Published in final edited form as:
J Speech Lang Hear Res. 2011 December ; 54(6): 1597–1608. doi:10.1044/1092-4388(2011/10-0213).

Concurrent and Construct Validity of Oral Language Measures with School Age Children with Specific Language Impairment

LaVae M. Hoffman,
University of Virginia

Diane Frome Loeb,
University of Kansas

Jayne Brandel, and
Fort Hays State University

Ronald B. Gillam
Utah State University

Abstract

Purpose: This study investigated the psychometric properties of two oral language measures that are commonly used for diagnostic purposes with school age children who have language impairments.

Methods: 216 children with SLI were assessed with the *Test of Language Development-Primary, 3rd edition* (TOLD-P:3) and the *Comprehensive Assessment of Spoken Language* (CASL) within a three-month period. The concurrent and construct validity of these two published tests were explored through correlation analysis and principle component factor analysis.

Results: The TOLD-P:3 Spoken Language Quotient and CASL Core Composite scores were found to have an inter-test correlation value of $r = .596$ within this sample, and a paired samples t-test revealed a statistically significant difference between these scores. Principle component factor analyses revealed a two factor structure solution for the TOLD-P:3, while data from the CASL supported a single factor model.

Conclusions: Analyses of assessment measure performance data from a sample of school age children with specific language impairment revealed concurrent validity values and construct validity patterns that differed from those found in the norming samples as cited in examiner's manuals. Implications for practice patterns and future research are discussed.

Correspondence can be sent to LaVae M. Hoffman at lmh3f@virginia.edu..
Author Note

The research reported in this article was funded by a grant from the National Institutes of Health, National Institute on Deafness and Communication Disorders (NIH 1-RO1-DC04560).
The authors extend their appreciation to Randall R. Robey at the University of Virginia for his support and encouragement during the final manuscript preparation.

Publisher's Disclaimer: This is an author-produced manuscript that has been peer reviewed and accepted for publication in the *Journal of Speech, Language, and Hearing Research* (*JSLHR*). As the "Papers in Press" version of the manuscript, it has not yet undergone copyediting, proofreading, or other quality controls associated with final published articles. As the publisher and copyright holder, the American Speech-Language-Hearing Association (ASHA) disclaims any liability resulting from use of inaccurate or misleading data or information contained herein. Further, the authors have disclosed that permission has been obtained for use of any copyrighted material and that, if applicable, conflicts of interest have been noted in the manuscript.

Keywords

language impairment; school age children; construct validity; concurrent validity

Introduction

Published, norm-referenced assessment instruments are often considered essential diagnostic tools for practitioners and researchers. These diagnostic measures have standardized administration, scoring, and interpretative procedures that have been developed through psychometric processes that establish their validity, reliability, and applicability for specified populations. Consistency across diagnostic measures has implications for the selection and interpretation of instruments and procedures, particularly with regard to determining eligibility for services and planning intervention. The greater the variability among measures, the more knowledgeable practitioners must be about the theoretical bases, psychometric properties, and clinical impact of each assessment instrument, and the greater the need for research evidence that supports each measure.

The amount of research evidence available to support speech-language pathologists' (SLPs') use of published tests is highly variable across diagnostic measures and age groups. Investigations of concurrent validity for published tests of oral language have been actively undertaken with toddler and preschool age children and are regularly reported in professional journals (e.g., Allen & Bliss, 1987; Chiat & Roy, 2007; Dale, 1991; Feldman et al., 2000; Gray, Plante, Vance, & Hendrichsen, 1999; Marchman & Martinez-Sussmann, 2002; Perona, Plante, & Vance, 2005; Plante & Vance, 1995; Rescorla, Ratner, Jusczyk, & Jusczyk, 2005; Restrepo et al., 2006; Roman, 1980; Sturner, Heller, Funk, & Layton, 1993; Thal, O'Hanlon, Clemmons, & Fralin, 1999; Tomblin, Shonrock, & Hardy, 1989; Wetherby, Allen, Cleary, Kublin, & Goldstein, 2002). However, independent investigations that would expand our understanding of the psychometric properties of standardized tests that are commonly administered to school age children are lacking. This circumstance is particularly concerning given that a recent American Speech-Language-Hearing Association (ASHA) report indicated that more than half of all SLPs in the United States are employed in school settings (ASHA, 2009).

