Journal of Communication Disorders 45 (2012) 59–68

Contents lists available at SciVerse ScienceDirect

Journal of Communication Disorders

Comparison of severity ratings on norm-referenced tests for children with specific language impairment

Tammie J. Spaulding *

University of Connecticut, Department of Communication Sciences, 850 Bolton Road, Storrs, CT 06269, United States

ARTICLE INFO

Article history:
Received 12 August 2011
Received in revised form 2 November 2011
Accepted 12 November 2011
Available online 25 November 2011

Keywords:
Specific language impairment (SLI)
Assessment
Severity

ABSTRACT

Purpose: This study evaluated the consistency in severity classifications for children with language impairment on tests of child language.
Methods: The TELD-3 and the UTLD-4 were administered to 16 preschool children with specific language impairment (SLI) and 16 typical controls. The boundaries described in the test manuals were used to assign language proficiency ratings to these children and to subsequently evaluate the consistency in these designations.
Results: Performance categories were more consistent for the typical children than for the children with SLI. When evaluating how children perform on the two tests, the severity category remained consistent for only 19% of the children with SLI when using the severity category boundaries recommended within the test manuals.
Conclusions: Clinicians should be cautious in assigning severity of impairment classifications to children with language impairment based, in part or in whole, on their performance on norm-referenced tests.

Learning outcomes: Readers will see the importance of relying on empirical evidence to support their clinical decisions, specifically in the area of severity of impairment determinations. Readers will learn of the lack of stability in severity of language impairment classifications for children with language impairment on tests of child language. Consequently, readers will learn to be cautious in the selection of norm-referenced tests of child language for the purposes of informing severity of impairment determinations.

1. Introduction

A diagnosis of language impairment in a child is typically followed by another clinical judgment, the determination of severity. Practical clinical guidelines for determining severity of impairment have not been established. In the absence of a gold standard, it is important to evaluate the usefulness of currently available tools for this purpose. A prior study by Spaulding, Swartwout Szulga, and Figueroa et al. (in press) found that a number of U.S. State Departments of Education indicate that clinicians are to use scores on norm-referenced assessments to determine or to assist in determining children's degree of language impairment. The authors also found that eleven norm-referenced test manuals of child language provide severity categories for clinicians to use based on children's performance on these tests. This investigation was designed to determine if there is empirical support for identifying severity of language impairment using norm-referenced test scores. Specifically, the consistency of severity designations assigned to

* Corresponding author. Tel.: +1 8604861665; fax: +1 8604865422.
E-mail address: tammie.spaulding@uconn.edu.

0021-9924/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcomdis.2011.11.003

preschool children with specific language impairment (SLI) was determined using the procedures identified within the corresponding norm-referenced test manuals of child language.

A child's language proficiency is the characterization of his or her language skills relative to some benchmark or expectation. When assessing the language proficiency of children with language impairment, this comparison is typically made relative to their same-aged peers. A given child's language proficiency can be similar to that expected for their age, above expectations for their age, or in the case of language impairment, below developmental expectations. Descriptive terms are often used to characterize children's degree of language proficiency. For example, a child with acceptable language skills for their age may be characterized as exhibiting "average" language skills or their language skills may be classified as "above average", "advanced", or "very advanced" for their age. In the case of language impairment, proficiency labels represent severity of impairment. For example, a language impaired child's language skills may be characterized as "delayed", "very delayed", "below average", "low", "very low", or their language impairment may be classified as "mild", "moderate", "severe", or "profound".

One way to measure a child's language proficiency is to compare a child's score on a norm-referenced test to proficiency criteria available within the examiner's manual. Prior work has demonstrated that 11 of the most recent editions of tests of child language provide tables within their manuals to convert a child's given score to a proficiency category (Spaulding et al., in press). Each one of the tables in the test manuals provides a range of scores associated with each proficiency category. A child's language proficiency category is determined based on which range of scores encompasses their score on the norm-referenced test.

For speech language pathologists who typically administer such tests when there is a suspected or confirmed language impairment, these proficiency characterizations are assumed to be useful for describing the language skills of children with language impairment relative to their peers, as children with language impairment are not developing language in an appropriate manner or at an adequate pace relative to children who are becoming proficient, competent language users. The goal of the severity labels provided by a test score to proficiency label conversion is to determine how disparate, or off track a given child's language skills are relative to his or her peers. This information is used to inform clinical judgment, as it goes one step further than identifying if an impairment exists by providing the clinician with information as to how impaired a child's linguistic skills are relative to age expectations.

