UCONN LIBRARY

University of Connecticut
OpenCommons@UConn

Master's Theses | University of Connecticut Graduate School

5-7-2011

The Effect of Test Revision: Comparing the Performance of Preschool Children with SLI and Typical Controls on the PPVT-III and the PPVT-IV

Sabrina E. Jara
University of Connecticut - Storrs, sabrina.hosmer@gmail.com

Recommended Citation
Jara, Sabrina E., "The Effect of Test Revision: Comparing the Performance of Preschool Children with SLI and Typical Controls on the PPVT-III and the PPVT-IV" (2011). *Master's Theses*. 89.
https://opencommons.uconn.edu/gs_theses/89

This work is brought to you for free and open access by the University of Connecticut Graduate School at OpenCommons@UConn. It has been accepted for inclusion in Master's Theses by an authorized administrator of OpenCommons@UConn. For more information, please contact opencommons@uconn.edu.

THE EFFECT OF TEST REVISION: COMPARING THE PERFORMANCE OF PRESCHOOL CHILDREN WITH SLI AND TYPICAL CONTROLS ON THE PPVT-III AND PPVT-IV

Sabrina Elizabeth Jara

B.A., University of Connecticut

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

at the

University of Connecticut

2011

APPROVAL PAGE

Master of Arts Thesis

THE EFFECT OF TEST REVISION: COMPARING THE PERFORMANCE OF PRESCHOOL CHILDREN WITH SLI AND TYPICAL CONTROLS ON THE PPVT-III AND PPVT-IV

Presented by

Sabrina Elizabeth Jara, B.A.

Major Advisor_____
Dr. Tammie J. Spaulding

Associate Advisor_____
Dr. Bernard Grela

Associate Advisor_____
Dr. Frank Musiek

University of Connecticut
2011

ACKNOWLEDGMENTS

I would first like to acknowledge Dr. Tammie J. Spaulding for her guidance as my advisor and professor. Her fervor for research, unmatched patience, and enthusiasm in training new clinicians such as myself fueled my interest in attending UConn's graduate program in speech-language pathology. With her help I was able to carry this project to completion and gained a strong appreciation for research as a result.

I also appreciate the time and effort taken by Dr. Bernard Grela and Dr. Frank Musiek as members of my thesis committee. Their knowledge and experience has helped me gain new insight into my project and strengthen my resolve to seek publication.

I would also like to thank the people of the Language Lab. Lab managers Calli Schechtman and, previously, Beverly Collisson, are proven leaders who organized our team, conferenced with school administrators, and worked with the preschool participants right alongside everyone else. I will miss belonging to the lab and, of course, its wonderful graduate and undergraduate assistants who ended up screening over 200 participants in total. Your commitment to the team made this project possible.

Finally, I will always be grateful to my family for their unwavering support. My parents emphasized a strong foundation of love and support ever since I can remember, and I will strive to carry it on to the next generation. My sister is an inspiration to me every day because she works harder than anyone I know. I am also fortunate to have found my best friend and fiancé Seth Hosmer, who is a better partner in mind and spirit than I would have thought possible.

TABLE OF CONTENTS

ABSTRACT

There are numerous assessments available for evaluating the language skills of children with specific language impairment (SLI). Given the substantial body of research identifying word learning deficits in this population of children (e.g., Gray, 2004; Oetting, Rice, & Swank, 1995; Paul, 1995; Rescorla, Roberts, & Dahlsgaard, 1997), norm-referenced assessments which assess receptive vocabulary may be useful for diagnostic purposes. The Peabody Picture Vocabulary Test is the most widely used assessment of receptive vocabulary for children with language impairment, as evidenced by both clinical report and research investigations (e.g. Betz, Sulllivan, & Eickhoff, 2010; Preston & Edwards, 2010; Evans et al., 2009). Given the inadequate diagnostic utility of the PPVT-III for identifying presence or absence of language impairment in preschool children (Gray, Plante, Vance, & Henrichsen, 1999), it was important to determine if this was improved for the most recent edition of this test, the PPVT-IV. This study compared the performance of preschool children with SLI and controls on the PPVT-III and PPVT-IV to determine the effect of test revision on identification of language impairment. A secondary purpose was to determine if children performed consistently on these two tests, as this would provide empirical evidence for readily substituting one for the other in both clinical and research practice.

Methods. Forty preschool children, 20 with SLI and 20 typically-developing (TD) controls, formed the exploratory sample. Children in the SLI and TD groups were matched for age, sex, and socioeconomic status. In order to determine the generalizability of the results to a new sample, a confirmatory sample was obtained. The

confirmatory sample was composed of 5 children with SLI and 20 TD controls. All participants were administered both the PPVT-III and the PPVT-IV.

Analysis. A MANOVA was conducted with Group (SLI, TD) as the between-subjects variable and Version (PPVT-III, PPVT-IV) as the within-subjects variable. The dependent variable was standardized test scores. Discriminate analyses were also conducted to identify the maximum discriminate accuracy of each test version and corresponding standard score cut-offs.

Results. A significant group effect was found between the experimental and the control group. Children with SLI performed significantly worse than TD peers on both test versions, although they performed well-within 1SD of the mean (standard score of 93.55 on the -III and 94.15 on the -IV). There was no version effect, meaning that on average, there was no difference in performance between the PPVT-III and PPVT-IV. No group x version effect existed either, meaning that the difference in performance between the PPVT-III and PPVT-IV was similar for both groups of children. However, an individual differences analysis found that 35% of children performed differently on the PPVT-III and -IV, 8/20 in the SLI group and 6/20 in the TD group. Half the children performed better on the PPVT-III while the remaining half performed better on the PPVT-IV. Discriminate analyses revealed an optimal cut-off of 103 for both tests. Using this cut-off, sensitivity of both remained consistent at 80% while the specificity dropped from 75% on the PPVT-III to 70% on the PPVT-IV in both the exploratory and confirmatory groups. Posterior probability analysis indicated that none of the misclassified children were strongly misclassified.

Discussion. The differences in performance between the two test versions for a subset of children suggests that clinicians and researchers should not consider the two test versions as interchangeable for determining impairment, for documenting change, or for other purposes. The lower diagnostic accuracy of the PPVT-IV relative to the PPVT-III highlights the need to avoid assuming newer versions are superior to older in identifying presence or absence of language impairment. Furthermore, the high cutoff for maximizing diagnostic accuracy provides further support that children with SLI are unlikely to score as low as clinicians may expect on norm-referenced tests. Both clinicians and researchers should approach tests, including newer versions, in a critical manner and evaluate evidence supporting their diagnostic utility if they are to be used for this purpose. Empirical evidence to date does not support the use of the PPVT-III nor the PPVT-IV for diagnosing language impairment in preschool children.

INTRODUCTION

The publication of newer versions of norm-referenced assessments is commonplace, and tests of child language are no exception. Speech language pathologists frequently use tests of child language to assist in determining if a child presents with a language impairment. While a newer version of a norm-referenced assessment may be developed to reflect more recent norms (Johnson, 1995; McFadden, 1996) or in response to academic and clinical criticism of the prior version (Adams, 2000), it is important to determine if the newer version is superior to the old for the purpose in which it is intended (Bush, 2010). Previous research has suggested that this may not be the case for the identification of language deficits in children with specific language impairment (SLI) (Ballantyne, Spilkin, & Trauner, 2007). This study compared the performance of preschool children with and without SLI on the Peabody Picture Vocabulary Test-Third Edition (PPVT-III; Dunn & Dunn, 1997) and the Peabody Picture Vocabulary Test-Fourth Edition (PPVT-IV; Dunn & Dunn, 2007), to determine whether the most recent version is superior to the prior for identifying presence and absence of language impairment in young children.

Speech language pathologists have many assessments of child language available for selection to assist in this process. A survey of school-based clinicians in California found that the clinicians reported using 59 different tests for the diagnosis of language disorders in children ages 4-9 at least once (Wilson, Blackmon, Hall, & Elcholtz, 1991). This indicates that a wide variety of tests are used in clinical practice for assessment of child language alone. The survey also found that the vast majority (263 of 266) of speech language pathologists used at least one norm-referenced test as part of their assessment of

children's language functioning. In contrast, a more recent survey found that school-based speech language pathologists in Michigan were less likely to use norm-referenced tests than informal procedures (Caesar & Kohler, 2009). However, the results of these and another more recent survey (Betz, Sullivan, & Eickhoff, 2010) found that the Peabody Picture Vocabulary Test was the most-widely used vocabulary measure for children. In fact, Betz et al. (2010) found that the Peabody Picture Vocabulary Test was the third most commonly employed norm-referenced tests that clinicians used for the diagnosis of children with SLI.

Peabody Picture Vocabulary Tests

The PPVT was first developed in 1959 and was subsequently revised three times. The third and fourth editions, the subject of this investigation, resemble one another in terms of presentation. These surface similarities include the use of four drawings per page in which one corresponds to the target word, repetition of a majority of stimulus items, and brevity of administration (11-12 minutes). In contrast to its predecessor, the PPVT-IV features full color illustrations, a larger physical display, a normative sample of increased size, and updated items (e.g., the target word "typewriter" was replaced by "computer").

Importantly, the PPVT-III remains relevant due to its noted popularity among both clinicians (Caesar & Kohler, 2009) and researchers. The PPVT-III is frequently used as part of an assessment battery when investigating children with documented language difficulties including SLI, autism spectrum disorder, and dyslexia (e.g., Condouris, Meyer, Tager-Flusberg, 2003; Farrar, Johnson, Tompkins, et al., 2009; Gray, 2004; Wise, Sevcik, Morris, et al., 2007). In addition, the PPVT-III is frequently employed as part of

the participant matching criteria when attempting to equate children on receptive vocabulary knowledge (e.g., Seiger-Gardener & Brooks, 2008, Silliman, Diehl, Bahr, et al., 2003; Sutherland & Gillon, 2005). The PPVT-III is also used in longitudinal studies documenting vocabulary growth (e.g., Hart, Petrill, & Dush, 2010; Rvachew, Chiang, & Evans, 2007).

