within 3 points of a DS participant or they differed in sex).

Twenty-six individuals with DS completed the eligibility session, of which seven did not meet the eligibility criteria. Of the individuals with DS who were not eligible to participate, one individual was not able to successfully complete the screening battery, and the remaining six were excluded due to limited word-level reading abilities (M = 6.00, SD = 6.42, range = 0 – 15; scores from the word-level reading screener with max score = 20). Hence, 19 DS children were included as participants and were matched to a TD control participant.

The eligibility criteria detailed below assisted in identifying participants who were able to participate in study procedures (e.g., adequate hearing and vision, ability to attend to assessments, demonstrated requisite word-level reading skills). The criteria also ensured inclusion of a representative sample of the clinical population of interest. Across both groups, participants must have been monolingual English speakers to ensure that performance on language and reading measures was not influenced by exposure to multiple languages.

Study 1. Nineteen children with TD, aged 6;6 to 8;5 (*M* = 86.11 months, *SD* = 6.54 months) participated in the study. Children were eligible to participate if they (a) demonstrated oral language skills within normal limits and neurotypical development as reported by parents, (b) were monolingual English speakers, (c) successfully completed the screening battery (i.e., listened to directions, completed assessments), (d) passed hearing screening in at least one ear, unaided using ASHA standards (ASHA, 2019), and (e) had normal or corrected to normal vision based on parent report. Exclusionary criteria included: (a) reading fewer than 80% of words correctly on the Phonological Awareness Literacy Screening-Kindergarten primer list (PALS-K; Invernizzi, Meier, Swank, & Juel, 1997), or (b) more than one standard deviation below the normative mean on the measure of nonverbal cognition. The exclusionary criteria ensured that participants in the control group presented with nonverbal intelligence within the average range. See Table 1 for participant demographic information.

Study 2. Nineteen individuals with DS, aged 10;0 to 22;11 (*M* = 206.63 months, *SD* = 41.68 months) participated in the study. Individuals were eligible to participate if they (a) had been diagnosed with DS by a physician as reported by the parent, (b) were monolingual English speakers and used spoken language as a primary form of communication, (c) successfully completed the screening battery (i.e., listened to directions, completed assessments), and (d) had normal or corrected to normal vision based on parent report. The hearing status

inclusionary criteria was not included for the DS group to ensure inclusion of a representative sample of participants with DS, who frequently present with mild to moderate hearing loss (Roizen, Wolters, Nicol, & Blondis, 1993). DS participant hearing screening thresholds are reported (See Table 2). Exclusionary criteria included: (a) reading fewer than 80% of words correctly on the PALS-K (Invernizzi, Meier, Swank, & Juel, 1997) or (b) uncontrolled seizures as reported by the parent. See Table 1 for participant demographic information.

Table 1

Participant Demographic Information

	DS Group (n = 19)	TD Group (n = 19)
Sex		
Male	8	6
Female	11	13
Race		
American Indian/Alaska Native	0	0
Asian	0	0
Black/African American	1	1
Hispanic	0	2
Native Hawaiian/Other Pacific Islander	0	0
White	17	15
Multiple races	1	1
Not reported	0	0
Ethnicity		
Hispanic or Latino	1	3
Not Hispanic or Latino	17	15
Not reported	1	1
Mother's education level		
Some high school	0	0
High school diploma/GED	1	0
Some college	2	3
Associate's degree	3	0
Bachelor's degree	6	9
Master's degree	5	4
Professional degree	2	3

Procedures

Participants completed two individual sessions (eligibility and assessment) at the university lab, school, community location (e.g., public library), or in their home. To determine participant eligibility and ascertain whether an individual could comply with assessment procedures, minimally invasive eligibility and descriptive assessments were administered at the initial session. Standardized and criterion-referenced assessments were administered by (a) the principal investigator, a graduate student and certified speech-language pathologist or (b) a graduate research assistant under the supervision of the principal investigator.

Eligibility Session. Parents or guardians provided written consent (or participants/power of attorneys for individuals over the age of 18), and participants provided written assent. Each participant's guardian provided demographic background information by completing an intake questionnaire. Eligibility assessments included a hearing screening, word-level reading screening, and measure of nonverbal cognition. Additional assessments administered at the eligibility session for descriptive purposes included measures of oral language (receptive and expressive vocabulary and grammar comprehension), word-level reading, and speech accuracy. All eligibility session measures were administered in the same fixed order. The eligibility session lasted 45-60 minutes.

Hearing screening. Pure tone audiometry was used to screen hearing acuity in both ears at frequencies of 500, 1000, 2000, and 4000 Hz at 30dB. A standard hand-raising response was used. For Study 2, when the individual with DS failed to respond to a particular frequency at 30 dB, the intensity of the tone was increased until a reliable response was obtained. The highest intensity necessary to elicit a passing response (two out of three presentations) was recorded (see Table 2).

Table 2

Hearing Screening Thresholds for DS Participants (n = 19)

	Mean (dB)	*SD*	Range
Right ear			
500 Hz	36.47	9.48	30-60
1000 Hz	35.59	9.98	30-60
2000 Hz	32.65	6.40	30-55
4000 Hz	39.71	13.17	30-70
Left ear			
500 Hz	35.29	10.07	30-70
1000 Hz	33.24	6.83	30-50
2000 Hz	32.06	6.14	30-55
4000 Hz	32.94	7.92	30-60

Note. dB = Decibels; *SD* = Standard deviation. Two DS participants wore bilateral hearing aids which parents reported to be tested regularly and in working conditioning at the time of the study sessions.

Nonverbal intelligence. The Kaufman Brief Intelligence-Second Edition Matrices subtest (KBIT-2; Kaufman, 2004) was administered as a brief measure of nonverbal intelligence. The K-BIT is normed for individuals ages 4 to 90 and ideal for those with limited language ability. Test takers identified a relation or rule in a set of pictures or patterns and pointed to the picture or pattern that best fit the relation or rule. The KBIT-2 includes simple oral instructions and only requires test takers to answer with a meaningful gesture such as pointing.

Oral language. The Receptive and Expressive One Word Picture Vocabulary Tests-Fourth Editions (ROWPVT-4 and EOWPVT-4; Martin & Brownell, 2011a; Martin & Brownell, 2011b) were administered as measures of receptive and expressive semantic knowledge. These vocabulary measures are normed for individuals ages 2 to 70. For the ROWPVT-4, test takers pointed to the picture (out of a field of four) that corresponded with the word the examiner said aloud. For the EOWPVT-4, test takers named pictures. The Test of Auditory Comprehension of Language-Fourth Edition Grammatical Morphemes subtest (TACL-4;

Carrow-Woolfolk & Allen, 2014) was administered as a measure of grammar comprehension. The TACL-4 is normed for individuals ages 3;0 to 12;11. Due to limited grammar comprehension characteristic of the DS phenotype, individuals with DS did not reach ceiling levels on this measure despite the DS participant age range extending beyond the intended age range. Test takers pointed to the picture (out of a field of three) that corresponded to items of increasing grammatical complexity presented orally by the examiner.

Word-level reading. To determine eligibility, word-level reading was screened with the PALS-K primer list (Invernizzi, Meier, Swank, & Juel, 1997). Test takers read a list of 20 isolated, real words. Individuals were excluded from participating if they read fewer than 80% of words (16 words) correctly on the PALS-K primer list. For descriptive purposes, the Woodcock Reading Mastery Test-Third Edition Word Identification subtest (WRMT-III; Woodcock, 2011) was administered as a measure of word-level reading. The WRMT-III is normed for individuals ages 4;6 to 79;11. Test takers read isolated, real words.

Speech*.* The Arizona Articulation Proficiency Scale-Fourth Edition (AAPS-4; Fudala & Stegall, 2017) was administered as a measure of speech accuracy. The AAPS-4 is normed for individuals ages 1;6 to 21;11. Test takers named or repeated stimulus items and the examiner noted speech sound production errors.

Assessment Session. Participants completed the measures for each domain described below. Assessment order was counterbalanced across participants in each group to control for order effects. Participants were given breaks between tasks as needed to maintain attention and on-task behavior. The assessment session for each participant lasted 75-100 minutes.

The selected methods represent a range of text and response formats (See Table 3). The nonverbal response method eliminated the need for a verbal response; determining the

accuracy of verbal responses can be confounded by speech unintelligibility in individuals with DS. Although the remaining methods do not minimize verbal demands, they represent methods of engaging in listening comprehension and reading comprehension that may frequently be encountered in academic and vocational settings. The specific measures were selected because the initial test items at lower levels of difficulty and complexity and the amount of scaffolding provided (i.e., illustrated items on the WRMT-III Passage Comprehension subtest) were expected to reduce task demands to minimize floor effects.