An important issue that arises when using children's scores on norm-referenced assessments to determine or to assist in determining severity of impairment is that a child's performance may differ depending on which assessments are selected for this purpose. A review of 43 tests of child language indicates that there is substantial variability in the mean performance of children with language impairment, in particular, across tests of child language (Spaulding, Plante, & Farinella, 2006). This may be expected given that some tests assess language more broadly, while others assess a single language domain. However, a number of studies have found that, even when assessing the identical language domain, the same sample of children with language impairment perform differently depending on which norm-referenced test is administered (e.g., Ballantyne, Spilkin, & Trauner, 2007; Gray, Plante, Vance, & Henrichsen, 1999; Merrell & Plante, 1997; Rice, Ash, Abel, & Lee, 2008). If differences are sufficiently large the severity designation, if based on test scores, may also differ depending upon which test is selected for interpretation.

One investigation to date has documented differences in severity ratings assigned to children with language impairment. Ballantyne et al. (2007) compared the performance of children with language impairment on the *Clinical Evaluation of Language Fundamentals-Revised* (CELF-R; Semel, Wiig, & Secord, 1987) and the *Clinical Evaluation of Language Fundamentals-Third Edition* (CELF-3; Semel, Wiig, & Secord, 1995). Their investigation consisted of a cross-sectional study and a longitudinal study of school-age children with language impairment. The results of both found that children with language impairment (and typical language peers) performed better on the newer version (the CELF-3) than the previous version (the CELF-R). Applying the cut-offs for severity categories provided within the test manuals resulted in a decrease in documented severity for school-age children with language impairment on the CELF-3 relative to the CELF-R. Thirteen children, which constituted the longitudinal sample of children with SLI, were given both the CELF-R and the CELF-3. However, substantial time had elapsed between administration of these two tests. For example, the children ranged from 5 years, 0 months to 11 years, 9 months when they were given the CELF-R, but were between 8 years, 4 months and 15 years, 8 months of age when administered the CELF-3. The average time between administration of the CELF-R and the CELF-3 was four years. Spaulding et al. (2006) documented substantial variability in how children with language impairment, relative to typical language peers, performed on the same test across different ages. In addition, a study by Conti-Ramsden, Botting, Simkin, and Knox (2001) found that language profiles varied dramatically for many children with SLI when comparing skills from 7 to 11 years of age, consistent with a four year duration. Given such variation, whether or not severity ratings would change or remain consistent if children with SLI were administered both tests at the same age is still an open question.

The findings of inconsistent severity classifications by Ballantyne et al. (2007), however, do suggest that caution may be needed when assigning severity categories to children with language impairment based, in part or in whole, on their performance on norm-referenced tests. This caution, however, does not appear to be heavily considered by a number of U.S. State Departments of Education. Spaulding et al. (in press) conducted a recent review of U.S. State Departments' of Education guidelines for ascertaining severity of impairment, and found that eight states provided specific criteria for using norm-referenced test scores to inform severity of language impairment determinations. Unfortunately, seven of the eight states provided criteria for converting a child's test score to a severity rating using specific boundaries without indicating which norm-referenced tests are appropriate for the boundaries specified. Without this information, an assumption would be to

apply such boundaries on whichever tests a clinician selects to administer. At the present time, we are lacking empirical data to support applying the same boundaries across tests of child language for determining severity of impairment. Likewise there is insufficient data to date to argue the contrary with certainty, specifically that it is inappropriate to use the same boundaries across tests of child language to determine or to assist in determining severity of impairment; hence the motivation for the current work.

In sum, because eleven tests of child language provide specific criteria in their manuals for converting a child's score to a severity rating (see Spaulding et al., in press), at a minimum these test developers are indicating that they should be used for this purpose. However, at the present time there is currently no empirical support for this practice. The research to date provides evidence that children with SLI score differently on different tests (e.g., Gray et al., 1999; Merrell & Plante, 1997; Rice et al., 2008) although whether the difference in performance is large enough to constitute a variation in severity label is unclear. Prior work has provided preliminary evidence that severity ratings may change for school-age children with language impairment when administered different versions of the same test (Ballantyne et al., 2007). At the current time, it is unknown if severity ratings assigned by virtue of norm-referenced test performance would differ for younger, preschool-age children. In addition, there is no empirical data evaluating if severity ratings for children with SLI change if administered two different norm-referenced tests at the same age, which would provide stronger evidence of such variation if it exists. Clearly a number of U.S. State Departments of Education are suggesting that the same boundaries can be applied across tests of child language to determine severity of impairment. Therefore, this area deserves further investigation.