Although there is no independent evidence to document the usefulness of Peabody Picture Vocabulary Test-Fourth Edition (PPVT-IV; Dunn and Dunn, 2007) for children with language disorders, it is gaining popularity in the current literature. In addition, a number of research investigations have included the PPVT-IV as part of their assessment battery with children. It has been used to document cognitive ability (e.g., Cutuli, Herbers, Rinaldi et al., 2010; Lam, Mahone, Mason et al., 2010), to measure verbal ability in general (Meador, Baker, Browning et al., 2011), and to describe receptive vocabulary skills (e.g., Alt, 2011; Hanson, Nasir, & Fong, 2010; Kulkofsky, 2010).

The motivation to critically examine the performance of children with and without SLI on both test versions arises from the popularity of the Peabody Picture Vocabulary Tests and previous investigations evaluating their lack of diagnostic utility, despite their popularity, with individuals with language-based disorders. The PPVT-III in particular has faced scrutiny regarding its utility in diagnosing language impairment. In comparison to its predecessor, the PPVT-R, children as well as adults with language-based learning disabilities both obtained significantly higher scores on the PPVT-III relative to the prior version (Williams, 1998; Pankratz, Morrison, & Plante, 2004). The developers of the PPVT-III (Dunn & Dunn, 1997) note this increase in the PPVT-III and provide a conversion table to convert a raw score from the PPVT-R to a raw score

equivalent on the PPVT-III. Ukrainetz and Duncan (2000) indicate, however, that even with this adjustment the standard score for the PPVT-III remains higher.

Pankratz et al. (2004) found that the elevated scores on the PPVT-III relative to the PPVT-R actually diminish the diagnostic accuracy for differentiating between adults with and without language-based learning disorders when the PPVT-III replaces the PPVT-R as part of a battery of language. The adults were identified with a battery of assessments Although no specific comparisons of the PPVT-R and PPVT-III have been conducted for children with SLI, Gray et al. (1999) found that the diagnostic accuracy of the PPVT-III for preschool children was modest at best, with 74% sensitivity and 71% specificity. Given the differences in performance between the PPVT-R and PPVT-III for children in general and for adults with language impairment, differences in performance may also be apparent between the PPVT-III and PPVT-IV. Unlike the PPVT-III, there is no conversion table identified in the manual for adjusting scores between the two tests, suggesting that scores are likely expected to be comparable.

Comparisons between the PPVT-III and IV are important to consider in order to evaluate the interchangeability of these two test versions. The manual provides data to indicate that scores on the PPVT-III and PPVT-IV are not significantly different for a sample of 322 children, including children of preschool-age. In addition to a lack of significant mean differences, the manual of the PPVT-IV also reports correlational analyses for the different age groups. For the purposes of this investigation, the correlations identified for children between the ages of 3-5 years were of interest. There were strong positive correlations identified between these two test versions, specifically .82 and .83 for children aged 2-4 years, and 5-6 years respectively. Despite these high

correlations, clinicians and researchers should be cautious in considering that the PPVT-III and PPVT-IV are interchangeable because, based on information provided in the test manual, between 33% and 31% of the variance is still unaccounted. In addition, no information is provided about the language functioning of the sample who were administered both the PPVT-III and PPVT-IV. Importantly, differences in performance between these two test versions may be apparent for clinical populations, including children with specific language impairment (SLI).

Vocabulary Acquisition in Children with SLI

A definition of SLI is the presence of language impairment in the absence of hearing difficulties, cognitive impairment, psychological or frank neurological disorders (Leonard, 1998). Based on an epidemiological study by Tomblin, Records, Buckwalter, et al. (1997), roughly 7% of children exhibit this disorder. One challenge in identifying children with this disorder is that heterogenous profiles of language skills result in the same diagnosis of SLI. The linguistic difficulties of children with SLI are typically characterized by deficits in morphosyntax development (e.g., Grela & Leonard, 1997; Reilly, Losh, Bellugi, et al., 2004; Rice, Wexler, & Cleave, 1995). Therefore it is no surprise that studies investigating preschool children with SLI have noted high diagnostic accuracy on tests which assess morphosyntax skills. These include the Patterned Elicitation Syntax Test (Merrell & Plante, 1998; Young & Perachio, 1993), Test for Examining Expressive Morphology (Merrell & Plante, 1998; Shipley, Stone, & Sue, 1983), and different versions of the Structured Photographic Expressive Language Test (SPELT-P2: Dawson, Eyer, & Fonkalsrud, 2005; Greenslade, Plante, Vance, 2009; SPELT-3: Dawson, Stout, & Eyer, 2003; Perona, Plante, Vance, 2005: SPELT-2: Plante &

Vance, 1994; Werner & Krescheck, 1983).

A number of investigations have also identified word learning deficits in children with this disorder (e.g., Alt, 2011; Alt, Plante, Creusere, 2004; Alt & Plante, 2006; Gray, 2003; Gray, 2004; Gray et al., 1999; McGregor, Newman, Reilly, et al., 2002; Nash & Donaldson, 2005; Storkel & Rogers, 2000). Compared to age-matched typically developing peers, children with SLI exhibit slower vocabulary growth (Paul, 1995; Rescorla et al., 1997). In both fast mapping and quick incidental learning tasks, children with SLI learn fewer novel words than their peers (Alt, 2011; Alt, Plante, Creusere, 2004; Gray, 2004; 2006; Oetting, Rice, & Swank, 1995; Rice, Cleave, & Oetting, 2000; Rice, Oetting, Marquis et al., 1995). Consequently, it is no surprise that they also exhibit smaller lexicons (Gray, 2006; McGregor et al., 2002, Watkins, Kelly, Harbers, et al., 1995). Therefore, the diagnostic utility of available tests of vocabulary skills is also of interest when investigating this population of children.

Evidence Needed to Support a Test's Diagnostic Utility

Prior to using recent editions of norm-referenced tests, including the PPVT-IV, for determining presence or absence of language impairment, evidence in support of a test's ability to determine who is and who is not impaired needs to be established empirically. A test's diagnostic accuracy is its ability to accurately identify impaired language development as impaired and its ability to accurately identify non-impaired language development as not impaired. Sensitivity refers to... while specificity is... Ultimately, acceptable levels of sensitivity and specificity should depend on clinician's personal preferences (de Beaman, Beaman, & Garcia-Peña, 2004; Emmons & Alfonso, 2005). However, several researchers have adopted the recommended cut-offs of Plante and

Vance (2004), who consider 80-89% sensitivity and specificity to be “fair” diagnostic accuracy and 90-100% sensitivity and specificity to be “good” diagnostic accuracy (Gray et al., 1999; Greenslade, Plante, & Vance, 2009; Jessup, Ward, Cahill, et al., 2008; O'Neill, 2007; Restrepo, 1998).

A norm-referenced test’s diagnostic accuracy is dependent on the cut-off score used to determine whether or not a child presents with a language impairment. A cut-off score is the standardized score used to differentiate between typically developing children and children with SLI. With respect to a test’s diagnostic accuracy, children who score above the cut-off score are classified as non-language impaired (or typically developing language), while children who score below the cut-off score are classified as language impaired.

Positive and negative likelihood ratios can be calculated from sensitivity and specificity data. Similar to sensitivity and specificity, likelihood ratios depict the amount of confidence that a norm-referenced test score distinguishes between individuals who test positive, in this case have a language impairment, and individuals who test negative, in this case do not have a language impairment. In other words, a positive likelihood ratio signifies the amount of confidence that test scores identify disordered individuals correctly and a negative likelihood ratio equates with the amount of confidence that a test score identifies typically developing individuals correctly. Dollaghan (2004) suggested using likelihood ratios, rather than sensitivity and specificity, because they are less reliant on the sample from which the sensitivity and specificity data are derived. Dollaghan further recommended that acceptable diagnostic accuracy translates to positive likelihood ratios (sensitivity/(1-specificity) greater than 10 and negative likelihood ratios (1-

sensitivity/specificity) of less than 0.2.

In contrast to positive and negative likelihood ratios which, like sensitivity and specificity, report a general degree of confidence with respect to a test's ability to differentiate impaired from unimpaired children, posterior probabilities determine the confidence associated with each individual child's impaired or unimpaired classification. Posterior probabilities are particularly useful to clinicians for determining the amount of confidence that should be placed with an individual child's language status classification derived from the assessment. To date, sensitivity and specificity are typically of focus in the research literature when describing the diagnostic accuracy of tests for children with language impairment, although posterior probabilities may be mentioned (e.g., Merrell & Plante, 1997; Pankratz, Vance, & Insalaco, 2007; Perona, Plante, & Vance, 2005; Plante & Vance, 1994; Spaulding, Plante, & Farinella, 2006). This is likely because studies investigating the diagnostic accuracy of assessments for differentiating between children with and without SLI tend to emphasize group-level analyses. Researchers across the medical and social sciences are strongly advocating for a more widespread adoption of posterior probability in lieu of sensitivity and specificity due to its higher degree of accuracy beyond an individual study (see Diamond and Forester, 1983; Chapman, Mapstone, Porsteinsson, et al., 2010). An additional benefit of posterior probabilities is that their child-specific, as opposed to group level focus, facilitates critical diagnostic decisions which clinicians typically make on an individual child basis.

Current Evidence to Support Diagnostic Utility of Tests for Children with SLI

There is no gold standard for the diagnosis of children with SLI. A test that would meet this qualification would be able to accurately identify this population with 100%

accuracy. This would mean that cultural differences would be accounted for, error would be non-existent, and there would be no grounds upon which to question the final diagnosis. However, there are no definitive tests in the social sciences due to the abstract nature of human behavior and the wide range of individual variation. This is particularly true for children with SLI, who by nature of their definition, represent a very heterogeneous population. Despite the gold standard, norm-referenced tests are often used to assist in determining whether or not a child presents with a language impairment (Betz et al., 2010).

Speech language pathologists may feel pressured to select the most recent version of a test to evaluate children suspected of having a language impairment. Consequently, independent evidence of a newer version's superiority over the prior version for diagnosing presence or absence of impairment is needed to justify new test adoption. Only one study to date has compared the diagnostic utility of two versions of a test for diagnosing presence or absence of language impairment in SLI, and this study was conducted on school-age children. Ballantyne et al. (2007) compared the ability of the CELF-R (Semel, Wiig, & Secord, 1987) and CELF-III (Semel, Wiig, & Secord, 1995) to diagnose language impairment in children. Typically developing children, children with SLI, and children with focal brain damage all exhibited higher mean scores on the newer version of this test. With respect to children with SLI specifically, those rated in the moderately to severely impaired range on the CELF-R were classified as exhibiting mild to moderate language impairment on the CELF-III. Importantly, many children with SLI who would have been identified as needing language intervention if given the CELF-R would be less likely to receive services if judgments were based on the CELF-III. As this

research suggests, clinicians should critically evaluate these tests based on empirical evidence.