Table 3

Methods of Measuring Listening Comprehension and Reading Comprehension

Method	Text format	Response format	Listening comprehension measure	Reading comprehension measure
Nonverbal response	Phrase and sentence	Nonverbal (pointing, acting out)	WJ IV Test of Oral Language Understanding Directions subtest	KABC Reading/ Understanding subtest
Cloze-procedure	Sentence and paragraph	Verbal, one word	WJ IV Test of Oral Language Oral Comprehension subtest	WRMT-III Passage Comprehension subtest
Passage-level with close-ended questions	Paragraph	One word, verbal or pointed	TILLS Listening Comprehension subtest	TILLS Reading Comprehension subtest
Passage-level with open-ended questions	Paragraph	Verbal	WIAT-III Listening Comprehension subtest	WIAT-III Reading Comprehension subtest

Listening comprehension. Four methods of measuring listening comprehension were administered. The Woodcock-Johnson IV Test of Oral Language (WJ IV; Schrank, Mather, & McGrew, 2014) Understanding Directions subtest was administered as a measure that required a nonverbal response. It is normed for individuals ages 2 to 90 years. On this subtest, test takers followed directions presented orally by the examiner to point to familiar objects with

varying characteristics (e.g., size, location) in a picture scene. The Oral Comprehension subtest was administered as a measure that used a cloze-procedure. On this subtest, test takers listened to a short audio-recorded passage and supplied the missing word at the end of the sentence using syntactic and semantic cues. The Test of Integrated Language and Literacy Skills (TILLS; Nelson, Plante, Helm-Estabrooks, & Holtz, 2015) Listening Comprehension subtest was administered as a measure that used passage-level text paired with close-ended questions. It is normed for individuals ages 6;0 to 18;11. On this subtest, test takers selected 'yes', 'no', or 'maybe' to answer questions about passage-level text read aloud by the examiner. A card with the three choices (yes, no, maybe) was placed on the table in front of the examiner as additional visual support and to provide a nonverbal response option. Lastly, the Wechsler Individual Achievement Test-III (WIAT-III; Wechsler, 2009) Listening Comprehension subtest was administered as a measure that used passage-level text paired with open-ended questions. It is normed for individuals ages 4;0 to 50;11. On this subtest, test takers answered open-ended questions about passage-level text read aloud by the examiner.

Reading comprehension. Four reading comprehension measures with parallel measurement methods to the listening comprehension measures were administered. The Kaufman Ability Battery for Children (KABC; Kaufman & Kaufman, 1983) Reading/Understanding subtest was administered as a measure that required a nonverbal response. The Reading/Understanding subtest is normed for individuals 7;0 to 12;6. On this subtest, test takers acted out written directions. The WRMT-III (WRMT-III, Woodcock, 2011) Passage Comprehension subtest was administered as a measure that used a cloze-procedure. It is normed for individuals ages 4;6 to 79;11. On this subtest, test takers supplied the missing word located anywhere in the sentence to complete the meaning of a sentence or paragraph that they read. The TILLS Reading Comprehension subtest was administered as a measure that used passage-level text paired with close-ended questions. It is normed for individuals ages 6;6

to 18;11. On this subtest, test takers read passage-level text and questions and then selected 'yes', 'no', or 'maybe' to answer the questions. In accordance with the manualized directions, the TILLS Reading Comprehension subtest was discontinued if test takers made seven or more miscues when reading the first passage. For the purpose of this study, if the discontinue rule was met, the participant was considered to score at the floor level, though a score was not included on this measure for the feasibility and construct validity analyses. Lastly, the WIAT-III Reading Comprehension subtest was administered as a measure that used passage-level text paired with open-ended questions. It is normed for individuals ages 4;0 to 50;11. On this subtest, test takers read passage-level text and then answered open-ended questions read aloud by the examiner. For participants with typical development, the entry point was based on their current grade level, and for participants with DS, the entry point was based on their word-level reading grade equivalent as measured by the Word Identification subtest of the WRMT-III. Because WIAT-III Reading Comprehension scores are based on the particular item set administered and the total raw scores from different item sets are not directly comparable, vertically scaled scores (i.e., weighted scores) were used as outlined in the assessment manual.

Reliability. A graduate student with formal training in psychoeducational assessment was trained on the scoring procedures for the measures used in the study. The graduate student reliability coder independently scored $\geq$ 25% of the participants assessments from video and audio recordings. Sessions scored for reliability were randomly selected from the subset of video recordings with camera angles that allowed for valid assessment scoring. The primary coder's scoring was used in the analyses. Interobserver reliability was estimated using intraclass correlation coefficients (ICCs). ICCs account for differences in scores between coders as well as the variance among participants on the measures of interest. The primary and secondary scorers double scored (93% inter-rater agreement) all measures used to characterize participants, and all discrepancies were resolved before data was double entered

for analysis. Raw scores were used as variables to capture incremental differences between children that would be obscured by using standard scores for children with intellectual disability.

Statistical Analyses

For Study 1 and 2, descriptive information was reported to answer the feasibility research questions (one, two, four, and five). To answer research questions three and six, separate MTMM were created for the group of children with TD and the individuals with DS. Within the matrices, four classes of cells are distinguished. *Monotrait-monomethod cells* (reliability diagonal, blue cells) constitute the main diagonal of the matrix and contain the reliability coefficient of each trait in each method, as measured by interclass correlations as an estimate of inter-rater reliability. Because a high consistency of scores is an essential requirement for test validity, the monotrait-monomethod cells are expected to be the highest values in the MTMM. *Monotrait-heteromethod cells* (validity diagonal, yellow cells) reflect the correlation between measures of the same trait measured using different methods (convergent validity). Because the two measures are of the same trait, strong correlations are expected. *Heterotrait-monomethod cells* (purple cells) reflect the correlation among measures that share the same measurement method. These values are considered an index of discriminant validity and thus should be weaker than the correlations in the yellow cells. If, however, these correlations are high, it is because measuring different constructs with the same methods results in correlated measures. *Heterotrait-heteromethod cells* (green cells) reflect the correlation among measures that differ in trait and method (discriminant validity). Because these correlations share neither trait nor method, the heterotrait-heteromethod cells are expected to be the lowest values in the MTMM. Summary level statistics are reported for each matrix to ascertain the extent to which the cells overlap or differ from another. Cook's distance was used to monitor for undue influence across analyses relevant to each cell within the MTMM. There was no evidence that any individual data points were leveraging regression lines. Because

scores were not reported for participants who met the discontinue rule on the TILLS Reading Comprehension measure, follow-up analyses demonstrated that the study results were robust to listwise deletion.

To answer research question seven, a chi-square test of independence was used to examine the relation between group membership and the proportion of participants who completed each measure and yielded a score above the floor level. To answer research question eight, a combined MTMM with data from both groups was created and regression analyses were used to evaluate whether the evidence of construct validity was moderated by group for each cell in the MTMM.

CHAPTER III

RESULTS

Participants completed assessments of nonverbal cognition, receptive and expressive vocabulary, grammar comprehension, word identification, and speech accuracy for descriptive purposes. See Tables 4 and 5 for participant raw scores and standard scores, respectively, on the descriptive measures. As Table 4 shows, the groups significantly differed on all descriptive measures except for word-level reading ($t(36) = -.21$, $p = .83$) which was the matching criteria for this study. Of the 19 participants in the TD group, 11 participants yielded a standard score within the average range (standard score 85–115) on the word identification measure. Of the remaining seven participants in the TD group, six scored in the above average range and one scored in the below average range. Because the age range of the DS group extended beyond the TACL-4 age range (normed for individuals 3;0 to 12;11), TACL-4 standard scores are not reported for the DS group. See Table 6 for participant raw scores on the measures evaluated for construct validity; raw scores and weighted raw scores for the WIAT-III Reading Comprehension measure were used in all analyses to capture incremental differences between participants that would be obscured by using standard scores. Again, standard scores are reported for the measures that were normed for the age range of the DS group (see Table 7). The age range of the DS group extended beyond the age range for the following measures: TILLS Listening Comprehension and Reading Comprehension, WIAT-III Listening Comprehension and Reading Comprehension, and KABC Reading/Understanding.

Table 4

Participant Characteristics in Raw Scores

	DS Group (n = 19)			TD Group (n = 19)			
	Mean	SD	Range	Mean	SD	Range	*p*
Age (months)	206.63	41.68	133-273	86.11	6.54	78-97	.00*
KBIT-2	16.21	5.02	10-28	25.42	5.63	14-34	.00*
ROWPVT-4	77.58	27.85	22-132	101.47	8.71	82-117	.00*
EOWPVT-4	82.95	19.40	50-117	96.79	14.32	68-122	.01*
TACL-4 Grammatical Morphemes	35.53	8.73	19-54	48.16	4.71	41-54	.00*
WRMT-3 Word Identification	21.32	6.79	50-117	20.84	6.90	11-34	.83
AAPS-4	88.92	7.27	74-100	97.90	3.34	88-100	.00*

Note. DS = Down syndrome; TD = Typically developing; SD = Standard deviation; KBIT-2 = Kaufman Brief Intelligence Test-2; ROWPVT-4 = Receptive One Word Picture Vocabulary Test-4; EOWPVT-4 = Expressive One Word Picture Vocabulary Test-4; TACL-4 = Test of Auditory Comprehension of Langauge-4; WRMT-3 = Woodcock Reading Mastery Test-3; AAPS-4 = Arizona Articulation Phonology Scale-4.