1.1. The current study

The objective of the current study is to evaluate the utility of assigning severity designations to preschool children with SLI based on norm-referenced test performance by comparing the consistency in severity classifications of children with SLI between two tests of child language. The tests selected for use were the *Test of Early Language Development-Third Edition* (TELD-3; Hresko, Reid, & Hammill, 1999) and the Utah *Test of Language Development-Fourth Edition* (UTLD-4; Mecham, 2003). These tests were selected because they were omnibus tests of language, included similar boundaries (or cut-scores) in their respective manuals for determining degree of impairment, and were appropriate for use for preschool children between 3 and 5 years of age. To avoid the potential confound of age of administration on the results, the children were administered both tests at the same age. Therefore, this work will assess whether severity of impairment designations remain the same for children when the same boundaries are applied across tests of child language to determine severity of impairment, as a number of U.S. State Departments of Education indirectly suggest, even when the two tests themselves provide similar boundaries for clinicians to use to determine children's language proficiency.

2. Methods

2.1. Participants

Thirty-two preschool children between the ages of 3 years, 2 months and 5 years, 5 months participated in this investigation. Half of these children exhibited typically developing (TD) language skills while the remaining children exhibited specific language impairment (SLI). All the participants were native, monolingual speakers of American English and were from a range of racial, ethnic, and socioeconomic backgrounds. Participants were individually matched for age (±3 months) and gender, and were group matched for socioeconomic status, as determined by maternal education level. See Table 1 for the demographic characteristics of the participants.

All participants in this investigation exhibited normal range hearing sensitivity as measured by an audiometric pure tone screening at 25 dB HL at 500 Hz, and 20 dB HL at 1, 2, and 4 kHz (ANSI, 2004). These levels were adjusted at times to accommodate ambient noise in the testing environment. The participants also exhibited normal range nonverbal cognitive skills measured by a standard score of 75 (70 + SEM) or above[1] on nonverbal index of the *Kaufman Assessment Battery for Children, Second Edition* (KABC-II; Kaufman & Kaufman, 2004). Participants exhibited no motor, behavior, frank neurological deficits, or other handicapping conditions as determined by parent and teacher questionnaires.

Children were designated as typically developing or language impaired based, in part, on their scores on the *Clinical Evaluation of Language Fundamentals-Preschool; Second Edition* (CELF-P2; Wiig, Secord, & Semel, 2004). This test includes diagnostic accuracy information in the examiner's manual. According to the manual, a standard score cut-off of 85 (or 1 SD below the mean), results in a sensitivity of 85% and a specificity of 82%. For this investigation, all typically developing children scored above a standard score of 85 while the children with specific language impairment scored at or below the 85 cut-off score. Confirmation of group classification was supported in each case by clinical judgment. Clinical judgment of typical or impaired language skills was made by a certified speech-language pathologist who was blind to the CELF-P2 results and was based on factors including (a) enrollment in services; (b) impressions of language impairment or no impairment during conversational speech; and (c) parent and teacher reports of impairment or no impairment on a brief questionnaire. Participants were also administered the *Peabody Picture Vocabulary Test-Fourth Edition* (PPVT-4; Dunn &

[1] Although a cut-off of 75 was used for the purpose of this investigation, the lowest score a child obtained was a standard score of 89.

Table 1
Demographic characteristics of participants.

	SLI ($n = 16$)	TD ($n = 16$)
Age	$M = 50.81$ months	$M = 51.38$ months
SD	8.89	8.79
Range	(38–64 months)	(38–65 months)
Sex	9 boys, 7 girls	9 boys, 7 girls
Mother's education level	$M = 14.28$ years	$M = 14.61$ years
SD	1.32	1.50
Range	(12–16 years)	(13–17 years)
Ethnicity (n)		
Not Hispanic	7	10
Hispanic	5	4
Not reported	4	2
Race (n)		
White	9	13
Black/African American	1	1
Multiracial	2	1
Not reported	4	1

Table 2
Norm-referenced test performance.

	SLI group			TD group		
	Mean	SD	Range	Mean	SD	Range
*CELF-P2	77.69	7.19	65–84	113.06	9.55	98–129
*PPVT-IV	91.06	9.15	77–111	113.75	10.70	92–126
KABC-II	102.88	8.64	89–115	106.09	7.51	98–119

Note: CELF-P2, Clinical Evaluation of Language Fundamentals – Preschool, second edition; PPVT-IV, Peabody Picture Vocabulary Test – fourth edition; KABC-II, Kaufman Assessment Battery for Children – second edition.
* Significantly different at $p < .05$.

Dunn, 2007) to further describe language skills. Given the lack of empirical evidence available to support this norm-referenced test in diagnosing children with SLI, it was not used for diagnostic purposes. See Table 2 for norm-referenced test performance of the SLI and TD groups.