Historically, vocabulary assessments in particular have had only modest utility for diagnosing language impairment in children. As stated previously, Gray et al. (1999) found that the PPVT-III exhibited 74% sensitivity and 71% specificity for determining language impairment in preschool children (four and five years old) with and without SLI. Therefore, the results of this study will help to determine whether the PPVT-IV offers improved diagnostic accuracy for determining presence or absence of language impairment relative to its predecessor.

The Present Study

In sum, both researchers and clinicians are confronted with test revisions on a regular basis. Confidence in adopting new assessments, including newer versions, depends on a variety of factors. However, if the purpose of administering a norm-referenced assessment is to identify whether or not a child presents with a language impairment, then empirical evidence of the test's diagnostic accuracy must be evaluated. Given that research has identified differences in performance on the PPVT-III and PPVT-R for children and adults with language disorders, it was important to determine whether similar differences were apparent on the fourth edition relative to the third edition for individuals with language impairment. In addition, because diagnostic accuracy is insufficient for preschool children with SLI on the PPVT-III (Gray et al., 1999), it is important to determine whether or not it improves for children of this age on the newer edition.

The purpose of this investigation was not to identify which test version, the

PPVT-III or PPVT-IV is a more accurate reflection of receptive vocabulary; rather it was to determine whether there is a difference in performance for children with SLI and TD children between the two test versions. The second purpose was to determine each test version's ability to discriminate between children with SLI and TD peers. This test was chosen because the Peabody Picture Vocabulary Test has historically been widely adopted in clinical practice (Betz et al., 2010; Caesar & Kohler, 2009; Wilson et al., 2001). The population of interest, children with SLI, was investigated because children with SLI represent a large percentage of a speech language pathologist's caseload (see Tomblin et al., 1997). The diagnostic utility of these tests was explored for preschool-age children because many children with SLI are identified as language impaired during the preschool years (Scarborough, 1990, van der Lely & Marshall, 2010). The specific research questions are:

1. Do preschool children in general score differently on the PPVT-IV as compared to the PPVT-III?
2. Is there a difference in performance between the PPVT-III and PPVT-IV for typically developing (TD) preschool children?
3. Is there a difference in performance between the PPVT-III and PPVT-IV for preschool children with SLI?
4. What is the diagnostic accuracy of the PPVT-III?
5. What is the diagnostic accuracy of the PPVT-IV?

METHODS

Participants

The participants in this investigation were recruited from local pre-kindergarten classrooms and daycare centers and completed this study in their respective school or home settings. The exploratory group consisted of forty preschool-age children. Twenty children formed the SLI cohort. They ranged in age from 43 months to 63 months, with a mean age of 51.85 (SD=)months. The 20 remaining children served as typically developing controls. They ranged in age from 45 months to 64 months, with a mean age of 52.65 (SD=) months. The participants were matched for age (+/- 3 months), sex, and socioeconomic status (+/- 3 years maternal education level). The confirmatory group consisted of 5 children with SLI and 20 typically developing peers. The participants in both the exploratory and confirmatory groups represented a variety of racial, ethnic, and socioeconomic backgrounds. See Table 1 for a description of the demographic characteristics of the participants.

Additional participant characteristics were extracted from teacher/caregiver and parent reports. All children were monolingual native English speakers. None were diagnosed with physical or psychological disorders, including attention-deficit/hyperactivity disorder. No developmental concerns, with the exception of communication concerns for the SLI group, were noted. See Table 2 for performance on norm-referenced assessments.

Inclusionary criteria for all children in this study included passing hearing and colorblind vision screenings as well as ruling out intellectual disability by virtue of a nonverbal cognitive assessment. Hearing was screened at 25 dB HL for 500 Hz and at 20

dB HL for 1000, 2000, and 4000 Hz in each ear (ANSI, 1989). The presence of colorblindness was an exclusionary criterion for participation in this study because a major difference between the two assessments under evaluation was that one presented the stimuli in black and white pictures while the other presented the stimuli in color. The criterion for passing was correct identification of 8 out of 9 stimulus items during initial administration or 9 out of 9 items during subsequent administration on the Color Vision Testing Made Easy, color vision test (Waggoner, 2002). In addition to ruling out sensory problems, it is essential to measure nonverbal intelligence because intellectual disability is an exclusionary criterion for both the typically developing and SLI diagnoses. Therefore, all participants completed and obtained a standard score of 75 or higher on the Nonverbal Cognition Index of the Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004).

The participants' language functioning was evaluated using a multi-method approach, a combination of norm-referenced testing, parent and teacher/caregiver report, and clinical judgment by a certified speech language pathologist. Parents and/or teachers/caregivers of children in the SLI group indicated concerns regarding their language development. In addition, children in the SLI obtained a standard score below 85 on the Core Language Scale of the Clinical Evaluation of Language Fundamentals – Preschool, Second Edition (CELF-P2; Wiig, Secord, and Semel, 2004). Based on information provided in the CELF-P2 examiner’s manual, this cutoff score results in 85% sensitivity for the identification of language impairment in preschool children. Confirmation of language impaired status was obtained through clinical judgment of impaired language skills based on a sample of the participants’ conversational speech.

Parents and teachers/caregivers of children in the TD group reported no concerns regarding the children's development and indicated that the children had no history of special education or related services. In addition, children in the TD group obtained a standard score of 85 or above on the CELF-P2, representing a specificity of 82% according to information provided in the test's manual.

Materials

Peabody Picture Vocabulary Test, Third Edition (PPVT-III; Dunn & Dunn, 1997)

The PPVT-III is a measure of receptive vocabulary knowledge and a screening test of verbal ability. It was standardized on a sample of 2000 children and adolescents and over 700 adults, and is intended for use with individuals between 2 years to 90+ years. The PPVT-III is composed of 204 stimulus words and takes on average of 11-12 minutes to complete. The examiner speaks a prompt, and the examinee has to determine which of four black-and-white pictures best represents the stimulus. Individuals are asked to point to the picture that best matches the stimulus word presented. Form A was used exclusively to ensure that the all children were exposed to the same stimulus words.

Peabody Picture Vocabulary Test, Fourth Edition (PPVT-IV: Dunn & Dunn, 2004)

The purpose for the PPVT-IV was also to measure receptive vocabulary and as a screening tool for verbal ability. The PPVT-IV was standardized on a larger population (3540 individuals) than the PPVT-III representative of the U.S. population as measured by the 2004 census in terms of age, sex, racial diversity, socioeconomic status and geographic region. The PPVT-IV contains 228 stimulus words. Although some words are identical to the previous version's stimuli, others are not and additional words are included. In addition, the picture choices are presented in color, which contrasts with the

black and white picture choices on the older, PPVT-III. Additional changes include the replacement of picture choices to exclude outdated technology represented on the earlier version. Otherwise the presence of two forms and the test administration procedures are consistent with the prior version. In this study, Form A of both versions was utilized.

Procedures

Each version of the PPVT was administered to participants on separate days. For the SLI and TD children in the exploratory sample and the TD group in the confirmatory sample, half of the participants in each group received the PPVT-III first and the PPVT-IV second while the remaining half completed the PPVT-IV first and the PPVT-III second. Given the odd number of participants in the SLI group of the confirmatory sample, three children received the PPVT-IV and then the PPVT-III while two received the PPVT-III followed by the PPVT-IV.

Test administration took place in a quiet, isolated area of the children's preschool, daycare, or home settings. All children were tested individually by undergraduate and graduate students thoroughly trained on test administration and scoring procedures. Children received stickers after the completion of each test. At the end of their participation in the study, each child selected two small prizes.

Ten percent of the norm-referenced tests administered were double scored by trained undergraduate students for reliability purposes. One examiner would administer, record responses, and make correct/incorrect item judgments while another individual would record responses and make correct/incorrect judgments only. Point-to-point reliability for individual items was calculated to be .95. Rare discrepancies were resolved by inter-rater discussion and reference to the test manual for scoring procedures.

RESULTS

Correlational Analyses

Given that the PPVT-III and the PPVT-IV are both intended to measure receptive vocabulary and reflect similar content, a high statistical correlation was expected between these test versions. Calculation of Pearson's product moment correlation for all participants resulted in a statistically significant association between performance on the PPVT-III and performance on the PPVT-IV, r = .886, $p<.001$ (two-tailed), with a 95% confidence interval of .794 to .938. These results indicate that approximately 78% of the variance in either measure can be accounted for by the other measure. This leaves 22% of the variance unaccounted. The correlation between the PPVT-III and PPVT-IV is slightly higher than the correlation presented in the PPVT-IV test manual of r =.79 for children aged 2-4 years. However, it is within the expected range for the r-value of .82 reported in the manual for children between the ages of 5-6 years. Children in this investigation overlapped with both of these age groups.

Separate correlational analyses were also conducted for each group. Calculation of Pearson's product moment correlation for the TD group resulted in a statistically significant association between performance on the PPVT-III and performance on the PPVT-IV, r = .871, $p<.001$ (two-tailed), with a 95% confidence interval of .698 to .948. A similar correlational analysis for the SLI group resulted in a statistically significant association as well, r = .715, $p<.001$ (two-tailed), with a 95% confidence interval of .399 to .879. These results indicate that 76% and 51% of the variance in performance in one version can be accounted for by the other version for both the TD and SLI groups respectively.

PPVT-III versus PPVT-IV Differences

To examine differences in performance on these two assessments, a Mixed ANOVA was conducted with Group (TD, SLI) as the between subjects factor and Test Version (PPVT-III, PPVT-IV) as the within subjects factor. A Mixed ANOVA is a repeated measures analysis used to determine if there are differences in performance across two or more groups. The dependent variable was test performance based on standard scores. The performance of both groups on the PPVT-III and PPVT-IV is displayed in Figure 1.