Table 5

Participant Characteristics in Standard Scores

	DS Group (n = 19)			TD Group (n = 19)		
	Mean	SD	Range	Mean	SD	Range
KBIT-2	52.37	12.25	40-80	109.47	13.26	82-127
ROWPVT-4	59.63	7.87	55-81	112.89	7.80	96-127
EOWPVT-4	62.67	10.81	55-86	111.32	14.57	85-131
TACL-4 Grammatical Morphemes*				11.58	2.22	8-15
WRMT-3 Word Identification	61.68	11.07	55-86	110.21	15.91	75-138
AAPS-4	88.92	7.27	74-99.5	96.95	6.70	76-100

Note. DS = Down syndrome; TD = Typically developing; SD = Standard deviation; KBIT-2 = Kaufman Brief Intelligence Test-2; ROWPVT-4 = Receptive One Word Picture Vocabulary Test-4; EOWPVT-4 = Expressive One Word Picture Vocabulary Test-4; TACL-4 = Test of Auditory Comprehension of Langauge-4; WRMT-3 = Woodcock Reading Mastery Test-3; AAPS-4 = Arizona Articulation Phonology Scale-4; *TACL-4 Scores not reported for DS Group because the age range of the DS group extended beyond the TACL-4 age range.

Table 6

Participant Characteristics on Construct Validity Measures in Raw Scores

	DS Group (n = 19)			TD Group (n = 19)		
	Mean	SD	Range	Mean	SD	Range
Listening Comprehension						
WJ IV TOL Understanding Directions	17.74	9.89	2-37	35.47	6.53	22-50
WJ IV TOL Oral Comprehension	7.84	4.62	0-17	15.26	2.62	10-20
TILLS Listening Comprehension	9.05	3.55	0-16	13.26	3.87	7-20
WIAT-III Listening Comprehension	4.95	4.13	0-15	11.42	1.90	8-16
Reading Comprehension						
KABC Reading/Understanding	8.58	5.32	0-18	10.58	5.64	2-19
WRMT-III Passage Comprehension	8.68	4.41	2-17	13.32	3.73	9-22
TILLS Reading Comprehension	5.12	3.82	0-11	9.53	3.80	4-15
WIAT-3 Reading Comprehension	27.42	14.67	2-55	46.37	9.71	30-64

Note. DS = Down syndrome; TD = typically developing; WJ IV TOL = Woodcock-Johnson IV Tests of Oral Language; TILLS = Test of Integrated Language and Literacy; WIAT-III = Wechsler Individual Achievement Test–Third Edition; KABC = Kaufman Assessment Battery for Children; WRMT-III = Woodcock Reading Mastery Test–Third Edition.

Table 7

Participant Characteristics on Construct Validity Measures in Standard or Scaled Scores

	DS Group (n = 19)			TD Group (n = 19)		
	Mean	SD	Range	Mean	SD	Range
Listening Comprehension						
WJ IV TOL Understanding Directions	47.79	10.91	40-70	103.89	12.4	80-128
WJ IV TOL Oral Comprehension	46.32	8.95	40-69	104.74	11.43	80-119
TILLS Listening Comprehension				9.37	3.37	2-15
WIAT-III Listening Comprehension				102.58	10.47	80-118
Reading Comprehension						
KABC Reading/Understanding				108.36	12.23	86-132
WRMT-III Passage Comprehension	57.16	4.06	55-70	107.79	11.83	89-134
TILLS Reading Comprehension				10.27	2.89	5-15
WIAT-3 Reading Comprehension				101.84	11.54	79-124

Note. DS = Down syndrome; TD = typically developing; WJ IV TOL = Woodcock-Johnson IV Tests of Oral Language; TILLS = Test of Integrated Language and Literacy; WIAT-III = Wechsler Individual Achievement Test–Third Edition; KABC = Kaufman Assessment Battery for Children; WRMT-III = Woodcock Reading Mastery Test–Third Edition.

Evaluating Feasibility in TD Group

To evaluate the feasibility of administering four parallel measures of listening comprehension and reading comprehension, we calculated the percentage of children who completed all the listening comprehension and reading comprehension measures and yielded a score above the floor level. Additionally, we calculated the mean length of administration time and scoring time for each measure. In the TD group, 78.95% of children completed the listening comprehension and reading comprehension measures and yielded a score above the floor level (i.e., raw score > 0) for all the measures across the various methods. Four participants with TD did not yield a score above the floor level on the passage-level with close-ended questions (TILLS) reading comprehension measure. All participants with TD completed and yielded a score above the floor level for the seven remaining measures. Table 8 displays the percentage of children completing the measures and yielding scores above the floor level as well as the mean, standard deviation, and range for administration time for each measure. For the TD group, the average administration time ranged from five to 13 minutes and the average scoring time was less than 100 seconds for each measure. Taken together, the results demonstrate that all eight measures are feasible to administer to TD children, though the passage-level with close-ended questions (TILLS) reading comprehension measure was less feasible to administer given that only 79% of participants yielded a score above floor level.

Table 8

Feasibility Data for Four Parallel Measures of Listening Comprehension and Reading Comprehension

		DS Group (n = 19)			TD Group (n = 19)			
Measure	Trait and Method	% participants above floor level	Administration time in min : sec Mean (*SD*) Range	Scoring time in seconds Mean (*SD*) Range	% participants above floor level	Administration time in min : sec Mean (*SD*) Range	Scoring time in seconds Mean (*SD*) Range	Administration time effect size (*d*)
WJ IV Test of Oral Language Understanding Directions subtest	LC Nonverbal	100.0	9:00 (3:03) 5:24-15:33	61.71 (6.97) 52-76	100.0	12:29 (2:28) 5:56-15:59	76.42 (16.27) 59-115	1.282
KABC Reading/ Understanding subtest	RC Nonverbal	89.4	4:55 (2:05) 1:55-8:16	53.67 (4.04) 49-56	100.0	5:14 (2:22) 2:28-11:22	45.00 (10.41) 26-62	0.145
WJ IV Test of Oral Language Oral Comprehension subtest	LC Cloze Procedure	89.4	7:14 (2:24) 4:09-12:44	58.82 (6.36) 24-51	100.0	7:34 (1:15) 5:16-10:13	71.42 (17.65) 50-110	0.179
WRMT-III Passage Comprehension subtest	RC Cloze Procedure	100.0	5:41 (2:33) 2:18-12:33	35.58 (6.36) 24-51	100.0	8:56 (4:46) 3:03-25:29	40.79 (12.43) 29-85	0.870
TILLS Listening Comprehension subtest	LC Passage-level/ CE questions	94.7	6:43 (1:17) 3:33-9:06	40.08 (7.89) 33-58	100.0	6:29 (1:00) 4:39-8:28	50.61 (9.98) 39-74	-0.213
TILLS Reading Comprehension subtest	RC Passage-level/ CE questions	84.2	10:10 (2:57) 5:27-15:22	27.20 (6.94) 19-39	79.0	12:26 (6:21) 5:43-25:23	41.67 (10.79) 29-74	0.479
WIAT-III Listening Comprehension subtest	LC Passage-level/ OE Questions	94.7	5:15 (2:47) 1:43-12:22	58.60 (13.67) 38-80	100.0	7:34 (1:02) 5:03-9:42	68.47 (9.75) 49-84	1.141
WIAT-III Reading Comprehension subtest	RC Passage-level/ OE Questions	100.0	10:22 (2:59) 6:00-17:44	71.93 (16.68) 52-110	100.0	9.50 (4:28) 4:13-20:35	92.95 (16.81) 63-132	-0.145

Note. DS = Down syndrome; TD = typically developing; LC = Listening Comprehension; RC = Reading Comprehension; CE = Close-ended; OE = Open-ended; WJ IV = Woodcock-Johnson IV; KABC = Kaufman Assessment Battery for Children; WRMT-III = Woodcock Reading Mastery Test; TILLS = Test of Integrated Language and Literacy; WIAT-III = Wechsler Individual Achievement Test.

Evaluating Construct Validity in TD Group

The TD group MTMM is displayed in Figure 1. The reliability diagonal marked in blue reflects the interclass correlation values as a measure of inter-rater reliability for each measure. The interclass correlation values for all measures ranged from good to excellent (ICCs = .80–1.00). Next, the monotrait-heteromethod cells marked in yellow reflect the correlation between measures of the same trait using different measurement methods. Statistically significant and strong ($r \geq .5$; Cohen, 1988) correlations (rs = .62–.88, $p < .05$) were observed in 75% of the monotrait-heteromethod yellow cells, thus reflecting good convergent validity across all measures. One notable exception of non-significant and weak correlations (rs = .15–.39, $p > .05$) was observed for the passage-level with close-ended questions (TILLS) listening comprehension measure. Given that this listening comprehension measure is not converging with other measures of listening comprehension, it appears that the TILLS Listening Comprehension measure is not tapping the construct that it purports to measure. It also may be the case that the TILLS Listening Comprehension measure is tapping a different dimension of the listening comprehension construct when compared with the other measurement methods. Because of the questionable construct validity of the TILLS, we will hone in on the cells reflecting only the associations of the remaining measures from this point forward.