2.2. Materials

2.2.1. Test of Early Language Development-Third Edition (TELD-3)

The *Test of Early Language Development-Third Edition* (TELD-3; Hresko et al., 1999) was normed on 2217 children who were representative of the U.S. population based on the U.S. Bureau of the Census (1997) *Statistical Abstracts of the United States*. It is designed to measure the spoken language of children aged 2 years, 0 months to 7 years, 11 months. The TELD-3 consists of a Receptive Language subtest and an Expressive Language subtest, both of which include items that assess semantic, morphology, and syntax skills. Performance on these subtests is converted to a Receptive Language quotient and Expressive Language quotient respectively, which combine to form a Spoken Language Composite score. Each quotient as well as the composite score is based on a mean of 100 and a standard deviation of 15. This test contains two forms, Form A and Form B. Form A was used exclusively for the purpose of this investigation.

The examiner's manual indicates that the TELD-3 serves five purposes. These include to identify children who are significantly below their peers in language development,[2] to identify strengths and weaknesses in language development, to document children's language progress resulting from intervention, to serve as a measure of language functioning for research purposes, and to supplement additional assessment procedures.

The examiner's manual of the TELD-3 reports criterion-related validity correlation values to eight other tests of child language. The correlation values for the Spoken Language Composite of Form A relative to other norm-referenced tests, particularly those which assess language broadly, are of interest to this investigation. The TELD-3 reports that these correlation values are statistically significant. The values reported include .77 with the Total Language Composite of the *Clinical Evaluation of Language-Fundamentals, Preschool* (CELF-P2; Wiig, Secord, & Semel, 1992), .76 with the Spoken Language Composite of the *Test of Language Development-Primary, Third Edition* (TOLD-P3; Newcomer & Hammill, 1997), and .61 with the Total Language Composite of the *Preschool Language Scale-Third Edition* (PLS-3; Zimmerman, Steiner, & Pond, 1992).

[2] The TELD-3 test manual does not provide information (i.e., sensitivity and specificity) to be able to determine the test's diagnostic accuracy.

Table 3
Descriptive ratings for standardized test scores.

TELD-3		UTLD-4	
Score range	Classification	Score range	Classification
>130	Very superior	131–165	Very superior
121–130	Superior	121–130	Superior
111–120	Above Average	111–120	Above average
90–110	Average	90–110	Average
80–89	Below average	80–89	Below average
70–79	Poor	70–79	Poor
<70	Very poor	35–69	Very poor

Note: TELD-3, Test of Early Language Development, third edition; UTLD-4, Utah Test of Language Development, fourth edition; no children in this investigation obtained a standard score below 35 or above 165 on the UTLD-4. Therefore, this minor discrepancy between the TELD-3 and the UTLD-4 did not affect the results of this investigation.

A table is provided within the TELD-3 manual to convert a child's standard scores (or quotients) on the Receptive Language and Expressive Language subtests as well as the Spoken Language Composite to a proficiency category. This conversion does not differ for either the subtests of the Spoken Language Composite. For the purpose of this study, Spoken Language Composite scores, which encompassed both subtests, were used. Table 3 specifies the standard score boundaries and associated language proficiency labels provided within the TELD-3 test manual.

2.2.2. Utah Test of Language Development-Fourth Edition (UTLD-4)

The *Utah Test of Language Development-Fourth Edition* (UTLD-4; Mecham, 2003) was normed on 841 children who were representative of the U.S. population based on that reported in the U.S. Bureau of the Census (1999) *Statistical Abstracts of the United States*. It is designed to measure the oral language skills of children between the ages of 3 years, 0 months and 9 years, 11 months. It consists of five subtests, Picture Identification, Word Functions, Morphological Structures, Sentence Repetition, and Word Segmentation[3] that combine to assess semantics, morphology, syntax, and phonology. The subtests combine to form a Content Composite, a Form Composite, and a Total Language Composite. The Content Composite is formed by combining the standard scores of the Picture Identification and Word Functions subtests. The Form Composite is determined by combining the Morphological Structures, Sentence Repetition, and Word Segmentations subtests. The Total Language Composite is formed by combining all five subtests and is purported by the author to be the best estimate of a child's spoken language ability. The Content Composite, Form Composite, and Total Language Composite scores are based on a mean of 100 and a standard deviation of 15.

The examiner's manual indicates that the UTLD-4 can be used to identify children who need special assistance due to their language delay or disorder,[4] to diagnose a child's strengths and weaknesses in language functioning, to be used as a screening metric, to guide intervention efforts in a general manner, and to objectively assess the effectiveness of a treatment program over time. The examiner's manual also indicates that this test can be used as a tool for research investigations.