The mean standard score on the PPVT-III was 112.50 (*SD* = 11.52) and 93.55 (*SD* = 8.77) for the TD group and SLI group respectively. The mean standard score on the PPVT-IV was 113.15 (*SD* = 12.60) for the TD group and 94.15 (*SD* = 11.55) for the SLI group. The mean of the TD group on both the PPVT-III and PPVT-IV was higher than the tests' normative samples, likely because the normative samples of both tests included disordered subjects. In addition, the mean of the SLI group on both of these tests was within 1 *SD* of the normative samples' mean which is consistent with that reported in the respective test manuals.

The results of the Mixed ANOVA revealed a main effect of Diagnosis, $F(1,38) = 32.01$, $p < .001$, $n^2_p = .457$. The SLI group performed significantly worse overall relative to the TD group. There was no Test Version effect, $F(1, 38) = .30$, $p = .587$, $n^2_p = .008$ indicating that, when the participants were combined, there was no difference in performance between the PPVT-III and PPVT-IV. There was also no Test Version x Group effect, $F(1,38) = .00$, $p = .983$, $n^2_p = .000$, indicating that the difference in performance between the PPVT-III and PPVT-IV was similar for both groups of children.

In addition to the Mixed ANOVA, which analyzes the data using mean performance as a reference, an individual difference approach was taken to document variation in scores between the PPVT-III and PPVT-IV for each participant in this investigation. This is displayed in Figure 2. for the TD group and Figure 3. for the SLI group. Although there was a high inter-test version correlation and no differences in performance of either group for the test versions administered based on the ANOVA results, the individual data suggests test version variability in performance for some children. However, some variability in performance can be expected from one test administration to another. Therefore, score differences between the two test versions for each individual child relative to the standard error of measure (SEM) reported in each test's manual were compared. The range of scores within 1 standard error of measure overlapped for 14 out of 20 participants in the TD group. Therefore, 6 out of 20 TD children's scores fell outside the 1 SEM range, indicating that their test scores were independent. Two of these children exhibited a higher score on the PPVT-III and the remaining four obtained a higher score on the PPVT-IV. For the SLI group, the range of scores within 1 SEM overlapped for 12 out of 20 participants, indicating that 8 children with SLI exhibited independent test version performance. Five of these children presented with higher scores on the PPVT-III and the remaining three exhibited higher scores on the PPVT-IV.

Diagnostic Accuracy

Exploratory group results. This investigation also examined each test version's ability to differentiate the two groups of children based on their respective performance. To assess this, discriminate analyses were conducted to determine classification accuracy

of the TD and SLI groups in the exploratory sample on both the PPVT-III and PPVT-IV. See Table 3 for a summary of the results. The discriminate analysis for the PPVT-III yielded a standard score cutoff of 103 for maximally differentiating between children with SLI and TD children (see Figure 4). This cutoff resulted in a sensitivity of .80 and a specificity of .75 on this test. These result in a negative likelihood ratio of .27 and a positive likelihood ratio of 3.20. The results of the discriminate analysis for the PPVT-IV also yielded a standard score cutoff of 103 for maximally differentiating between children with SLI and TD children (see Figure 5). This cutoff resulted in a sensitivity of .80 and a specificity of .70 on this test. These result in a negative likelihood ratio of .29 and a positive likelihood ratio of 2.67.

Based on their individual standard scores on the PPVT-III and PPVT-IV, sixteen out of the 20 children with SLI were correctly classified as SLI, with 4 misclassified, resulting in an error rate of 20%. On the PPVT-III, fifteen out of the 20 TD children were correctly classified, with 5 misclassified, resulting in an error rate of 25%. In contrast, on the PPVT-IV, fourteen of the 20 TD children were correctly classified, with 6 misclassified, resulting in an error rate of 30%. The characteristics of the misclassified children are reported in Tables 4 and 5.

In addition, posterior probabilities were determined for each individual child. The posterior probability of classification refers to the probability that each child was correctly classified into the diagnostic group. Of the 4 children with SLI misclassified on the PPVT-III, the posterior probabilities were as follows: .41, .66, .81, and .96. Therefore, the posterior probability results indicate that these misclassified children had between a 4% and 59% chance of being classified into the wrong group. The five TD

children misclassified on the PPVT-III exhibited the following posterior probabilities: .47, .66, .74, .74, and .89. Therefore, these misclassified children had between an 11% and 53% chance of being wrongly classified into the appropriate group. Posterior probabilities were also classified for individual children on the PPVT-IV. For the SLI group, the posterior probabilities for the 4 misclassified children were as follows: .45, .52, .86, and .99. This indicates that these misclassified TD children had between a 1% and 55% probability of being wrongly classified. For the TD group, the posterior probabilities for the 6 misclassified children were as follows: .57, .57, .63, .69, .86 , and .88. These results indicate that the TD children had between a 12% and 43% chance of being wrongly classified. The results of the posterior probability analyses indicate that none of the children misclassified on the PPVT-III or PPVT-IV were strongly misclassified.

Confirmatory group results. In addition to the exploratory group, a confirmatory group was needed to assess the external validity of the classification accuracy obtained from the exploratory group analyses. Therefore, the cut-off of 103 derived from the exploratory analyses on the PPVT-III and PPVT-IV was applied to the standard scores of the confirmatory group participants in order to calculate sensitivity and specificity and negative and positive likelihood ratios. For the confirmatory group, the mean standard score on the PPVT-III was 109.90 (*SD* = 7.25) and 91.6 (*SD* = 13.94) for the TD group and SLI group respectively. The mean standard score on the PPVT-IV was 109.95 (*SD* = 13.04) for the TD group and 88.80 (*SD* = 15.14) for the SLI group. Similar to the exploratory group, the mean of the TD group on both the PPVT-III and PPVT-IV was higher than the each tests' normative samples and the mean of the SLI group on both of

these tests was within 1 *SD* of the normative samples' mean.

One out of the 5 children in the SLI confirmatory group received a standard score above the 103 cut-off on the PPVT-III and on the PPVT-IV. This single misclassification resulted in a sensitivity of .80 for both test versions, with a positive likelihood ratio of 3.2 for the PPVT-III and 2.67 for the PPVT-IV. In the TD confirmatory group, fifteen out of 20 children received a standard score above 103 on the PPVT-III, with 5 misclassified because they scored below this cut-off. This resulted in a specificity of .75, with a negative likelihood ratio of .27. On the PPVT-IV fourteen out of the 20 TD children obtained a standard score above 103, with 6 misclassified because they scored below the cut-off. This resulted in a specificity of .70 and a negative likelihood ratio of .29. The characteristics of the misclassified children in the exploratory group are reported in Tables 4 and 5 for the SLI and TD groups respectively.

DISCUSSION

In accordance with speech-language pathologists' ethical responsibility to utilize evidence-based practice, norm-referenced assessments should be critically evaluated for the purpose in which they are intended prior to their clinical application. Given previous work documenting insufficient diagnostic accuracy of the PPVT-III for preschool-age children (Gray et al., 1999), it was important to determine whether the most recent edition of this assessment, the PPVT-IV, demonstrated improved diagnostic utility for the purpose of identifying language impairment in this population. A secondary purpose of this investigation was to determine the consistency with which preschool children with and without SLI perform between these two assessments. Issues concerning the utility of these assessments is particularly poignant given their widespread use in both clinical and research settings.

Given the lexical acquisition difficulties characteristic of children with SLI (e.g., Alt & Plante, 2006; Gray, 2003;2004;2005; McGregor, Newman, & Reilly, 2002), it was not surprising that they performed significantly worse than their TD peers on both the PPVT-III and PPVT-IV, which are both designed to assess receptive vocabulary knowledge. Despite the commonly-held notion that children with SLI perform low on tests of child language (see Spaulding, Plante, & Farinella, 2006), the mean performance of the SLI group on these assessments was 93.55 for the PPVT-III and 94.15 for the PPVT-IV, representing -.40 and -.33 standard deviations below the mean respectively. The finding that children with SLI, on average, score relatively well on both of these tests is consistent with the performance specified of the language impaired group represented in each test's examiner manual. The performance of children with language impairments

documented in the manuals along with the independent confirmation observed within this investigation suggests that clinicians should be cautious in assuming that children with language impairment will score low on these measures. Clearly, preschool children with SLI do not.

While the test manuals do indicate that children with language impairment were given these assessments as part of the test development process, they are lacking information describing the tests' ability to differentiate between children with and without language impairment. While speech-language pathologists recruit both formal and informal measures when evaluating children's language skills (Caesar & Kohler, 2009), clear understanding of the diagnostic utility of an assessment measure is critical for understanding how confident a clinician should be in using the results to help determine whether or not a child is language impaired. Previous research by Gray et al. (1999) found that the PPVT-III exhibited only modest diagnostic accuracy for discriminating between preschool children with and without SLI. However, as Gray and colleagues indicate, further analysis of the PPVT is warranted prior to making final determinations. This is especially true because the sensitivity and specificity determinations were sample-dependent and, in contrast to this investigation, were not confirmed by an additional independent sample. The results of this investigation do, however, validate the findings of Gray and colleagues and extend their findings to preschool children as young as 3 years of age. Per Plante and Vance (2004) guidelines of reference, the PPVT-III's diagnostic utility is unacceptable for differentiating between preschool children with SLI and their TD peers.

Given widespread adoption of newer versions of norm-referenced tests by both

clinicians (Caesar & Kohler, 2009) and researchers alike (e.g. Alt, 2011; Hanson et al., 2010; Kulkofsky, 2010), it was important to evaluate whether the PPVT-IV's diagnostic utility was improved relative to its predecessor. The results of this investigation indicate that this is clearly not the case. In fact, while the sensitivity remained consistent between these test versions (.80), the specificity dropped from .75 for the PPVT-III to .70 for the PPVT-IV. Similar to the PPVT-III results obtained in this study, these sensitivity and specificity rates were confirmed with an additional independent sample, the confirmatory sample, providing further support for their external validity. Importantly, the posterior probabilities for the misclassified children are .47 or greater, and the greater the number the less likely an individual has been misclassified. The posterior probability results are concerning, as clinicians may be unlikely to second judge the accurate classifications of children whose language ability is wrongly classified. A descriptive analysis was used in this investigation to pinpoint whether the misclassified children varied systematically from other children in the sample. Their demographic characteristics and test scores did not vary from the accurately classified children in a systematic way. This suggests that there is no clear manner for clinicians to predict who will be correctly and who will be wrongly misclassified.