Figure 1

TD Group MTMM

		Nonverbal response		Cloze procedure		Passage-level/ close-ended questions		Passage-level/ open-ended questions	
		LC	RC	LC	RC	LC	RC	LC	RC
Nonverbal response	LC	.99							
	RC	.41 [-.06, .73]	.99						
Cloze procedure	LC	.62* [.22, .84]	.43 [-.03, .74]	.96					
	RC	.41 [-.05, .73]	.78* [.50, .91]	.58* [.17, .82]	1.00				
Passage-level/ close-ended questions	LC	.15 [-.33, .56]	.49* [.04, .77]	.39 [-.08, .72]	.57* [.16, .81]	1.00			
	RC	.66* [.22, .88]	.64* [.18, .87]	.66* [.23, .88]	.63* [.17, .86]	.53* [.02, .82]	1.00		
Passage-level/ open-ended questions	LC	.70* [.36, .88]	.38 [-.09, .71]	.66* [.29, .86]	.44 [-.02, .74]	.17 [-.30, .58]	.43 [-.10, .77]	.80	
	RC	.55* [.13, .80]	.88* [.70, .95]	.60* [.21, .83]	.81* [.56, .92]	.59* [.19, .82]	.76* [.41, .92]	.57* [.16, .81]	.84

Note. Multitrait-multimethod matrix for typically developing group. Monotrait-monomethod cells (reliability diagonal) marked in blue, monotrait-heteromethod cells (validity diagonal) marked in yellow, heterotrait-monomethod cells marked in purple, and heterotrait-heteromethod cells marked in green.

The heterotrait-monomethod cells marked in purple reflect the correlations between listening comprehension and reading comprehension measures using the same method. Statistically significant and strong correlations (*r*s = .53–.58, $p < .05$) were observed between the two traits of interest—listening comprehension and reading comprehension—for three out of the four measurement methods (i.e., cloze procedure, passage-level with close-ended questions, and passage-level with open-ended questions). However, the two traits of interest were not significantly correlated ($r = .41$, $p > .05$) for the nonverbal response (KABC Reading/Understanding and WJ IV TOL Oral Comprehension) measurement method. This pattern of strong correlations (Cohen, 1988) demonstrates shared method variance, that measuring different constructs with the same methods results in correlated measures. It is also

important to note that the values in monotrait-heteromethod (yellow) cells were not significantly stronger than the values in the heterotrait-monomethod cells, as evidenced by the overlapping confidence intervals. Lastly, the heterotrait-heteromethod cells marked in green reflect the correlation between different traits measured using different methods. Statistically significant and strong correlations (*r*s = .49–.66, $p < .05$) were observed in 58% of the heterotrait-heteromethod green cells, with the remaining cells reflecting moderate correlations (*r*s = .38–.44, $p > .05$). Given this range of values, the values in the heterotrait-heteromethod (green) cells were not all significantly weaker than the values in the heterotrait-monomethod or the monotrait-heteromethod cells. In summary, the results in the TD group provide some support that these measures, with the exception of the TILLS Listening Comprehension measure, are measuring the same constructs. Within this sample, the results also do not provide evidence of discriminant validity, in other words, that the listening comprehension and reading comprehension constructs are separable.

Evaluating Feasibility in DS Group

To evaluate the feasibility of administering four parallel measures of listening comprehension and reading comprehension, we calculated the percentage of individuals who completed all the listening comprehension and reading comprehension measures and yielded a score above the floor level. Additionally, we calculated the mean length of administration time and scoring time for each measure. In the DS group, 63.16% of individuals with DS completed the listening comprehension and reading comprehension measures and yielded a score above the floor level (i.e., raw score > 0) for all the measures across the various methods. Similar to the TD group, three participants with DS did not yield a score above the floor level on the passage-level with close-ended questions (TILLS) reading comprehension measure. Additionally, two participants with DS did not yield a score above the floor level on the following measures: the nonverbal response (KABC) listening comprehension and the cloze procedure

(WJ IV TOL) listening comprehension measure. One participant with DS did not yield a score above the floor level on the following measures: the passage-level with close-ended questions (TILLS) measure and the open-ended questions (WIAT) listening comprehension measure. Out of the seven participants with DS who did not complete and yield a score above the floor level, five participants did not score above floor level on one measure and the remaining two participants did not score above floor level on two measures. All participants with DS completed and yielded a score above the floor level for three out of the eight measures: the nonverbal response (WJ IV TOL Understanding Directions) listening comprehension measure, the cloze procedure (WRMT-III) reading comprehension measure, and the passage-level with open-ended questions (WIAT-III) reading comprehension measure.

The percentage of individuals completing the measures and yielding scores above the floor level as well as the mean, standard deviation, and range for administration time for each measure are displayed in Table 8. For the DS group, the average administration time ranged from five to ten minutes and the average scoring time was less than a minute and a half for all measures. Taken together, the results demonstrate that all eight measures are feasible to administer for the vast majority of individuals with DS.

Evaluating Construct Validity in DS Group

The DS group MTMM is displayed in Figure 2. The reliability diagonal marked in blue reflects the interclass correlation values as a measure of inter-rater reliability for each measure. Similar to the TD group, the interclass correlation values for all measures were excellent (ICCs = .94–1.00). Next, the monotrait-heteromethod cells marked in yellow reflect the correlation between measures of the same trait using different measurement methods. Statistically significant and strong correlations (rs = .77–.90, $p < .05$) were observed in half of the monotrait-heteromethod yellow cells, thus reflecting good convergent validity across the measures except for the passage-level with close-ended questions (TILLS) measures. Similar to the TD group,

non-significant and weak correlations (*r*s = .05–.18, *p* > .05) were observed for the passage-level with close-ended questions (TILLS) listening comprehension measure. Additionally, non-significant and weak correlations (*r*s = .12–.40, *p* > .05) were observed for the passage-level with close-ended questions (TILLS) reading comprehension measure. Given that these measures are not converging with other measures of the same construct, it appears that the TILLS measures are not tapping the construct that it's purporting to measure for the DS group. Because of the questionable construct validity of the TILLS, we will hone in on the cells reflecting only the associations of the remaining measures from this point forward.

Figure 2

DS Group MTMM

		Nonverbal response		Cloze procedure		Passage-level/ close-ended questions		Passage-level/ open-ended questions	
		LC	RC	LC	RC	LC	RC	LC	RC
Nonverbal response	LC	1.00							
	RC	.79* [.52, .92]	.99						
Cloze procedure	LC	.90* [.75, .96]	.71* [.37, .88]	1.00					
	RC	.75* [.45, .90]	.86* [.65, .94]	.70* [.36, .87]	.99				
Passage-level/ close-ended questions	LC	.05 [-.41, .49]	.08 [-.39, .52]	.18 [-.30, .59]	.15 [-.33, .57]	.98			
	RC	-.15 [-.59, .36]	.12 [-.39, .57]	-.11 [-.56, .39]	.26 [-.26, .66]	.09 [-.41, .55]	1.00		
Passage-level/ open-ended questions	LC	.78* [.51, .91]	.65* [.27, .85]	.88* [.71, .95]	.68* [.32, .86]	.15 [-.32, .57]	-.01 [-.49, .47]	1.00	
	RC	.64* [.26, .85]	.77* [.49, .91]	.68* [.33, .87]	.81* [.56, .92]	.06 [-.41, .50]	.40 [-.10, .74]	.69* [.35, .87]	.94

Note. Multitrait-multimethod matrix for Down syndrome group. Monotrait-monomethod cells (reliability diagonal) marked in blue, monotrait-heteromethod cells (validity diagonal) marked in yellow, heterotrait-monomethod cells marked in purple, and heterotrait-heteromethod cells marked in green.

The heterotrait-monomethod cells marked in purple reflect the correlations between listening comprehension and reading comprehension measures using the same method. Statistically significant and strong correlations (*r*s = .69–.79, $p < .05$) were observed between the two traits of interest—listening comprehension and reading comprehension—for three out of the four measurement methods (i.e., nonverbal response, cloze procedure, and passage-level with open-ended questions). However, the two traits of interest were not significantly correlated ($r = .09$, $p > .05$) for the passage-level with close-ended question (TILLS) measurement method. This pattern of strong correlations (Cohen, 1988) demonstrates shared method variance, that measuring different constructs with the same methods results in correlated measures. Similar to the TD group, the values in monotrait-heteromethod (yellow) cells were not significantly stronger than the values in the heterotrait-monomethod (purple) cells. Lastly, the heterotrait-heteromethod cells marked in green reflect the correlation between different traits measured using different methods. When excluding the associations related to the TILLS measures, statistically significant and strong correlations (*r*s = .64–.75, $p < .05$) were observed in the remaining heterotrait-heteromethod green cells. The values in the heterotrait-heteromethod (green) cells were not all significantly weaker than the values in the heterotrait-monomethod or the monotrait-heteromethod cells. In summary, the results in the DS group again provide some support that these measures, with the exception of the TILLS measures, are measuring the same constructs. Within the DS group, the results also do not provide evidence of discriminant validity, in other words, that the listening comprehension and reading comprehension constructs are separable.

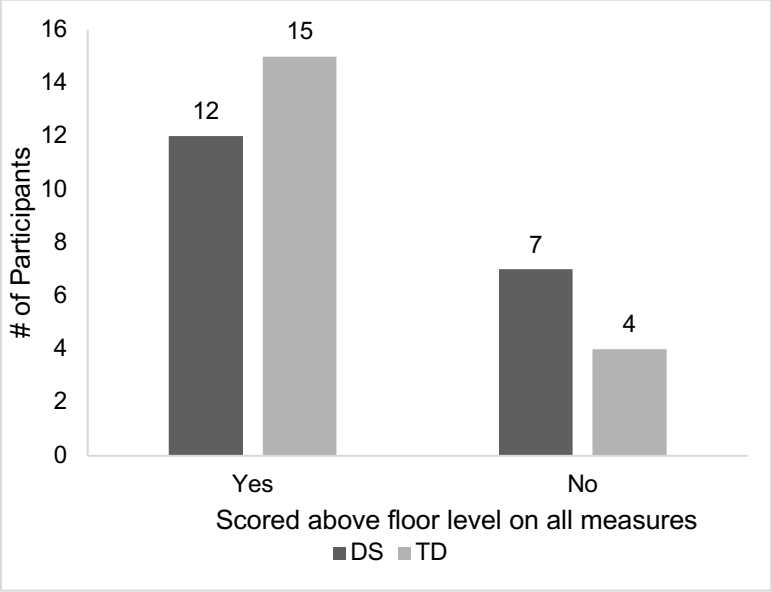
Group Comparisons – Feasibility and Construct Validity

A chi-square test of independence was performed to examine the relation between group membership and completing each measure and yielding a score above the floor level.