The examiner's manual of the UTLD-4 reports criterion-related validity correlation values to two other test's of child language, specifically the *Test of Language Development-Primary, Third Edition* (TOLD-P3; Newcomer & Hammill, 1997) and the *Kindergarten Language Screening Test, Second Edition* (KLST-2; Gauthier & Madison, 1998). Relevant to this investigation, the correlation values for the Total Language Composite scores of the UTLD-4 were of interest. The correlation values of the Total Language Composite of the UTLD-4 were .84 and .87 with the Spoken Language Composite of the TOLD-P3 and the Total Language Composite of the KLST-2, respectively.

Two tables are provided within the UTLD-4 examiner's manual to convert standard scores to language proficiency categories. One table provides these descriptive categories for children's subtest scores. The second table provides these descriptive proficiency categories for the composite scores including the Content Composite, Form Composite, and Total Language Composite. Participants' scores on the Total Language Composite were used in this investigation. Table 3 specifies the standard score boundaries and associated language proficiency labels provided within the UTLD-4 test manual.

2.3. Procedures

The majority of participants completed the assessments in a separate room at their preschool and daycare settings. Four participants, three children with SLI and one typically developing child completed the assessment protocol in their home setting. The children were tested individually by a person trained in test administration. Consistent with the purpose of this study, 16 typically developing (TD) children and 16 children with SLI were administered the *Test of Early Language Development; Third Edition* (TELD-3; Hresko et al., 1999) and the *Utah Test of Language Development; Fourth Edition* (UTLD-4;

[3] The Word Segmentation subtest is only given to children who are age 4 years, 0 months or older. Consequently, if the child is younger than 4, the Total Language Composite consists of the remaining 4 subtests only.

[4] The UTLD-4 test manual does not provide information (i.e., sensitivity and specificity) to be able to determine the test's diagnostic accuracy.

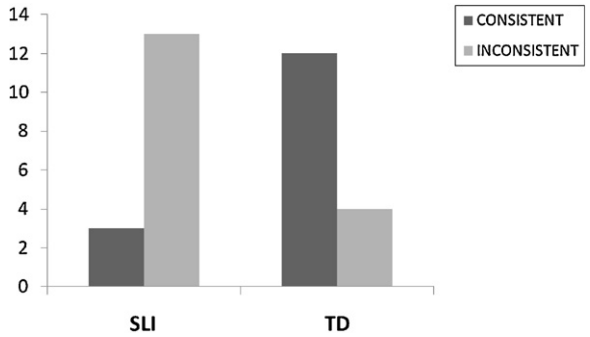

Fig. 1. Severity consistency: TELD-3 vs. UTLD-4.

Mecham, 2003) in counterbalanced order. Half the children in both the TD and SLI groups received the TELD-3 followed by the UTLD-4, while the remaining half were administered these tests in reverse order. Children were administered the TELD-3 and UTLD-4 on separate days. The average length of time between the administrations of these tests was 2.56 days (SD = 1.36) for the SLI group and 2.63 days (SD = 1.45) for the TD group.

The boundary cut-offs described within the respective test manuals were used to classify each child's language proficiency based on their performance on each of these tests. The consistency of proficiency categorization between the two norm-referenced tests was determined for each child in this study.

2.4. Reliability measures

Inter-examiner reliability data were collected for 9 of the 36 TELD-3 tests and 8 out of the 36 UTLD-4 tests administered, for a combined total of 23%. Point-to-point reliability was calculated to be 94.3%. Discrepancies in responses recorded were resolved by referring to the test manual for proper scoring procedures or by deferring to the individual with the better view of the child's response when this was deemed the source of the discrepancy.

3. Results

3.1. Mean performance

The mean performance of the SLI group was 84.75 (SD = 16.82) on the TELD-3 and 80.56 (SD = 6.75) on the UTLD-4. The mean performance of the TD group was 114.31 (SD = 13.81) on the TELD-3 and 109.75 (SD = 12.20) on the UTLD-4. The mean of the SLI group on the TELD-3 in this investigation is slightly higher than that for the "Delayed Language" group (mean = 77, SD = 13) reported in the examiner's manual, while the mean of the SLI group on the UTLD-4 in this study is slightly lower than that reported for the "Low Language Achievement" group reported in the UTLD-4 examiner's manual (mean = 85, SD = unreported). The TD group's performance on both of these tests was higher than that reported for the normative samples in the respective test manuals. However, the normative samples included disordered subjects, which depresses the group mean relative to a typically developing only normative samples (Peña, Spaulding, & Plante, 2006). Therefore, we would expect our means for the TD group to be higher than that of the normative samples.