The decrease in diagnostic accuracy of the PPVT-IV relative to the PPVT-III may be partially attributable to the characteristics of the normative samples within their respective manuals. Both tests included disordered subjects in the normative sample. In the PPVT-III, the normative sample consisted of 11.33% of children with documented disorders while the PPVT-IV's normative sample consisted of 13.4% of children with disorders. Peña, Spaulding, & Plante (2006) conducted a simulation and child language

test manual review study, and found that including disordered subjects in the normative sample resulted in more overlap in performance between children with language impairment and the normative sample used for comparison. This is because including subjects with impairments in the normative sample decreased the mean performance and increased the variability of performance within the normative sample. Accordingly, this resulted in a drop in diagnostic accuracy for tests which included impaired children in the normative sample relative to tests including only typically developing children in the normative sample. Therefore, the higher frequency of disordered subjects in the normative sample of the PPVT-IV relative to the PPVT-III may contribute to the drop in diagnostic accuracy for the more recent PPVT edition observed in this investigation.

Given that the mean performance of the SLI groups on both of these tests was well within one standard deviation of the mean, it was not surprising that the cut-off for maximizing the sensitivity and specificity observed was high. Discriminate analyses of the exploratory group identified an optimal cut-off of 103 for both test versions. This is particularly high relative to cut-offs employed in common clinical practice and research investigations (e.g., Eickhoff, Betz, & Ristow, 2010; Leonard, 1998; Tomblin et al., 1997; Tomblin et al., 1996), but not unexpected. While preschool children with SLI do present with lexical acquisition deficits (e.g., Alt & Plante, 2006; Gray, 2003; 2004; 2005; McGregor, Newman, & Reilly, 2002), their greatest area of weakness tends to be in morphosyntax (Leonard, Eyer, Bedore, & Grela, 1997; Rice & Oetting, 1993; Van der Lely, 2005). Studies documenting the sensitivity and specificity of tests of morphosyntax on preschool children with and without SLI have found much higher levels of discriminate accuracy (e.g., Greenslade, Plante, & Vance, 2009; Merrell & Plante, 1997;

Perona, Plante, & Vance, 2005). In addition to the relative diminished gravity of the word learning deficits of children with SLI relative to their morphosyntax difficulties, the poor diagnostic accuracy of vocabulary assessments in general (see Gray et al., 1999) may be due to how these tests are assessing children's vocabulary knowledge. In the case of the PPVTs, children are asked to point to one of four pictures when provided with a label. This format fails to assess the depth of their knowledge concerning the stimulus presented. Given prior research documenting that children with SLI have difficulty encoding the relevant features when learning new lexical items compared to typically developing peers, particularly in a fast-mapping scenario (Alt & Plante, 2006), the gross assessment of vocabulary knowledge offered by the PPVT tests would likely fail to detect these vocabulary acquisition weaknesses for children with this disorder.

In addition to the diagnostic accuracy, an additional purpose of this investigation was to evaluate consistency in performance between the two test versions. Researchers and clinicians alike would benefit from knowing whether these tests can be used interchangeably for score comparison purposes. Although the results of the discriminate analysis indicate slightly more overlap between how children with SLI and children with TD perform on the PPVT-IV relative to the prior version, the mean scores of the SLI group and the mean scores of the TD group did not differ between the two test versions. This combined with a strong positive correlation between performance on the PPVT-III and PPVT-IV suggests that, on average, children can be expected to perform similarly between these two tests. However, this finding was somewhat misleading. While the average performance did not differ, further inspection at the individual level indicated that some children did perform differently between these two tests. Thirty five percent of

the children in each group exhibited test score differences that exceeded the variability expected from one administration to another. Therefore, their test scores can be considered independent. While some of the children in each group performed better on the PPVT-III, others performed better on the PPVT-IV. This suggests that children who do perform differently between these two tests do not score consistently higher on one version relative to the other.

In sum, the results of this study indicate that neither the PPVT-III nor the PPVT-IV are acceptable for identifying presence and absence of language impairment in preschool children with and without SLI. In addition, while approximately two thirds of children perform consistently between these two tests, nearly one third of children do not. Therefore, these tests are not interchangeable for clinical or research purposes. Future studies may want to determine whether demographic characteristics of the participants, including race, ethnicity, or socioeconomic status, contribute to the version differences observed. Prior research has suggested that typically developing African-American children score, on average, -1.5SD below the mean on the PPVT-III (Kaiser, Milan, & Hancock, 2006). However, an investigation by Washington and Craig (1999) concluded that the PPVT-III was less biased towards at risk African-American children than the earlier PPVT-R edition. Given prior findings of demographic influences on PPVT performance, it continues to be a worthy avenue of exploration as newer editions, such as the PPVT-IV, are published and adopted for use by both researchers and practitioners.

Disordered populations such as children with SLI are particularly vulnerable to the effects of test revision given that both access to and continuation of language services may hinge, in part, on their test performance. Given such high stake decisions, clinicians

place heavy importance on the accuracy of their evaluations. The present investigation adds to the evidence available to date that, although psychometric assessments of child language are frequently revised, they are not necessarily interchangeable with prior versions and do not necessarily result in improved diagnostic utility by virtue of their more recent development.

Furthermore, clinicians should be wary of utilizing existing vocabulary assessments for the diagnosis of children with SLI. As evidenced in this investigation, the PPVT-III and -IV are both lacking in diagnostic utility for this population. However, word learning is still a challenge for children with SLI. In fast mapping and incidental learning tasks, children with SLI typically do not learn as many words as their peers and, if they do, they need more exposures and exhibit slower learning rates relative to controls (Gray, 2004; Oetting et al., 1995; Rice, 1994). Therefore, assessments designed specifically to assess what children with SLI have difficulty with with respect to word learning may help to elucidate the word learning deficits apparent in this population. Specifically, dynamic assessment of the word learning process, a form of testing that measures an individual's potential for learning across several sessions, may prove to be a superior approach to identify language impairment in children with SLI than traditional receptive vocabulary tests currently available.

The generalizability of this investigation is subject to certain limitations. Participants were administered the CELF-P2 in order to assist in determining whether or not they were to be placed in the SLI or TD groups. Given that the CELF-P2 consists of both receptive and expressive subtests, children could perform poorly on this assessment if they had an expressive only or mixed language impairment. Considering the

heterogeneity of this population, it is likely that the SLI group consisted of some children with expressive language impairment and some with both expressive and receptive language impairment. Children with expressive language impairment alone would likely perform well on both versions of the PPVT as they are receptive-based language measures. Future studies may wish to consider evaluating the utility of norm-referenced tests for diagnosing SLI according to the subtype or profile of SLI expressed. For example, Conti-Ramsden, Botting, Simkin, et al. (2001) used cluster analysis to identify five featured subtypes of children with SLI in a sample of 242 school-age children. However, the language profiles of the children evolved with time. Only 55% of the children in their sample retained the same language profile when they were reevaluated a year later. Therefore, as Law, Tomblin, and Zhang (2008) indicate, the difficulty in isolating stable qualities of children in each profile makes it challenging to devise an acceptable paradigm for differentiating language profiles in individuals with SLI. Until well-defined, non-temporally delineated profiles of SLI are established, the results of this investigation provide data which can be generalized to the broader SLI population.

Although generalizability may be improved by having a heterogeneous sample, this study was limited by the small sample size and regional data location sites. It is important to continue to gather additional TD and SLI participants for the purposes of this investigation to generalize the findings to the wider population of preschool children and raise confidence in the study's results. Finally, the participants in this study were all from the state of Connecticut, and consequently the results may not generalize beyond state boundaries. However, given that tests are developed to represent the national population at large, they rarely align well with how children will perform in a particular

region. Therefore, as Merrell and Plante (1998) indicate, it is important to develop local norms, like the ones obtained in this investigation, for comparative purposes.

REFERENCES

Adams, K.M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12*(3), 281-286.

Alt, M. (2011). Phonological working memory impairments in children with specific language impairment: Where does the problem lie? *Journal of Communication Disorders, 43*(2), 173-185.

Alt, M., Plante, E., & Creusere, M. (2004). Semantic features in fast-mapping: Performance of preschoolers with specific language impairment versus preschoolers with normal language. *Journal of Speech, Language, and Hearing Research, 47*(2), 407-420.

Alt, M. & Plante, E. (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49*(5), 941-954.

American National Standards Institute. (1989). Specifications of audiometers. (ANSI S3.6-1989). New York: ANSI.

Ballantyne, A.O., Spilkin, A.M., & Trauner, D.A. (2007). The revision decision: Is change always good? A comparison of CELF-R and CELF-3 test scores in children with language impairment, focal brain damage, and typical development. *Language, Speech, and Hearing Services in Schools, 38*(3), 182-189.

Betz, S.K., Sullivan, S.F., & Eickhoff, J. (2010, June). Factors impacting the selection of standardized tests for the diagnosis of SLI. Poster session presented at the Symposium of Research on Child Language Disorders in Madison, WI.

Bush, S.S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *Clinical Neuropsychologist, 24*(1), 7-16.

Caesar, L.G. & Kohler, P.D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly, 30*(4), 226-236.

Chapman, R.M., Mapstone, M., Porsteinsson, A.P., Gardner, M.N., McCrary, J.W., Degrush, E., Reilly, L.A., & Guillily, M.D. (2010). Diagnosis of Alzheimer's disease using neuropsychological testing improved by multivariate analyses. *Journal of Clinical and Experimental Neuropsychology, 32*(8), 793-808.

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology, 12*(3), 349-358.

Conti-Ramsden, G., Botting, N., Simkin, Z., & Knox, E. (2001). Follow-up of children attending infant language units: Outcomes at 11 years of age. *International Journal of Language and Communication Disorders, 36*(2), 207-219.

Cutuli, J.J., Herbers, J.E., Rinaldi, M., Masten, A.S., & Oberg, C.N. (2010). Asthma and behavior in homeless 4- to 7-year-olds. *Pediatrics, 125*(1), 145-151.

Dawson, J., Eyer, J.A., & Fonkalsrud, J. (2005). *Structured Photographic Expressive Language Test-Preschool: Second Edition.* DeKalb, IL: Janelle Publications.