The percentage of participants that completed each measure and yielded a score above the floor level did not differ by group χ^2 (1, n = 38) = 1.15, p = .28 (See Figure 3).

Figure 3

Scoring Above Floor Level by Group



Regression analyses were performed to test whether the associations of interest within the MTMM (excluding the reliability diagonal) varied according to group. Only seven associations were significantly different, all but one of which were within heterotrait cells. See Figure 4. Four associations reflected that correlations were slightly stronger in the DS group, although all correlations for both groups ranged from moderate to strong (*r*s = .38-.79). Further, these correlations did not yield a meaningful interpretation given that they all index associations between indices purported to tap different constructs (i.e., associations moderated by group were all in heterotrait cells). The remaining three correlations moderated by group suggest that associations were attenuated in the DS group compared with the TD group. However, two of these correlations were expected to be small given that they were values contained in

heterotrait-heteromethod cells. That is, they reflect correlations that were purported to tap different constructs using different methods. In summary, only a few associations were moderated by group and thus the moderated associations on the whole do not suggest variable construct validity for these measures in the TD group compared with the DS group. The pattern of results was not moderated in any way that is meaningful for interpretation within the MTMM.

Figure 4

Combined Groups MTMM

		Nonverbal response		Cloze procedure		Passage-level/ close-ended questions		Passage-level/ open-ended questions	
		LC	RC	LC	RC	LC	RC	LC	RC
Nonverbal response	LC								
	RC	.54*							
Cloze procedure	LC	.91	.53*						
	RC	.74	.78	.75					
Passage-level/ close-ended questions	LC	.42	.35	.51	.51				
	RC	.44	.43	.43*	.57	.50			
Passage-level/ open-ended questions	LC	.89	.49*	.92	.71	.45	.43		
	RC	.78*	.73*	.80	.86	.50*	.67	.80	

Note. Multitrait-multimethod matrix for combined groups. Correlation values for monotrait-heteromethod cells (validity diagonal) marked in yellow, heterotrait-monomethod cells marked in purple, and heterotrait-heteromethod cells marked in green. Significance of regression interaction term (predictor x group) indicated as * = $p < .05$.

CHAPTER IV

DISCUSSION

In this study, we assessed the feasibility and construct validity of four parallel measures of listening comprehension and reading comprehension for individuals with DS and their TD peers. Evaluation of psychometric properties is important to validate the use of commonly used norm-referenced measures for various clinical populations. Though establishing measures as demonstrating strong reliability and validity is essential in development, it is also essential to determine whether those characteristics hold true for each research sample of interest. Further, given that there are a variety of methods for assessing listening comprehension and reading comprehension, it is important to determine whether measures of the same traits using different methods demonstrate convergent validity (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008). Researchers can make informed decisions regarding assessment and outcome measure selection based on the empirical evidence regarding feasibility and psychometric properties of commonly used measures.

Demonstrating Feasibility

To ascertain which methods of listening comprehension and reading comprehension assessment are optimal for individuals with DS and children with TD, feasibility of administration in combination with evidence of construct validity should be considered. We determined optimal methods of assessment based on whether they were feasible to collect and valid measures of the construct of interest. In this study, we were interested in evaluating floor effects as well as the average administration and scoring times as indicators of feasibility for the TD and DS groups. Floor effects occur when many participants score at the minimum values of a measure, and suggest that the measure has an inadequate range to capture the variability present in a

given population (Maenner et al., 2013). When assessing individuals with intellectual disabilities, standardized language and literacy measures are prone to floor effects (Mervis & Robinson, 2005). By intentionally including measures with a range of text and response formats, we demonstrated limited floor effects in the TD and DS groups.

A large percentage (i.e., $\geq$ 89%) of participants in both groups completed and yielded scores above floor level (i.e., raw score > 0) for all measures. The single measure for which 21% of TD participants and 16% of DS participants were not able to complete and yield a raw score above floor level was the passage-level with close-ended questions (TILLS) reading comprehension measure. As a reminder, in accordance with the manualized test directions, four TD participants and two DS participants met the discontinue rule on this measure. Three of the four TD participants and two of the three DS participants who did not score above floor level, obtained word-level reading scores in the lowest quartile in their respective groups. Additionally, two of the TD participants were also among the youngest children in the sample and thus were near the lower age limit for the TILLS Reading Comprehension subtest. As such, the results indicate that the passage-level with close-ended questions (TILLS) reading comprehension measure is less feasible for test takers with limited word-level reading especially for TD children near the lower age limit (i.e., 6;6 years old) compared with more proficient readers or older children. The first columns in Tables 9 and 10 indicate the rank ordering of the optimal methods of listening comprehension and reading comprehension assessment based solely on the data regarding floor effects. The rank order in the first column lists the measures from highest to lowest percentage of participants scoring above floor level.

Table 9

TD Group Rank Ordering of Optimal Methods for Assessing Listening Comprehension and Reading Comprehension by Criteria

	Floor effects	Floor effects + administration time	Floor effects + Administration time + Construct validity
Listening Comprehension			
WJ IV TOL Understanding Directions	1	4	3
WJ IV TOL Oral Comprehension	1	2	1
TILLS Listening Comprehension	1	1	4
WIAT-III Listening Comprehension	1	2	1
Reading Comprehension			
KABC Reading/Understanding	1	1	1
WRMT-III Passage Comprehension	1	2	2
TILLS Reading Comprehension	4	4	4
WIAT-3 Reading Comprehension	1	3	3

Note. Measures are ordered 1 to 4 as most to least optimal to administer. TD = typically developing; WJ IV TOL = Woodcock-Johnson IV Tests of Oral Language; TILLS = Test of Integrated Language and Literacy; WIAT-III = Wechsler Individual Achievement Test–Third Edition; KABC = Kaufman Assessment Battery for Children; WRMT-III = Woodcock Reading Mastery Test–Third Edition.

Table 10

DS Group Rank Ordering of Optimal Methods for Assessing Listening Comprehension and Reading Comprehension by Criteria

	Floor effects	Floor effects + administration time	Floor effects + Administration time + Construct validity
Listening Comprehension			
WJ IV TOL Understanding Directions	1	4	3
WJ IV TOL Oral Comprehension	4	3	2
TILLS Listening Comprehension	1	2	4
WIAT-III Listening Comprehension	3	1	1
Reading Comprehension			
KABC Reading/Understanding	3	2	2
WRMT-III Passage Comprehension	1	1	1
TILLS Reading Comprehension	4	4	4
WIAT-3 Reading Comprehension	1	3	3

Note. Measures are ordered 1 to 4 as most to least optimal to administer. DS = Down syndrome; WJ IV TOL = Woodcock-Johnson IV Tests of Oral Language; TILLS = Test of Integrated Language and Literacy; WIAT-III = Wechsler Individual Achievement Test–Third Edition; KABC = Kaufman Assessment Battery for Children; WRMT-III = Woodcock Reading Mastery Test–Third Edition.

The length of administration time has important implications for the test-taker and examiner. Given that children with DS demonstrate poor task persistence and a tendency to avoid tasks when faced with cognitive challenges, it is essential to consider administration time when planning an evaluation and administering assessments to this clinical population (Fidler, 2006; Wishart, 1996). Further, efficiency of test administration and scoring procedures has important implications for researchers and clinicians who are expected maintain high levels of productivity. When comparing the average administration time between groups, effect sizes ranged from -.145 to 1.282, indicating small to large group effects. Large effect sizes were observed for the nonverbal (WJ IV) listening comprehension measure, cloze procedure (WRMT-III) reading comprehension measure, and the passage-level with open-ended question (WIAT-III) listening comprehension measure. The longer administration time for the TD group likely reflects a greater number of items being administered before a ceiling was obtained or exposure to longer passages in the reading comprehension measure such that the test taker thus needed more time to read. Anecdotally, and contrary to reports for younger children with DS, we did not observe many avoidance behaviors or poor task persistence during testing in the DS group. However, it should be noted that start point selection was informed by developmental level (rather than based on age or grade) in an attempt to optimize the efficiency with which the measures were administered. Next, we compare the average administration time between measures. The second column in Tables 9 and 10 indicates the rank ordering from most to least feasible of the optimal methods of listening comprehension and reading comprehension assessment based on minimizing floor effects and administration time. The length of scoring time has important implications for the examiner, though limited variation in the average scoring time was observed—all measures were scored within an average of 100 seconds or less.

Taken together, we demonstrated the feasibility of administering all of the listening comprehension and reading comprehension measures for the TD and DS groups. The rank

ordering of measures based on the feasibility results can guide determination of the optimal methods of listening comprehension and reading comprehension assessment.