3.2. Severity category consistency: TELD-3 vs. UTLD-4

The consistency of classifications assigned to the children based on test scores on the TELD-3 and the UTLD-4 were examined. A two-way Chi-square analysis was conducted to determine if consistency in language proficiency determinations was related to group. The results indicated a significant effect ($X^2 = 10.17$, $df = 1$, $p = .001$) (see Fig. 1). Consistency in proficiency category assignments was more likely for the TD group than for the SLI group. When using the specified cut-offs for determining severity of impairment provided within the TELD-3 and UTLD-4 test manuals for the TD group, 12 out of the 16 (75%) proficiency category assignments were consistent, indicating that 25% of the language proficiency designations for the TD group differed between the two tests as a function of TD children's performance. In contrast, for the SLI group 3 (19%) of the children's classifications of language proficiency remained consistent between the two tests, while 13 (81%) of their language classifications differed depending on whether the language proficiency classification was derived from the TELD-3 or the UTLD-4.

For the SLI group, 8 children's scores were characterized as "average" on the TELD-3. Based on their UTLD-4 scores, 4 of these children were classified as having "below average" language skills while the remaining 4 were classified as having "poor" language skills. Three children with SLI were classified as exhibiting "below average" language skills based on their TELD-3 performance. Of these children, two were likewise classified with "below average" language skills on the UTLD-4,

while the remaining child's language skills were characterized as "average" by the UTLD-4. The language skills of one child with SLI were rated as "poor" by both the TELD-3 and the UTLD-4. An additional child in the SLI group who was classified with "poor" language skills on the TELD-3 was characterized as exhibiting "very poor" language skills on the UTLD-4. Finally, two children with SLI who were characterized with "very poor" language skills on the TELD-3 were characterized with "below average" language skills on the UTLD-4, while the remaining child with SLI characterized as "very poor" on the TELD-3 was classified with "average" language skills on the UTLD-4.

For the TD group, 6 children who achieved an "average" language skill proficiency rating on the TELD-3 obtained the same identical "average" rating on the UTLD-4. Likewise, 4 children in the TD group assigned an "above average" rating on the TELD-3 obtained an "above average" rating on the UTLD-4. Another consistency documented in the TD group was for two children who received a "superior" language proficiency characterization on both the TELD-3 and the UTLD-4. Two children in the TD group who were classified as presenting with "above average" language skills on the TELD-3 obtained "superior" classifications of their language functioning on the UTLD-4. One child in the TD group who was also classified with "above average" language skills on the TELD-3 obtained an "average" rating based on their performance on the UTLD-4. Finally, one child in the TD group who obtained a "superior" rating on the TELD-3 was classified as "very superior" using their performance on the UTLD-4.

4. Discussion

Speech language pathologists are encouraged by a number of U.S. State Departments of Education to use children's norm-referenced test performance to assist in determining severity of impairment (Spaulding et al., in press). It is therefore useful for practitioners to know how similarly children's language proficiencies are classified across these measures in order to accurately interpret severity results. Lack of consistency in severity classifications across these measures has implications, not only for the selection and interpretation of these instruments, but also for the development of evidence-based guidelines for evaluating severity of impairment.

Despite the ease with which clinicians can use information provided within a number of norm-referenced test manuals to convert a child's score to a severity rating, evidence supporting the selection and use of norm-referenced tests to determine impairment severity is lacking. The findings of this study indicate that language proficiency ratings change depending on which test, the TELD-3 or the UTLD-4, is selected for interpretation. Despite the fact that both tests use identical cut-off points for proficiency classifications, differences in language proficiency characterizations were apparent. Importantly, although most of the typically developing children's proficiency ratings remained consistent, the majority of severity of impairment designations for children with SLI changed depending on which test was selected for this purpose. Unlike what a number of U.S. State Departments of Education are suggesting to do (see Spaulding et al., in press), these results provide empirical evidence against applying the same boundaries to determine language proficiency, particularly severity of impairment for children with SLI, across tests of child language. This adds to prior work by Ballantyne et al. (2007), who identified differences in language severity classification for older, school-age children with SLI on different versions of the same test.