Dawson, J.I., Stout, C.E., Eyer, J.A. (2003) *Structured Photographic Expressive Language Test: Third Edition.* DeKalb, IL: Janelle Publications.

De Beaman, S.R., Beaman, P.E., Garcia-Peña, C., Villa, M.A., Heres, J., Cordova, A., & Jagger, C. (2004). Validation of a modified version of the Mini-Mental State Examination (MMSE) in Spanish. *Aging, Neuropsychology, and Cognition, 11*(1), 1-11.

Diamond, G.A., & Forrester, J.S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine, 98*(3), 385-394.

Dollaghan, C.A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37*(5), 391-400.

Dunn, L. M,, & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised.* Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–III.* Circle Pines, MN: American Guidance Service.

Dunn, L.M., & Dunn, D.M. (2007). *Peabody Picture Vocabulary Test-IV.* Circle Pines, MN: American Guidance Service.

Eickhoff, J., Betz, S.K., & Ristow, J. (2010, June). Clinical procedures used by speech language pathologists to diagnose SLI. Poster session presented at the Symposium of Research on Child Language Disorders in Madison, WI.

Emmons, M.R., & Alfonso, V.C. (2005). A critical review of the technical characteristics of current preschool screening batteries. *Journal of Psychoeducational Assessment, 23*(2), 111-127.

Farrar, M.J., Johnson, B., Tompkins, V., Easter, M., Zilisi-Medus, A., & Benigno, J.P. (2009). Language and theory of mind in preschool children with specific language

impairment. *Journal of Communication Disorders, 42*(6), 428-441.

Gray, S. (2003). Word learning by preschoolers with specific language impairment: What predicts success? *Journal of Speech, Language, and Hearing Research, 46*(1), 56-67.

Gray, S. (2004). Word learning by preschoolers with specific language impairment: Predictors and poor learners. *Journal of Speech, Language, and Hearing Research, 47*(5), 1117-1132.

Gray, S. (2005). Word learning by preschoolers with specific language impairment: Effect of phonological or semantic cues. *Journal of Speech, Language, and Hearing Research, 48*(6), 1452-1467.

Gray, S. (2006). The relationship between phonological memory, receptive vocabulary, and fast mapping in young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49*(5), 955-969.

Gray, S. Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests adminstered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), *1*96-206.

Greenslade, K.J., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construct validity of the Structured Photographic Expressive Language Test-Preschool: Second Edition. *Language, Speech, and Hearing Services in Schools, 40*(2), 150-160.

Grela, B.G., & Leonard, L.B. (1997). The use of subject arguments by children with specific language impairment. *Clinical Linguistics and Phonetics, 11*(6), 443-453.

Hanson, E., Nasir, R.H., Fong, A., Lian, A., Hundley, R., Shen, Y., Wu, B.L., Holm, I.A.,

& Miller, D.T. (2010). Cognitive and behavioral characteristics of 16p11.2 deletion syndrome. *Journal of Developmental and Behavioral Pediatrics, 31*(8), 649-657.

Hart, S.A., Petrill, S.A., & Kamp Dush, C.M. (2010). Genetic influences on language, reading, and mathematics skills in a national sample: An analysis using the national longitudinal survey of youth. *Language, Speech, and Hearing Services in Schools, 41*(1), 118-128.

Jessup, B. Ward, E., Cahill, L., & Keating, D. (2008). Teacher identification of speech and language impairment in kindergarten students using the Kindergarten Development Check. *International Journal of Speech-Language Pathology, 10*(6), 449-459.

Johnson, C.J. (1995). Expanding norms for narration. *Language, Speech, and Hearing Services in Schools, 26*(4), 326-341.

Kaufman, A.S., & Kaufman, L.N. (2004). *Kaufman Assessment Battery for Children Second Edition, Manual.* AGS Publishing, Circle Pines.

Kulkofsky, S. (2010). The effects of verbal labels and vocabulary skill on memory and suggestibility. *Journal of Applied Developmental Psychology, 31*(6), 460-466.

Lam, J.C., Mahone, E.M., Maston, T., & Scharf, S.M. (2011). The effects of napping on cognitive function in preschoolers. *Journal of Developmental and Behavioral Pediatrics, 32*(2), 90-97.

Law, J., Tomblin, J.B., & Zhang, X. (2008). Characterizing the growth trajectories of language-impaired children between 7 and 11 years of age. *Journal of Speech, Language, and Hearing Research, 51*(3), 739-749.

Leonard, L.B., Eyer, J.A., Bedore, L.M., & Grela, B.G. (1997). Three accounts of the grammatical morpheme difficulties of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 40*(4), 741-753.

McFadden, T.U. (1996). Creating language impairments in typically achieving children: The pitfalls of “normal” normative sampling. *Language, Speech, and Hearing Services in Schools, 27*(1), 3-9.

McGregor, K.K., Newman, R.M., Reilly, R.M., & Capone, N.C. (2002). Semantic representation and naming in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 45*(5), 998-1014.

Meador, K.J., Baker, G.A., Browning, N., Cohen, M.J., Clayton-Smith, J., Kalayjian, L.A., Kanner, A., Liporace, J.D., Pennell, P.B., Privitera, M., & Loring, D.W. (2011). Fetal antiepileptic drug exposure and verbal versus non-verbal abilities at three years of age. *Brain, 134*(2), 396-404.

Merrell, A.W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*(1), 50-58.

Nash, M., & Donaldson, M.L. (2005). Word learning in children with vocabulary deficits. *Journal of Speech, Language, and Hearing Research, 48*(2), 439-458.

O'Neill, D.K. (2007). The language use inventory for young children: A parent report measure of pragmatic language development for 18- to 47-month old children. *Journal of Speech, Language, and Hearing Research, 50*(1), 214-228.

Oetting, J.B., Rice, M.L., & Swank, L.K. (1995). Quick incidental learning (QUIL) of words by school-age children with and without SLI. *Journal of Speech and*

Hearing Research, 38(2), 434-445.

Pankratz, M., Morrison, A., & Plante, E. (2004). Difference in standard scores of adults on the Peabody Picture Vocabulary Test (Revised and Third Edition). *Journal of Speech, Language, andHearing Research, 47*(3), 714-718.

Pankratz, M.E., Vance, E.P.R., & Insalaco, D.M. (2007). The diagnostic and predictive validity of the Renfrew Bus Story. *Language, Speech, and Hearing Services in Schools, 38*(4), 390-399.

Peña, E.D. Spaulding, T.J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15*(3), 247-254.

Perona, K., Plante, E., & Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition (SPELT-3). *Language, Speech, and Hearing Services in Schools, 36*(2), 103-115.

Paul, R. (1995). Language disorders from infancy through adolescence: Assessment and intervention. Austin, TX: Pro-Ed.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24.

Preston, J., & Edwards, M.L. (2010). Phonological awareness and types of sound errors in preschoolers with speech sound disorders. *Journal of Speech, Language, and Hearing Research, 53*(1), 44-60.

Restrepo, M.A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research,* 41(6), 1398-1411.

Rice, M.L., & Oetting, J.B. (1993). Morphological deficits of children with SLI: Evaluation of number marking and agreement. *Journal of Speech and Hearing Research, 36*(6), 1249-1257.

Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88*(2), 229-247.

Rescorla, L., Roberts, J., Dahlsgaard, K. (1997). Late talkers at 2: Outcome at age 3. *Journal of Speech, Language, and Hearing Research, 40*(3), 556-566.

Rice, M.L., Oetting, J.B., Marquis, J., Bode, J., & Pae, S. (1994). Frequency of input effects on word comprehension of children with specific language impairment. *Journal of Speech and Hearing Research, 37*(1), 106-122.

Rice, M.L., Wexler, K., & Cleave, P.L. (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research, 38*(4), 850-863.

Rvachew, S., Chiang, P.Y., & Evans, N. (2007). Characteristics of speech errors produced by children with and without delayed phonological awareness skills. *Language, Speech, and Hearing Services in Schools, 38*(1), 60-71.

Scarborough, H.S. (1990). Very early language deficits in dyslexic children. *Child Development, 61*(6), 1728-1743.

Seiger-Gardener, L., & Brooks, P.J. (2008). Effects of onset- and rhyme-related distractors on phonological processing in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 51*(5), 1263-1281.

Semel, E., Wiig, E.H., & Secord, W. (1987). Clinical Evaluation of Language Fundamentals Revised. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E.H., & Secord, W.A. (1995). *Clinical Evaluation of Language Fundamentals Third Edition.* San Antonio, TX: The Psychological Corporation.

Shipley, K.G., Stone, T.A., & Sue, M.B. (1983). *Test for Examining Expressive Morphology (TEEM).* Tucson, AZ: Communication Skill Builders.

Silliman, E.R., Diehl, S.F., Bahr, R.H., Hnath-Chisolm, T., Zenko, C.B., Friedman, S.A. (2003). A new look at performance on theory-of-mind tasks by adolescents with autism spectrum disorder. *Language, Speech, and Hearing Services in Schools, 34*(3), 236-252.

Spaulding, T.J., Plante, E., Farinella, K.A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37*(1), 61-72.

Stockman, I.J. (2000). The new Peabody Picture Vocabulary Test-III: An illusion of unbiased assessment? *Language, Speech, and Hearing Services in Schools, 31*(4), 340-353.

Storkel, H.L., & Rogers, M.A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistic and Phonetics, 14*(6), 407-425.

Sutherland, D. & Gillon, G.T. (2005). Assessment of phonological representations in children with specific language impairment. *Language, Speech, and Hearing Services in Schools, 36*(4), 294-307.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children.

Journal of Speech, Language, and Hearing Research, 40(6), 1245-1260.

Ukrainetz, T.A., & Duncan, D.S. (2000). From old to new: Examining score increases on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools, 31*(4), 336-339.

van der Lely, H.K.J. (2005). Domain-specific cognitive systems: Insight from grammatical SLI. *Trends in Cognitive Science, 9*(2), 53-59.

van der Lely, H.K.J. & Marshall, C.R. (2010). Assessing component language deficits in the early detection of reading difficulty risk. *Journal of Learning Disabilities, 43*(4), 357-368.