Demonstrating Construct Validity

The MTMM approach proposed by Campbell and Fiske (1959) was chosen for the assessment and interpretation of construct validity. We were interested in evaluating whether measures of the same construct that use different methods were more strongly correlated than (a) measures of different constructs that use the same method and (b) measures of different constructs that use different methods. In other words, we were interested in evaluating whether the monotrait-heteromethod (yellow) associations were more strongly correlated compared with the heterotrait-monomethod (purple) and heterotrait-heteromethod (green) associations. Inspection of the MTMMs for both groups revealed that monotrait-heteromethod associations were not significantly different when compared with the heterotrait-monomethod and heterotrait-heteromethod associations, as evidenced by overlapping confidence intervals. Thus, the results indicate that the listening comprehension and reading comprehension constructs may not be separable or cannot be meaningfully differentiated for the study groups in this developmental period using these specific measures.

The current findings are consistent with the broader literature in which researchers have suggested that listening comprehension and reading comprehension are highly intercorrelated in readers (e.g., Sticht, Beck, Hauke, Kleinman, & James, 1974; Sinatra, 1990; Nation & Snowling, 2004). In a study of concurrent and longitudinal predictors of reading comprehension, Nation and Snowling (2004) examined reading development in 72 children at 8.5 and 13 years of age. Based on concurrent analyses at Time 1, they found that even after controlling for nonverbal cognition, phonological awareness, semantics, and expressive vocabulary; listening comprehension was the strongest contributor to reading comprehension, accounting for 31% of the unique variance. Based on longitudinal analyses, they found that even after controlling for

Time 1 nonverbal cognition, reading comprehension, nonword reading, phonological awareness, semantics, and expressive vocabulary; listening comprehension accounted for an additional 14% of the unique variance in reading comprehension at Time 2. Further, Ebert and Scott (2016) found statistically significant and strong correlations between listening comprehension and reading comprehension among younger (r = .47, p < .05; aged 6.0 to 8.11) and older (r = .47, p < .05; aged 9.1 to 16.7) school-aged children with TD. Listening comprehension and reading comprehension have been found to be highly intercorrelated (rs = .41–56, p < .05) in studies with children with DS as well, with some stronger correlations between listening comprehension and reading comprehension observed in children with DS compared with children with TD (Laws et al., 2016; Roch & Lovarto, 2009). Our results indicate a similar pattern with stronger correlations between listening comprehension and reading comprehension measured using the same method observed in the DS group (rs = .69–.79) compared with the TD group (rs = .41–.58).

It is important to consider the possible influences of development when interpreting these findings. Based on the simple view of reading model, it is not surprising that we observed strong intercorrelations between listening comprehension given that the participants in this study had achieved some level of proficiency with word recognition. However, the strength of the relation between listening comprehension and reading comprehension is likely to vary across development. Thus, as other researchers have suggested, measuring listening comprehension earlier in development may be useful in predicting future reading comprehension (e.g., Ebert & Scott, 2016; Nation & Snowling, 2004). Capturing the predictive power of listening comprehension may be particularly helpful when children are developing reading skills or may be considered emergent readers, which may be a prolonged process for individuals with DS. Additionally, despite establishing strong correlations between listening comprehension and reading comprehension, the constructs are not perfectly correlated, thus some unexplained variance remains. Although beyond the scope of this project, future research will evaluate how

other variables, such as those illustrated in Scarborough's (2001) reading rope model, contribute to reading comprehension.

Inspection of the MTMMs provides additional information regarding the construct validity of the measures evaluated in this study. For the TD and DS groups, the results reflect high inter-rater reliability (ICC > .5) which is a precursor to further evaluate validity. For the TD group, the pattern of results reflecting high monotrait-heteromethod correlations ($r \geq .5$) were strongly in favor of convergent validity for most of the measures. The results demonstrate that three of the measurement methods converged on a single listening comprehension construct: nonverbal response, cloze procedure, passage-level with open-ended questions and all of the measurement methods converged on a single reading comprehension construct. For the DS group, the pattern of results reflecting high monotrait-heteromethod correlations ($r \geq .5$) were strongly in favor of convergent validity for three listening comprehension and reading comprehension measures. The results demonstrate that three of the measurement methods converged on single listening comprehension and reading comprehension constructs: nonverbal response, cloze procedure, passage-level with open-ended questions. The third column in Tables 9 and 10 indicates the rank ordering of the optimal methods of listening comprehension and reading comprehension. The measures are listed in order of most to least optimal to administer based on minimizing floor effects, minimizing administration time, and whether strong construct validity of the measure was established.

Although the passage-level with close-ended questions (TILLS) listening comprehension and reading comprehension measures were feasible to administer and demonstrated a high degree of inter-rater reliability, inspection of the MTMM provides no evidence of construct validity. The correlations related to indices that were derived for the TILLS measures did not reflect strong construct validity for the TD group (TILLS Listening Comprehension subtest) and was replicated in the DS (TILLS Listening Comprehension and Reading Comprehension subtests) group. Next, we speculate some possible explanations for why the construct validity of

the TILLS measure may be compromised in our study sample. One possible explanation relates to the response format of the TILLS—passage-level text with close-ended questions. Anecdotally, some participants appeared to simply be randomly guessing the correct response given the close-ended or forced choice comprehension questions which in turn is not necessarily a true reflection of their listening comprehension or reading comprehension. Another possible explanation related to the response format is the presence of 'maybe' as a potential answer choice for the close-ended questions. Answering 'maybe' to a comprehension question reflects a certain degree of abstraction, which participants may not have fully understood, despite completing trial items with instructional feedback provided for the 'maybe' response if needed. In the TD group, one possible explanation for why differential evidence of construct validity was observed may be that the TILLS Listening Comprehension measure was more heavily influenced by verbal working memory compared with the TILLS Reading Comprehension measure. We acknowledge that we did not account for a number of other variables that may influence performance on these measures. For instance, perhaps the TILLS measures, when used with children in these age ranges, are more influenced by verbal working memory or place greater demands on the decoding skills of participants when compared with the other measurement methods. Lastly, we hypothesized that the associations related to the TILLS measures may have been moderated by chronological age. However, based on follow-up moderation analyses, the TILLS associations did not vary according to age, though this may be because the sample reflects a truncated age range. Prior tests of the construct validity of the TILLS by the test developers were not restricted to this early elementary period so additional research is needed in this particular age range to further evaluate the validity of this measure for TD children. Additional research evaluating the construct validity of the TILLS for individuals with DS is also warranted.

Group Comparisons

In comparing the TD and DS groups, we aimed to evaluate the association between the percentage of participants scoring above floor level and group membership, as obtaining scores above floor level is oftentimes a challenge for individuals with intellectual disability (Mervis & Robinson, 2005). There was not a significant association between group membership and scoring above floor levels. Thus, the considerations for selecting measures across a range of response and text formats to minimize floor effects proved worthwhile in that there was not a between groups difference in the percentage of participants scoring above floor level. The final group comparison involved evaluating whether the evidence of construct validity was moderated by group. Within the MTMMs, a similar pattern of construct validity was demonstrated between groups, and thus only seven or 25% of associations were moderated by group, though not in any particularly meaningful way. We only evaluated whether the associations were moderated by group because we hypothesized that group membership would capture any other potential moderators given that group differences on any other variables (e.g., nonverbal cognition, grammar comprehension) would be accounted for by group membership.

Limitations

The results of this study should be interpreted with the following limitations in mind. First, the MTMM approach that we used involves a primarily logical rather than analytical approach to guide interpretation of construct validity. Despite this limitation, analyzing and reporting the confidence intervals within the MTMMs enabled us to evaluate the extent to which the various classes of cells differed from one another according to Campbell and Fiske's (1959) guidelines. In addition, we demonstrated sufficient power ($\geq$ .80) to interpret the correlation coefficients within the MTMM. Further analyses using confirmatory factor analysis was not possible in this study, though this approach could be used to evaluate construct validity in a larger sample. Second, we acknowledge that not all individuals with DS demonstrate sufficient word-level

reading and reading comprehension skills necessary to complete the reading and language-related literacy tasks included in this study. In the current study, six individuals with DS did not meet the eligibility criteria to participate in the study due to limited word-level reading abilities. These individuals may have presented with some pre-reading skills, though they did not demonstrate sufficient reading given the established eligibility criteria and thus we cannot draw any conclusions regarding the reading abilities for those non-readers or emergent readers who were not included in the study. Further, the study results should be viewed as a minimal estimate of the feasibility and construct validity of the measures given that convenience sampling was used. In summary, the study results are specific to a particular subset of individuals with DS and the degree to which these results can be generalized for a broader and more representative sample in individuals with DS is unknown.