There are reasons why children with SLI, in particular, may perform sufficiently different enough on the TELD-3 and UTLD-4 to obtain different severity of impairment ratings. First, the TELD-3 and UTLD-4 were developed from different theoretical frameworks. Consequently, although both measure language proficiency in children, they differ to some extent in their format and content. For example, the TELD-3 is designed to assess semantics, morphology, and syntax skills while the UTLD-4 assesses semantics, morphology, syntax, and phonology. Given that the UTLD-4 assesses the phonological aspect of language while the TELD-3 does not, a child with phonological issues would likely perform worse on the UTLD-4 than the TELD-3. However, the word segmentation subtest of the UTLD-4, which is designed to assess the phonological aspect of language, is given only to children who are at least 4 years of age. Of the nine children with SLI who scored lower on the UTLD-4 than the TELD-3, five of them were old enough to be administered the word segmentation subtest. In order to determine whether or not performance on the word segmentation subtest negatively affected the severity rating of these five children with SLI, severity ratings were compared based on including or excluding the word segmentation subtest. The examiner's manual of the UTLD-4 provides explicit information on how to do such a conversion. Indeed, the severity rating did change for three of the children with SLI, evidenced by lower severity ratings with than without the word segmentation subtest. However, the addition of a phonological subtest on the UTLD-4 cannot account for the remaining 10 children with SLI (63% of the original SLI group) whose severity ratings differed based on their UTLD-4 and TELD-3 performance. In addition, factors beyond the content of test items can affect the consistency in performance between these two tests. For example, Merrell and Plante (1997) administered two norm-referenced tests to preschool children with SLI, and found inconsistent inter-test agreement on items which assessed the same morphosyntactic structures. Therefore, variation in how language targets are assessed between the TELD-3 and the UTLD-4 may also impact severity consistency. However, although it is important to consider underlying causes for the differences in severity ratings obtained on assessments of child language, it does not change the fact that the children's severity ratings differ depending upon which test is administered.

The criterion-related validity section of the test examiner's manual will describe the comparability of a given test to other norm-referenced assessments. The publishers of both the TELD-3 and UTLD-4 do, in fact, report high inter-test correlations as evidence to support the validity of these measures for clinical use. Neither manual, however, provides information on the relationship expected between these two tests of child language. The TELD-3 and the UTLD-4 each measure multiple

dimensions of language, so we would expect the relationship between performances on these two assessments to be relatively consistent with the correlations with other tests of child language documented within their examiner manuals. However, even if there was a perfect correlation between performance on the TELD-3 and the UTLD-4, scores would not have to be consistent with each other. They would just have to change at the same rate. Therefore, using criterion-related validity to determine consistency in severity classifications between assessments is not recommended.

In order for a norm-referenced test of child language to be able to successfully determine severity of language impairment, it has to be able to make fine distinctions in language skill. Consequently, it must include a high number of items in order to make such distinctions. It is unlikely that a test would be successful at making such determinations across the language proficiency spectrum, specifically ranging from very poor to very superior language skills as both tests in this investigation indicate, without having an exorbitant number of items. Therefore, a test developed to assess severity of language impairment would likely be better at doing so if it was normed on a distribution of children with language impairment in order to include items which would be focused on skill levels which differentiate children with different degrees of impairment. Likewise, a test developed to assign language proficiency categories to children with normal range or higher language skills would be better at doing so if it was normed on typically developing children and consequently target items which would differentiate children at these higher language levels.

Importantly, both the TELD-3 and the UTLD-4 base their language proficiency designations on a test taker's score relative to the normative sample. Given that the normative samples consist for the most part of typically developing children, it is not too surprising that proficiency characterizations for the typical language group in this study remained relatively consistent for the vast majority, 75%, of this group of children. We might have expected greater variability in proficiency categorization for children with impairments, including those with SLI who participated in this investigation, because fewer items were likely included for administration to make such fine distinctions at their language levels. Indeed this variability was apparent in the data.

The results of this investigation indicate that although a number of norm-referenced tests provide information for determining severity of impairment within their manuals, this does not mean that clinicians (or researchers) should give much weight to severity designations derived from test results. First, this investigation found that severity ratings assigned to the vast majority of children with SLI, in particular, differed as a function of which norm-referenced test was administered. Furthermore, the likelihood that severity ratings assigned to children, by virtue of their score on a static measure of language functioning, truly reflects or approximates the severity of their impairment is low. This is particularly true when severity ratings are based on single test batteries. The impact of the child's language skills on their everyday functioning would be a better metric of severity of impairment than the limited ecological validity offered by a standardized test score.

The importance of severity determinations for children with language impairment is ultimately determined by their impact on service provision. With respect to service provision in the United States, the Individuals with Disabilities Education Act, 2004 (IDEA, 2004) regulation part 300 A 300.8I11, states that in order for a child to be eligible for language services, the language impairment must "adversely affect a child's educational performance". This act says nothing about severity of impairment as a contributing factor to such determinations. Clearly, using severity decisions to determine qualification for service is not supported by this regulation. However, the determination of a child's severity of impairment may contribute to clinical decisions regarding the best treatment approach, the priorities for intervention, or prognostic expectations. Future work can help to clarify the impact of language severity determinations on clinical intervention.

5. Conclusions

This work underscores the importance of adhering to clinical practices with an evidence-base. The findings of the present study are valuable for clinicians and researchers who use norm-referenced test performance to determine or to supplement other assessment procedures for determining severity of language impairment in children. The inconsistency in severity of impairment for children with SLI determined through the procedures specified within two norm-referenced test manuals indicates that clinicians need to be judicious in selecting norm-referenced tests to inform severity of impairment decisions. In addition, future policies adopted by agencies which set standards for determining impairment severity should carefully consider the accumulating evidence against using identical boundaries across norm-referenced tests of child language to determine severity of language impairment in children without reference to how children with impairment perform on the tests.