The emphasis on evidence based practice in speech language pathology has spurred a new interest in empirically documenting how diagnostic measures operate clinically (e.g., Ballantyne, Spilkin, & Trauner, 2007; Bruckner, Yoder, Stone, & Saylor, 2007; Friberg, 2010; Gray, et al., 1999; Greenslade, Plante, & Vance, 2009; Pankratz, Plante, Vance, & Insalaco, 2007; Peña, Spaulding, & Plante, 2006; Perona, et al., 2005; Spaulding, Plante, & Farinella, 2006; Tomblin, 2008; Wetherby, et al., 2002). However, our professional evidence base does not have published accounts of the independent verification of concurrent and construct validity for published tests that are widely used in clinical practice with elementary school age, monolingual English-speaking children, despite the fact that these instruments are frequently used to support clinical decisions.

Practitioners have more assessment measures available to them than ever before, and federal mandates require that multiple sources of assessment data be considered when determining the presence of a disability and eligibility for special education services. Clinically, it would be helpful to practitioners to know how similarly children with language deficits can be expected to perform across various measures in order to accurately interpret assessment results. Instruments that have been developed by the same author or publisher, as are often used to establish validity during the norming process, are likely to be similar in theoretical orientation to language and measurement formats, and may therefore be more likely to yield higher degrees of association than measures that are developed and distributed by differing authors or publishers. Ideally, there should be independent documentation of how well instruments from different authors and publishers perform relative to one another, especially if they purport to measure similar language constructs. Moreover, practitioners often need to evaluate relative strengths and weaknesses of communication skills (McCauley & Swisher, 1984) across measures or procedures, including published tests. To make informed decisions, SLPs need independent evidence about how various diagnostic tools operate relative to each other, particularly with children who have language impairments. This information has been largely unavailable to date.

A study of concurrent validity involves extensive testing and data collection for a large number of children with well-defined language abilities. This is expensive and time consuming, so the lack of this information in our literature is not surprising. While completing a randomized controlled trial (RCT), we were able to systematically collect a variety of language data from a large group of school age children with specific language impairment (SLI) in three metropolitan areas. This dataset afforded us the unique opportunity to analyze the strength of association between measures of oral language ability. Specifically, this dataset provides the opportunity to examine two psychometric properties, concurrent validity and construct validity, of the assessment instruments that were used. *Concurrent validity* refers to the consistency of results between two tests that assess the same skill. *Construct validity*, on the other hand, is a measure's ability to accurately reflect the conceptual foundation upon which it has been developed. In other words, if a measure has good construct validity it appears to be measuring the abilities that it was designed to measure. Analyses of the RCT data with regard to concurrent and construct validity issues could provide some of the independent evidence needed to support clinical decisions during the diagnostic process.

Within the broad arena of contributing to the research-based evidence that supports the profession of speech language pathology, this study has two aims.

1. To independently test and empirically describe the concurrent validity of two norm-referenced, standardized measures of language skills as evidenced in school age children with specific language impairment.
2. To independently test and empirically describe the construct validity of two norm-referenced, standardized assessment instruments that are commonly used with school age children who have specific language impairment.

We hypothesized that because the *Test of Language Development-Primary, 3rd Edition* (TOLD:P-3: Newcomer & Hammill, 1997) and the *Comprehensive Assessment of Spoken Language* (CASL: Carrow-Woolfolk, 1999) are decontextualized, norm-referenced measures of overall oral language abilities, performance on these measures would be significantly and positively correlated when both tests are completed by children who have language learning deficits. However, these two instruments measure multiple dimensions of language from differing theoretical frameworks, so the degree of association between them may not be as high as the criterion validity correlation values reported in their respective manuals. For example, the examiner's manual of the TOLD-P:3 reported that the correlation values of its six primary subtests were statistically significant and ranged from .64 to .97 relative to scores on the *Bankson Language Test – Second Edition* (Bankson, 1990, p. 77) when administered to 30 students attending first, second, and third grades in an Austin, Texas public school. The examiner's manual of the CASL reported that its Core Composite scores had statistically significant and positive correlations with the Oral Composite score from the *Oral and Written Language Scales* (OWLS: Carrow-Woolfolk, 1995) at .80, and with the Total Score from the Test of Auditory Comprehension of Language –Revised