Third, we identified and analyzed the data using a narrow definition of scoring above floor effects—obtaining a raw score greater than zero. The utility of a broader definition of floor effects in reference to interpreting standard scores and percentiles, rather than raw scores ought to be considered in future research. However, norm-referenced standard scores and percentiles should be interpreted with caution, as research demonstrates that individuals with DS often score within the 1st to 5th percentile on norm-referenced measures when compared with same-age peers (e.g., Mervis & Robinson, 2005). Finally, we did not control for a number of variables that are known to contribute to listening comprehension and reading comprehension in the analyses evaluating whether the evidence of construct validity was moderated by group. As mentioned previously, we hypothesized that any between-group differences that may have moderated the evidence of construct validity would be accounted for in analyzing the effect of group membership. To evaluate this hypothesis, future research should explore the extent to which participant characteristics such as working memory, background knowledge, verbal reasoning, and executive functioning influence the construct validity of various measurement

methods across clinical populations (e.g., Miller, Cutting, & McCardle, 2013; Kintsch & Kintsch, 2005; Perfetti, Marron, & Foltz, 1996; Snow, 2002).

Strengths

Three strengths should be acknowledged. First, no other known studies have considered the DS phenotype to guide the evaluation of the feasibility and psychometric properties of commonly used assessment measures. Thus, the current findings provide a unique contribution to the literature by demonstrating the feasibility and evidence of strong construct validity for a number of measures of listening comprehension and reading comprehension. Second, this study included multiple measurement methods that varied in text and response formats. To adequately capture the relation between listening comprehension and reading comprehension and in line with simple view of reading, we evaluated parallel measures of the two constructs across the various methods (Gough & Tunmer, 1986). The included measures reflect a wide range of measurement methods as well as parallel measurement methods, thus facilitating like-comparisons across the two traits of interest—listening comprehension and reading comprehension. Finally, by including the TD word-level reading matched control group we were able to evaluate group differences regarding feasibility of administration as well as compare the evidence of construct validity within the MTMMs between groups. The TD control group provided a point of reference to guide the interpretation of the DS group data, which ultimately reflected similar patterns of construct validity between the study groups.

Implications

The methods employed in this study address many of the challenges to valid assessment of listening comprehension and reading comprehension for individuals with DS. Although challenges with how listening comprehension and reading comprehension are defined will likely persist, we have clearly operationalized these two constructs as well as selected

measures that align with those definitions. Reading comprehension was operationalized as constructing meaning from written text, and listening comprehension was operationalized as constructing meaning from read-aloud written text. In addition, the various measurement methods included in this study reflect a comprehensive range of methods for assessing listening comprehension and reading comprehension. Lastly, we also considered the DS phenotype in selecting which measures to evaluate. Given that floor effects are often observed when using norm-referenced assessments with individuals with DS, we included measures with a range of text formats so that participants, specifically those with limited word-level reading, would be able to complete at least some initial test items. Further, because individuals with DS often have limited speech intelligibility, we included measures with a range of response formats to limit the impact of poor speech intelligibility on understanding and scoring responses. We demonstrated a high degree of inter-rater reliability for the measurement methods that required a nonverbal or minimal verbal response. Not surprisingly we observed the lowest degree of inter-rater reliability, though still an acceptable value, for the most verbally robust (passage-level text with open-ended question) measurement method in both groups.

Consistent with the rationale for conducting this study, the results provide guidance on the optimal methods (i.e., feasible to collect and valid measures of the construct of interest) for assessing listening comprehension and reading comprehension for individuals with DS and their peers with TD. Overall, the results demonstrate the feasibility and inter-rater reliability of all the measures evaluated. However, strong evidence of convergent validity was only observed for three out of the four measurement methods, with no evidence of discriminant validity for the listening comprehension and reading comprehension constructs. Tables 9 and 10 indicate the rank ordering of the optimal methods of listening comprehension and reading comprehension assessment based on various criteria for the TD and DS groups, respectively. The feasibility results are of practical importance in regards to attempting to minimize floor effects to complete efficient evaluations that yield meaningful information. The construct validity results are of

critical importance in regards to using psychometrically sound assessment measures. One can see that the rank ordering of the measures changes as additional criteria are added. Further, for examiners who may not have experience assessing and making accuracy judgements for individuals with limited speech intelligibility, it may be important to consider the response format in addition to the evidence presented in these tables. Similarly, for test takers with limited reading proficiency, it may also be important to consider the text format; although the implications for selecting a measure based on text format may be captured in the floor effect rankings presented here. Taken together, the study results can help guide listening comprehension and reading comprehension assessment selection for individuals with DS and their peers with TD. By establishing the feasibility, inter-rater reliability, and construct validity of multiple listening comprehension and reading comprehension measurement methods, researchers and clinicians can have greater confidence in using these measures to identify disorders and characterize patterns of strengths and weaknesses.

Future Directions

This initial measurement investigation lays the foundation for developing and evaluating individualized reading interventions for individuals with DS. The current study results that established the feasibility and construct validity of multiple measurement methods, as well as identified the optimal methods of assessment, inform the outcome measure selection for studies of reading intervention for individuals with DS. Future analyses of these data will replicate previous studies (e.g., Laws et al., 2016; Roch & Levorato, 2009; Roch et al., 2011) evaluating the simple view of reading in individuals with DS based on this sample of participants. Kamhi (2007) proposes embracing a narrow view of reading which restricts the scope of reading to word recognition alone. Implications regarding this theoretical model will be explored in light of the findings in individuals with DS. In addition, future research will apply generalizability (G) theory to conduct a decision (D) study to determine the number of measures needed to obtain

stable estimates of listening comprehension and reading comprehension based on the current data of individuals with DS. The results of a G and D study will extend the current findings and enable researchers to ascertain how many of the evaluated (feasible and valid) measures should be administered to adequately capture an individual's listening comprehension and reading comprehension abilities. As mentioned previously, future research is also needed to further evaluate the construct validity of the passage-level with close-ended questions (TILLS) listening comprehension and reading comprehension measures.

Conclusion

The current study contributes to the evidence base regarding the feasibility, reliability, and validity of commonly used measures of listening comprehension and reading comprehension in terms of their utility for individuals with DS. Key findings include demonstrated feasibility of administration for all four measurement methods as well as strong evidence of reliability and construct validity for three out of four of the measurement methods (nonverbal response, cloze procedure, and passage-level with open-ended questions). These results support the use of these measurement methods in clinical practice and future studies of reading comprehension in individuals with DS.

REFERENCES

Afacan, K., Wilkerson, K. L., & Ruppar, A. L. (2018). Multicomponent reading interventions for students with intellectual disability. *Remedial and Special Education*, *39*, 229-242. doi: 10.1177/0741932517702444

American Speech-Language-Hearing Association (n.d.). *Childhood Hearing Screening*. Retrieved May 1, 2019, from https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935406&section=Key_Issues.

Byrne, A., Buckley, S., MacDonald, J., & Bird, G. (1995). Investigating the literacy, language and memory skills of children with Down's syndrome. *Downs Syndrome Research and Practice, 3*, 53-58. doi: 10.3104/reports.51

Byrne, A., MacDonald, J., & Buckley, S. (2002). Reading, language and memory skills: a comparative longitudinal study of children with Down syndrome and their mainstream peers. *British Journal of Educational Psychology*, *72*, 513-529. doi: 10.1348/00070990260377497

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105. doi: 10.1037/h0046016

Carrow-Woolfolk, E. (2014). *Test for Auditory Comprehension of Language-Fourth Edition (TACL-4)*. Austin, TX: Pro-Ed.

Catts, H. W., Fey, M. E., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45,* 1142-1157. doi: 10.1044/1092-4388(2002/093)

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research*, *49,* 278-293. doi: 10.1044/1092-4388(2006/023)

Catts, H. W., Herrera, S., Nielsen, D., & Bridges, M. (2015). Early prediction of reading comprehension within a simple view framework. *Reading and Writing, 28*, 1407–1425. doi: 10.1007/s11145-015-9576-x

Catts, H. W., & Kamhi, A. G. (1986). The linguistic basis of reading disorders: Implications for the speech-language pathologist. *Language, Speech, and Hearing Services in Schools*, *17*, 329-341. doi: 10.1044/0161-1461.1704.329

Catts, H., & Kamhi, A. (1999). Defining reading disabilities. In H. Catts & A. Kamhi (Eds.), *Language and reading disabilities,* (pp. 50-72). Toronto: Allyn & Bacon.

Chapman, R. S., & Hesketh, L. J. (2000). Behavioral phenotype of individuals with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, *6*, 84-95. Doi: 10.1002/1098-2779(2000)6:2<84::AID-MRDD2>3.0.CO;2-P

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cossu, G., Rossini, F., & Marshall, J. C. (1993). When reading is acquired but phonemic awareness is not: A study of literacy in Down's syndrome. *Cognition*, 46, 129-138. doi: 10.1016/0010-0277(93)90016-O

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*, 277-299. doi: 10.1207/s1532799xssr1003_5

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

Edwards, S., Fletcher, P., Garman, M., Hughes, A., Letts, C., & Sinka, I. (1997). *Reynell Developmental Language Scales-Third Edition.* Windsor: NFER-Nelson.

Ebert, K. D., & Scott, C. M. (2016). Bringing the simple view of reading to the clinic: Relationships between oral and written language skills in a clinical sample. *Journal of Communication Disorders*, *62*, 147-160. doi: 10.1016/j.jcomdis.2016.07.002

Elleman, A. M., & Compton, D. L. (2017). Beyond comprehension strategy instruction: What's next? *Language, Speech, and Hearing Services in Schools*, *48*, 84-91. doi: 10.1044/2017_LSHSS-16-0036

Fidler, D. J. (2005). The emerging Down syndrome behavioral phenotype in early childhood: Implications for practice. *Infants & Young Children*, *18*, 86-103. doi: 10.1097/00001163-200504000-00003

Fidler, D., Most, D., & Philofsky, A. (2009). The Down syndrome behavioural phenotype: Taking a developmental approach. *Down syndrome research and practice, 12,* 37-44. Doi: 10.3104/reviews/2069

Fidler, D. J. (2006). The emergence of a syndrome-specific personality profile in young children with Down syndrome. In J. Perera & J. Rondal (Eds.), *Down syndrome: Neurobehavioural specificity* (pp. 139-152). Chichester, West Sussex, England: Wiley.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. Paris & S. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369-394). Mahwah, NJ: L. Erlbaum Associates.