Waggoner, T.L. (2002). *Color Vision Testing Made Easy.* Elgin, IL: Good-lite Company.

Watkins, R.V., Kelly, D.J., Habers, H.M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*(6), 1349-1355.

Werner, E., & Kresheck, J.D. (1983). Structured Photographic Expressive Language Test-II. Sandwich, IL: Janelle Publications.

Wiig, E.H., Secord, W.A., & Semel, E. (2004). *Clinical evaluation of language fundamentals- Preschool, Second Edition.* Toronto: The Psychological Corporation.

Wilson, K.S., Blackmon, R.C. Hall, R.E. & Elcholtz, G.E. (1991). Methods of language assessment: A survey of California Public School clinicians. *Language, Speech, and Hearing Services in Schools, 22*(4), 236-241.

Wise, J.C., Sevcik, R.A., Morris, R.D., Lovett, M.W., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension,

pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language, and Hearing Research, 50*(4), 1093-1109.

Williams, K.T. (1998). Peabody Picture Vocabulary Test-III: What is new and different? *Clinical Connection, 11*(3), 6-8.

Young, E.C., & Perachio, J.J. (1993). The Patterned Elicitation Syntax Test with morphophonemic analysis. Tucson, AZ: Communication Skill Builders.

Figure 1. Mean Performance of TD and SLI groups

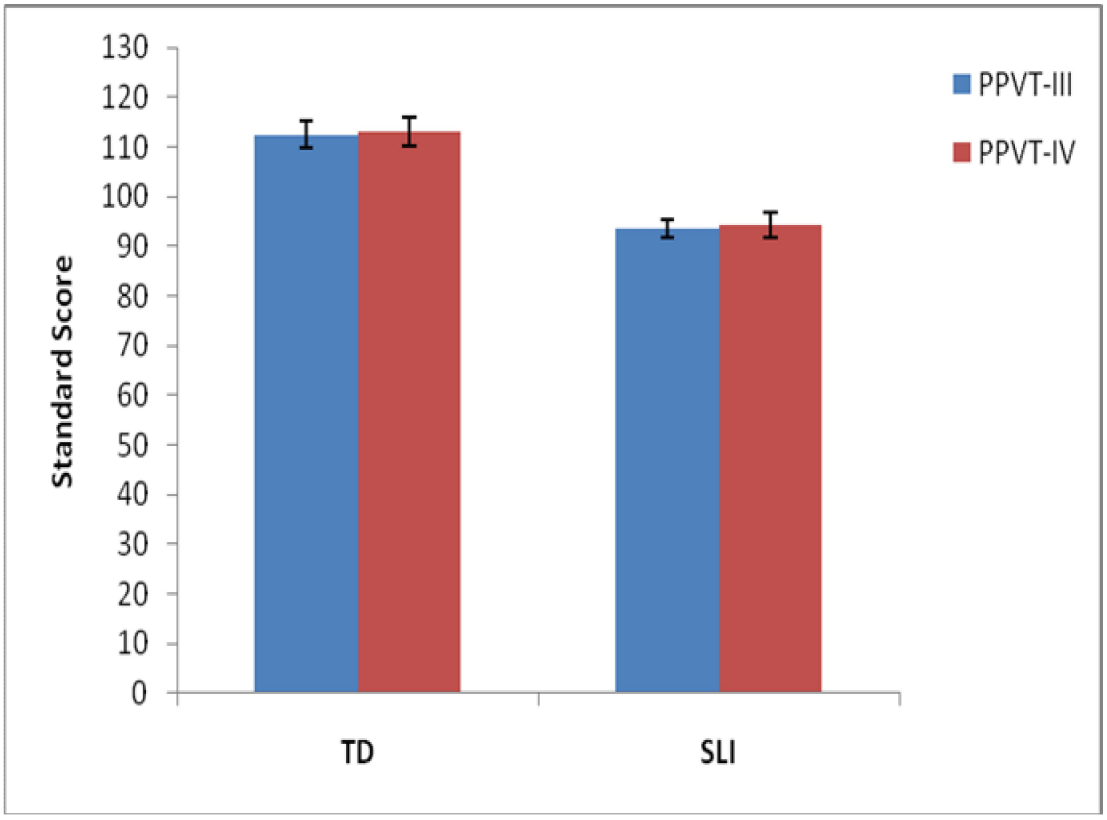

Figure 2. Individual Variability: Typically Developing Group

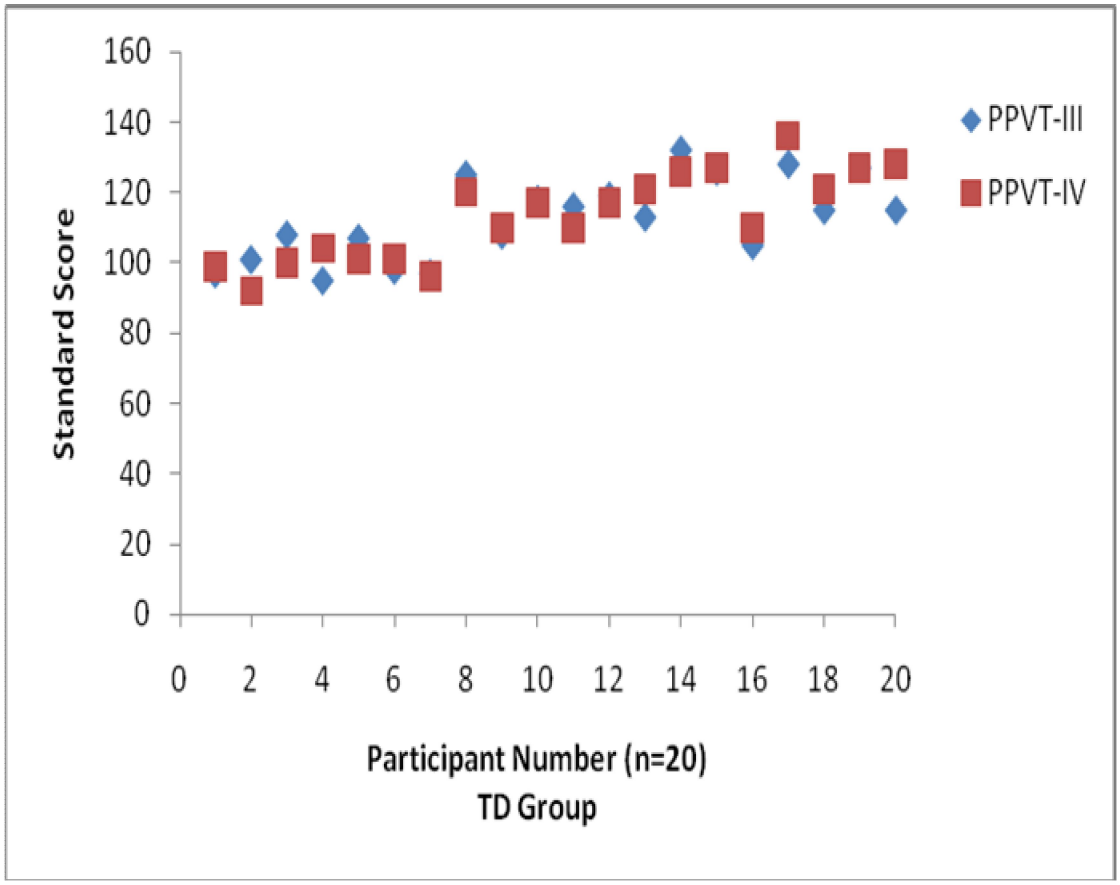

Figure 3. Individual Variability: SLI Group

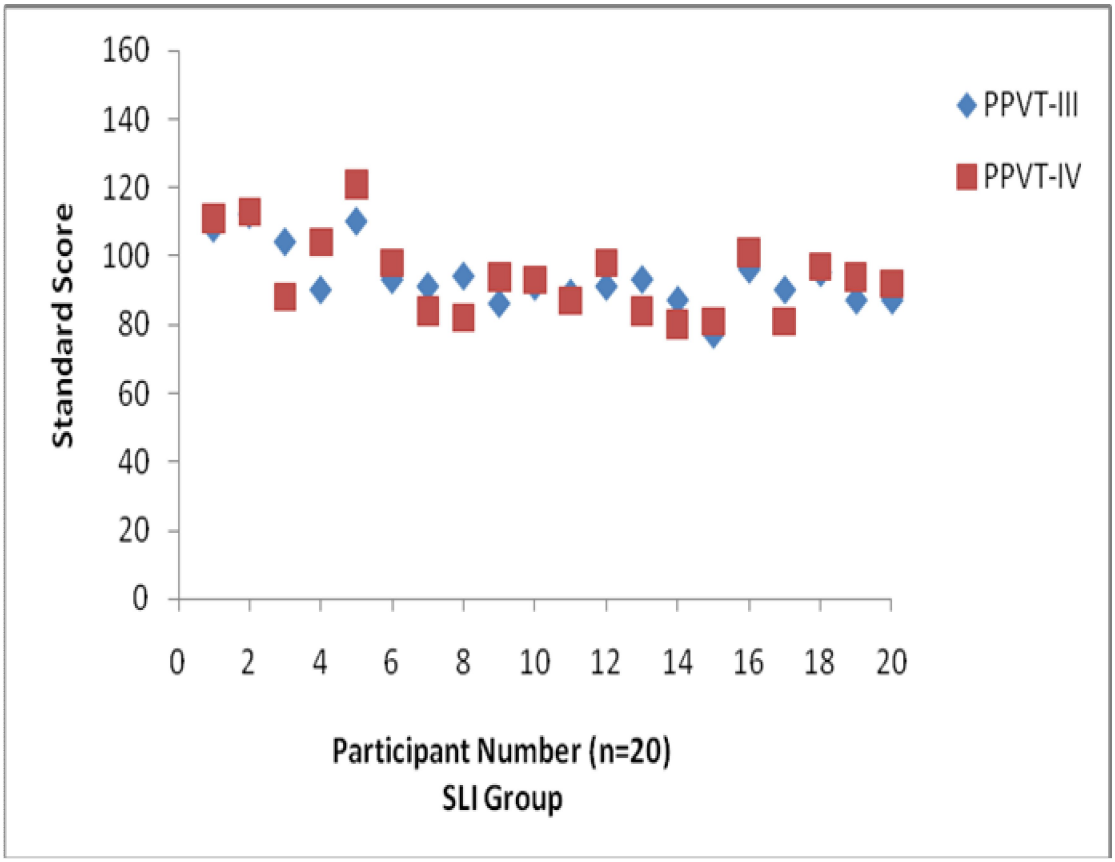

Figure 4. *Distribution of PPVT-III standard scores obtained by exploratory group participants. The distribution demonstrates the cutoff score of 103.*