Acknowledgement

This work was supported by a Large Faculty Grant from the UConn Foundation.

Appendix A. Continuing education

Comparison of severity ratings on norm-referenced tests for children with specific language impairment

1. Results from studies investigating the performance of children with language impairment on norm-referenced tests of child language have found all but which of the following?
   a. Studies have found that these children perform differently on different tests of child language
   b. Studies have found that their performance is reflective of their long-term prognosis
   c. Studies have found that they perform differently on the same test across different ages
   d. Studies have found that they obtain different severity ratings on different versions of the same test
2. The author was interested in assessing the utility of norm-referenced tests for determining severity of impairment because of which of the following policies?
   a. A number of State Departments of Education indicate for clinicians to use norm-referenced tests to assist in determining the severity of a child's language impairment
   b. The Individuals with Disabilities Education Act of 2004 specifies to use norm-referenced tests to assist in determining the severity of an impairment
   c. The Individuals with Disabilities Education Act of 2004 specifies that severity of impairment must be used, with certain exceptions, for determining whether or not a child with a language impairment qualifies for services
   d. All of the above
3. The author was interested in assessing the use of norm-referenced tests for determining severity of language impairment because a number of norm-referenced tests include which of the following?
   a. A table to convert a child's test score to a language severity classification
   b. A normative table depicting the distribution of performance of children with language impairment
   c. A sufficient number of items to finely differentiate language skill
   d. All of the above
4. The results of this investigation found which of the following?
   a. The vast majority of language proficiency classifications for typically developing children and children with language impairment remained stable between the two tests of child language
   b. The vast majority of typically developing children's language proficiency classifications remained stable, while the vast majority of language proficiency classifications for the children with language impairment differed between the two tests of child language
   c. The vast majority of language proficiency characterizations for the children with language impairment remained stable, while the vast majority of language proficiency classifications for the children with typical language differed between the two tests of child language
   d. The vast majority of language proficiency classifications for both typically developing children and children with language impairment differed between the two tests of child language
5. The authors suggest caution in using norm-referenced test performance to determine severity of language impairment for which of the following reasons?
   a. Both preschool and school-age children with language impairment obtain different severity ratings on tests of child language
   b. Norm-referenced test scores represent a static measure of language skill, not their potential for learning
   c. Norm-referenced test performance does not sufficiently capture the impact of the child's language skills on their everyday functioning.
   d. All of the above

References

American National Standards Institute. (2004). *Specifications of audiometers* (ANSI S3.6-2004). 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.

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.

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

Gauthier, S. V., & Madison, C. L. (1998). *Kindergarten Language Screening Test* (second edition). Austin, TX: Pro-Ed.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), 196–206.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). *Test of early language development* (third edition). Austin, TX: Pro-Ed.

Individuals With Disabilities Education Act. (2004). § 300 A 300.8I11 § 2004.

Kaufman, A., & Kaufman, N. L. (2004). *Kaufman assessment battery for children* (second edition). Circle Pines, MN: AGS Publishing.

Mecham, M. J. (2003). *Utah Test of Language Development* (fourth edition). Austin, TX: Pro-Ed.

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.

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development Primary* (third edition). Austin, TX: Pro-Ed.

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.

Rice, M., Ash, A., Abel, A., & Lee, J. (2008). A comparison of children with SLI and control children on the PPVT-R and PPVT-III: Effects of test revision on sensitivity to affectedness. Poster presentation at the Symposium on Research in Child Language Disorders, Madison, WI.

Semel, E., Wiig, E. H., & Secord, W. (1987). Clinical Evaluation ofLanguage 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.

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.

Spaulding, T. J., Swartwout Szulga, M. J., & Figueroa, C. (in press). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy-makers and test developers. Language, Speech, and Hearing Services in Schools.

The National Data Bank of U.S. Bureau of the Census. (1997). *Statistical abstracts of the United States*. Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census. (1999). *The statistical abstract of the United States*. Washington, DC: Author.

Wiig, E. H., Secord, W. A., & Semel, E. (1992). *Clinical evaluation of language fundamentals—Preschool*. San Antonio, TX: The Psychological Corporation.

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical evaluation of language fundamentals—Preschool* (second edition). San Antonio. TX The Psychological Corporation.

Zimmerman, I., Steiner, V., & Pond, R. E. (1992). *Preschool Language Scale* (third edition). San Antonio, TX: The Psychological Corporation.