Fuchs, D., Hendricks, E., Walsh, M. E., Fuchs, L. S., Gilbert, J. K., Zhang Tracy, W., Patton, S., Davis-Perkins, N., Kim, W., Elleman, A., & Peng, P. (2018). Evaluating a multidimensional reading comprehension program and reconsidering the lowly reputation of tests of near-transfer. *Learning Disabilities Research & Practice*, *33*, 11-23. doi: 10.1111/ldrp.12162

Fudala, J., & Stegali, S. (2017). *Arizona Articulation Proficiency Scale–Fourth Revision (Arizona–4)*. Los Angeles, CA: Western Psychological Services.

Gough, P. and Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6-10. doi: 10.1177/074193258600700104

Groen, M. A., Laws, G., Nation, K., & Bishop, D. V. (2006). A case of exceptional reading accuracy in a child with Down syndrome: Underlying skills and the relation to reading comprehension. *Cognitive Neuropsychology*, *23*, 1190-1214. doi: 10.1080/02643290600787721

Gray, S., Catts, H., Logan, J., & Pentimonti, J. (2017). Oral language and listening comprehension: Same or different constructs?. *Journal of Speech, Language, and Hearing Research*, *60*, 1273-1284. doi: 10.1044/2017_JSLHR-L-16-0039

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance on listening comprehension. *International Journal of Speech-Language Pathology, 16*, 199–207. doi: 10.3109/17549507.2014.904441

Invernizzi, M., Meier, J.D., Swank, K., & Juel, C. (1997). *Phonological Awareness Literacy Screening (PALS).* Charlottesville, VA: University Printing Services.

Kamhi, A. G., & Catts, H. W. (2017). Epilogue: Reading comprehension is not a single ability—Implications for assessment and instruction. *Language, Speech, and Hearing Services in Schools*, *48*, 104-107. doi: 10.1044/2017_LSHSS-16-0049

Kaufman, A. S. (2004). *Kaufman Brief Intelligence Test–Second Edition.* Circle Pines, MN: American Guidance Service.

Kaufman, A.S. & Kaufman, N.L. (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service. doi: 10.1037/t27677-000

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*, 281-300. doi: 10.1080/10888430802132279

Kim, Y., & Phillips, B. (2014). Cognitive correlates of listening comprehension. *Reading Research Quarterly, 49*, 269–281. doi: 10.1002/rrq.74

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, E., & Kintsch, W. (2005). Comprehension. In S. Paris & S. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369-394). Mahwah, NJ: L. Erlbaum Associates.

Laws, G., Brown, H., & Main, E. (2016). Reading comprehension in children with Down syndrome. *Reading and Writing*, *29*, 21-45. doi: 10.1007/s11145-015-9578-8

Lemons, C. J., & Fuchs, D. (2010). Phonological awareness of children with Down syndrome: Its role in learning to read and the effectiveness of related interventions. *Research in Developmental Disabilities*, *31*, 316-330. doi: 10.1016/j.ridd.2009.11.002

Lemons, C. J., King, S. A., Davidson, K. A., Puranik, C. S., Al Otaiba, S., Fulmer, D., Mrachko, A.A., Partanen, J., & Fidler, D. J. (2017). Developing an early reading intervention aligned with the Down syndrome behavioral phenotype. *Focus on Autism and Other Developmental Disabilities*, *32*, 176-187. doi: 10.1177/1088357615618941

Leslie, L., & Caldwell, J. (2001). *Qualitative Reading Inventory–3.* New York: Addison Wesley Longman.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates–MacGinitie Reading Tests–Fourth Edition.* Itasca, IL: Riverside.

Maenner, M. J., Smith, L. E., Hong, J., Makuch, R., Greenberg, J. S., & Mailick, M. R. (2013). Evaluation of an activities of daily living scale for adolescents and adults with developmental disabilities. *Disability and Health Journal*, *6*, 8-17. doi: 10.1016/j.dhjo.2012.08.005

Martin, N., & Brownell, R. (2011). *Receptive One-Word Picture Vocabulary Test, Fourth Edition (ROWPVT-4)*. Novato, CA: Academic Therapy Publications.

Martin, N., & Brownell, R. (2011). *Expressive One-Word Picture Vocabulary Test, Fourth Edition (EOWPVT-4)*. Novato, CA: Academic Therapy Publications.

Megherbi, H., Seigneuric, A., & Ehrlich, M.F., 2006, Reading comprehension in French 1st and 2nd grade children: Contribution of decoding and language comprehension. *European Journal of Psychology and Education,* 135-147. doi: 10.1007/BF03173573

Mervis, C. B., & Robinson, B. F. (2005). Designing measures for profiling and genotype/phenotype studies of individuals with genetic syndromes or developmental language disorders. *Applied Psycholinguistics*, *26*, 41-64. doi: 10.1017/S0142716405050058

Miller, B., Cutting, L. E., & McCardle, P. (2013). *Unraveling reading comprehension*. Baltimore, MD: Brookes Publishing.

Nash, H., & Heath, J. (2011). The role of vocabulary, working memory and inference making ability in reading comprehension in Down syndrome. *Research in Developmental Disabilities*, *32*, 1782-1791. doi: 10.1016/j.ridd.2011.03.007

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*, 342-356. doi: 10.1111/j.1467-9817.2004.00238.x

Nelson, N., Plante, E., Helm-Estabrooks, N., & Hotz, G. (2015). *Test of Integrated Language and Literacy Skills (TILLS)*. Baltimore, MD: Brookes Publishing.

Paris, S. G., & Stahl, S. A. (2005). *Children's reading comprehension and assessment*. Mahwah, NJ: L. Erlbaum Associates. doi: 10.4324/9781410612762

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 137-166). Routledge.

Petersen, D. B., & Stoddard, A. (2018). Psychometric requirements of oral and written language progress monitoring assessments. *Perspectives of the ASHA Special Interest Groups*, *3*, 180-197. doi: 10.1044/persp3.SIG1.180

Roch, M., Florit, E., & Levorato, C. (2011). A follow-up study on reading comprehension in Down syndrome: the role of reading skills and listening comprehension. *International Journal of Language & Communication Disorders*, *46,* 231-242. doi: doi.org/10.3109/13682822.2010.487882

Roch, M., & Levorato, M. C. (2009). Simple view of reading in Down's syndrome: The role of listening comprehension and reading skills. *International Journal of Language & Communication Disorders*, *44*, 206-223. doi: 10.1080/13682820802012061

Roizen, N. J., Wolters, C., Nicol, T., & Blondis, T. A. (1993). Hearing loss in children with Down syndrome. *The Journal of Pediatrics*, *123*, S9-S12. doi: 10.1016/S0022-3476(05)81588-4

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis) abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97-110). New York: Guilford Press.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Oral Language*. Rolling Meadows, IL: Riverside.

Semel, E., Wiig, E., & Secord, W. (2000). *Clinical Evaluation of Language Fundamentals—3UK*. London: Pearson Assessment.

Sinatra, G. M. (1990). Convergence of listening and reading processing. *Reading Research Quarterly*, 115-130. doi: doi.org/10.2307/747597

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation. Washington, DC: RAND Education.

Sticht, T. G., Beck, L., Hauke, R. ,Kleiman, G., & James, J. (1974). *Auding and reading: A developmental model.* Alexandria, VA: Human Resources Research Organization. doi: 10.1037/e457482004-001

Stoel-Gammon, C. (1997). Phonological development in Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, *3*, 300-306. doi: 10.1002/(SICI)1098-2779(1997)3:4<300::AID-MRDD4>3.0.CO;2-R

van Wingerden, E., Segers, E., van Balkom, H., & Verhoeven, L. (2018). Cognitive constraints on the simple view of reading: A longitudinal study in children with intellectual disabilities. *Scientific Studies of Reading*, *22*, 321-334. doi: 10.1080/10888438.2018.1446435

Wechsler, D. (2009). *Wechsler Individual Achievement Test–Third Edition.* San Antonio, TX: Psychological Corporation.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*, 89-106. doi: 10.1177/001440291107800106

Wiederholt, L., & Bryant, B. (1992). *Gray Oral Reading Test-Third Edition*. Austin, TX: Pro-Ed.

Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical evaluation of language fundamentals-Fifth edition (CELF-5)*. Bloomington, MN: NCS Pearson.

Wishart, J. G. (1996). Avoidant learning styles and cognitive development in young children. In B. Stratford & P. Gunn (Eds.), *New approaches to Down syndrome* (pp 173-205). London: Cassell.

Woodcock, R. W. (2011). *Woodcock Reading Mastery Tests-Third edition (WRMT-III)*. Pearson. doi: 10.1037/t15178-000

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement*. Itasca, IL: Riverside.