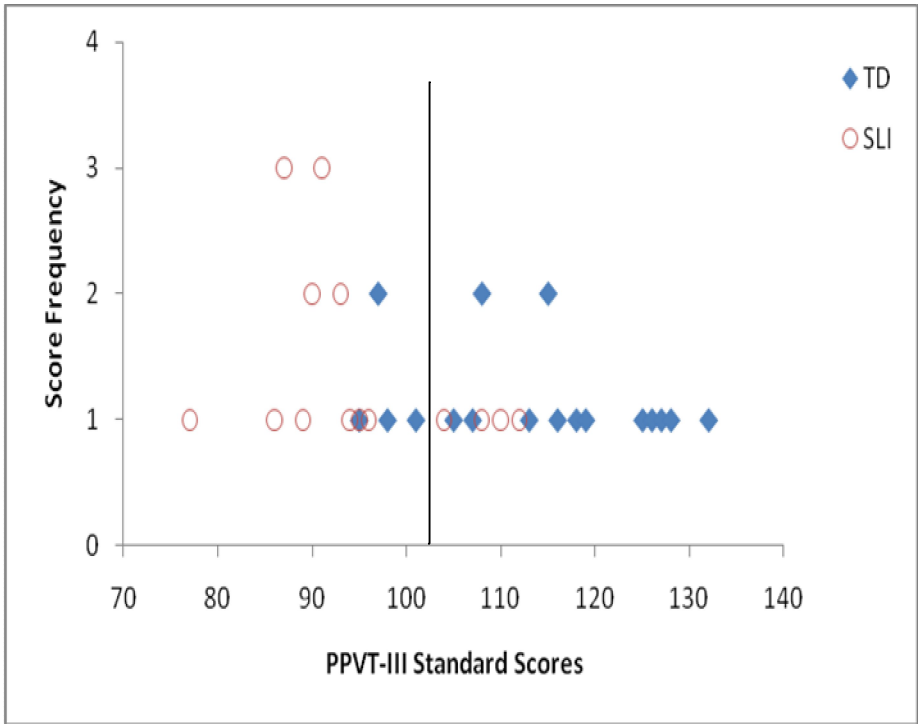

Figure 5. *Distribution of PPVT-IV standard scores obtained by exploratory group participants. The distribution demonstrates the cutoff score of 103.*

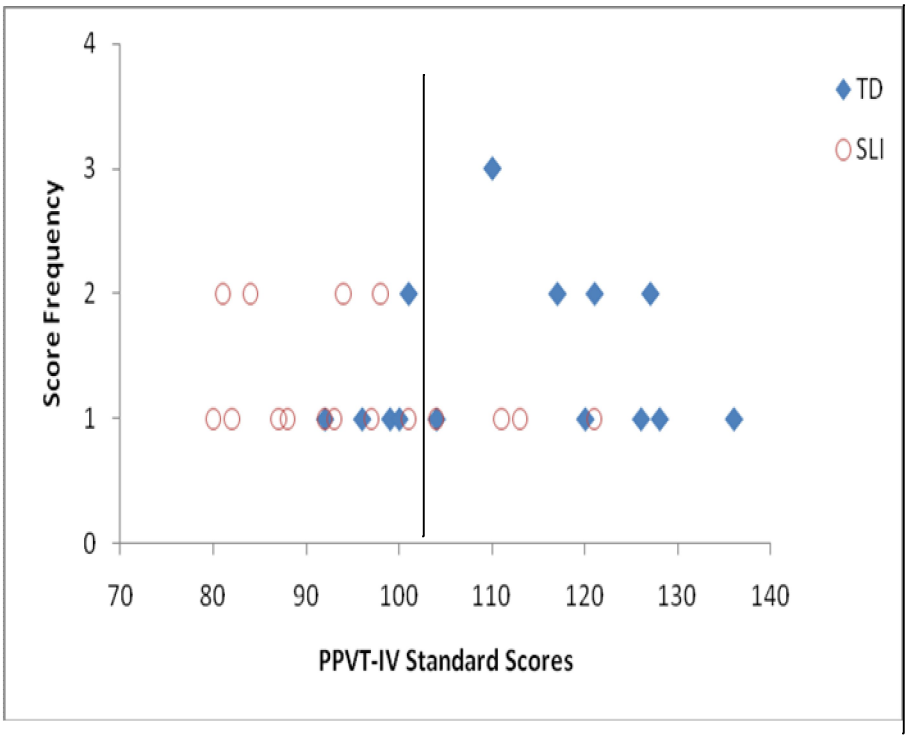

Table 1

Demographic Characteristics of Participants

	Exploratory Sample		Confirmatory Sample	
	TD	SLI	TD	SLI
Gender	10M, 10F	10M, 10F	12M, 8F	4M, 1F
Age				
Mean	52.65	51.85	52.05	51.20
Range	(45-64)	(43-63)	(42-59)	(46-55)
Race				
Afr-Am	2	4	7	1
Asian	0	0	1	0
Caucasian	13	11	11	3
Mixed	5	2	0	0
Not Reported	0	3	1	1
Ethnicity				
Hispanic	8	11	3	2
Not Hispanic	11	7	14	3
Not Reported	1	2	3	0
Maternal Education Level				
Mean	14.42	14.26	14.53	14.20
Range	(11-18)	(9-18)	(9-18)	(11-18)

Afr-Am = African American; Mixed = multi-racial

Table 2

Exploratory and Confirmatory Group Performance on Norm-Referenced Assessments

	TD Group			SLI Group		
	Mean	SD	Range	Mean	SD	Range
Exploratory Participants						
CELF-P2*	108.90	11.96	90-131	78.55	6.75	63-84
KABC-II	110.95	10.04	94-125	106.15	7.68	92-119
Confirmatory Participants						
CELF-P2*	103.40	6.16	94-114	79.00	4.30	73-84
KABC-II	110.20	10.38	91-128	103.60	9.07	95-119

Note: CELF-P2 = Clinical Evaluation of Language Fundamentals – Preschool, Second Edition (Wiig, Secord, & Semel, 2004); KABC-II = Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004)

* = significant difference at $p = .05$

Table 3

PPVT-III and PPVT-IV Sensitivity and Specificity Data for Exploratory and Confirmatory Samples

	Group categorization based on CELF-P2 scores and clinical judgment			
Group categorization based on discriminate analysis	PPVT-III		PPVT-IV	
	TD	SLI	TD	SLI
Exploratory Sample				
TD (n=20)	15(.75)	4(.20)	14(.70)	4(.20)
SLI (n=20)	5(.25)	16(.80)	6(.30)	16(.80)
Confirmatory Sample				
TD (n = 20)	15(.75)	1(.20)	14(.70)	1(.20)
SLI (n = 5)	5(.25)	4(.80)	6(.30)	4(.80)

Note: CELF-P2 = Clinical Evaluation of Language Fundamentals – Preschool, Second Edition (Wiig, Secord, & Semel, 2004); PPVT-III = Peabody Picture Vocabulary Test – Third Edition (Dunn & Dunn, 1997); PPVT-IV = Peabody Picture Vocabulary Test-Fourth Edition (Dunn & Dunn, 2007)

Table 4. Characteristics of children with TD misclassified as SLI on the PPVT-III and PPVT-IV.

	Test Performance				Demographic Characteristics			
Child	PPVT-III	PPVT-IV	CELF-P2	KABC-II	Age	Gender	Race/Ethnicity	SES
Exploratory Sample								
1	95	104	96	94	62	F	Mixed/NR	14
2	97	96	90	98	57	F	Mixed/Hispanic	12
3	97	99	90	116	49	F	Mixed/Hispanic	14
4	98	101	94	96	52	F	White/Hispanic	14
5	101	92	100	100	58	F	Mixed/Hispanic	12
6	107	101	104	104	47	F	AfrAm/NH	14
7	108	100	108	111	48	F	White/NH	14
Confirmatory Sample								
1	100	106	102	100	51	M	AfrAm/NH	18
2	100	114	108	100	57	M	AfrAm/NH	16
3	102	111	94	115	48	F	AfrAm/NR	16
4	102	112	112	111	54	M	White/NH	18
5	103	91	96	95	56	M	AfrAm/NH	11
6	104	90	98	91	53	F	White/NH	16
7	105	89	98	128	59	F	NR/ Hispanic	9
8	107	101	98	111	49	M	AfrAm/Hispanic	15
9	110	98	106	126	59	F	Asian/NR	NR
10	113	91	102	106	42	M	White/NR	14

Note: Test performance reported in standard scores (Mean = 100, SD= 15).
AfrAm = African American, NH= not Hispanic, NR = not reported.
PPVT-III = Peabody Picture Vocabulary Test –Third Edition (Dunn & Dunn, 1997); PPVT-IV = Peabody Picture Vocabulary Test-Fourth Edition (Dunn & Dunn, 2007)

Table 5. Characteristics of children with SLI misclassified as TD on the PPVT-III and PPVT-IV.

	Test Performance				Demographic Characteristics			
Child	PPVT-III	PPVT-IV	CELF-P2	KABC-II	Age	Gender	Race/Ethnicity	SES
Exploratory Sample								
1	104	88	83	95	62	F	Mixed/Hispanic	13
2	108	111	79	113	44	F	White/Hispanic	16
3	112	113	84	111	46	M	White/Hispanic	18
4	90	104	83	109	56	M	White/Hispanic	14
5	110	121	84	119	46	F	Mixed/Hispanic	18
Confirmatory Sample								
1	112	113	82	119	54	F	White/NH	16

Note: Test performance reported in standard scores (Mean = 100, SD= 15).
PPVT-III = Peabody Picture Vocabulary Test –Third Edition (Dunn & Dunn, 1997); PPVT-IV = Peabody Picture Vocabulary Test-Fourth Edition (Dunn & Dunn, 2007)