(TACL-R: Carrow-Woolfolk, 1985) “above .70 with the exception of the Syntax Construction score” (p. 131). CASL publishers reported that inter-test data were collected for the OWLS with 50 children in second through fifth grades from across the nation, and for the TACL-R with 35 children in preschool through first grade from across the nation. Hence, the publishers of both instruments cited high inter-test correlations as strong evidence to support the validity of their measures for clinical use.

The issue of construct validity is more complex. Based on a factor analysis of data from the measure’s standardization sample, publishers of the TOLD-P:3 reported that children’s performance on this test’s core subtests loaded on a single factor, which they interpreted to be a “measure of general spoken language ability.” Publishers of the CASL reported that their standardization sample also yielded a single significant factor loading for children up to the age of 6 years. However for older children, they reported a three factor structure (Lexical/Semantic, Syntactic, and Supralinguistic). Of particular relevance to the current investigation, the publishers described their CASL norming data for 7 to 10 year old children as fitting a complex, three structure model “only moderately well” (p. 130). On the basis of Tomblin and Zhang’s (2006) recent findings that data from longitudinal and cross-sectional studies fit a unidimensional model of language ability in elementary years, we hypothesized that TOLD-P:3 and CASL performance data for 6- to 8-year-old children with language impairment would probably reveal a single, undifferentiated language factor.

The independent examination of these psychometric properties will provide empirical evidence to inform our understanding of how the TOLD-P:3 and CASL operated within the clinical sample for which they were designed. Moreover, regardless of the outcome of the factor analyses, an exploration of factor structures in performance data from two measures completed by a large sample of school age children with SLI may further the theoretical discussion of underlying language constructs in school age children with language impairment.

Method

Participants

The participants for this study were identified and recruited as part of a randomized controlled trial (RCT) to investigate the efficacy of three language interventions with school age children in comparison to a control condition. Children were recruited by meeting with speech language pathologists, special educators, and classroom teachers in public schools within the regions of the study. To increase minority participation, we targeted public schools that had higher percentages of minority enrollment. We described the kinds of children we were looking for as those who were performing at or below the 10th percentile in listening and/or speaking skills, and who did not have intellectual disabilities, autism spectrum disorders, or other health impairments. SLPs, special educators, and classroom teachers were asked to send informational brochures home to children they considered similar to our description of the kinds of children for whom the study was designed. Parents who responded to the brochure were contacted by a researcher who explained the study and provided each parent with written material describing the study, a consent form for identification testing, and a case history questionnaire.

Children whose parents returned the permission form and the questionnaire received the preliminary identification testing consisting of the TOLD-P:3, the *Kaufman-Brief Intelligence Test* (K-BIT: Kaufman & Kaufman, 1990) Matrices scale, vision and hearing screenings, oral mechanism screening, and a review of school records. Inclusionary criteria included nonverbal intelligence within normal limits as demonstrated by standard subtest scores between 75 and 125 (+/−1.66 SD) on the K-BIT Matrices scale, and presence of a language deficit as demonstrated by standard scores at or below 81 on two or more composites of the TOLD-P:3. Descriptive statistics for the norm-referenced identification measures can be found in Table 1. In addition, adequate vision and hearing status was determined through educational records documenting screening results within the past year. When educational records could not provide documentation of vision or hearing status, live administration using Lea Symbols vision screening materials to ensure adequate vision in at least one eye with or without corrective lenses and/or hearing screenings at 20 dB HL at the frequencies of 1, 2, and 4 KHz in both ears were completed. Recruitment activities began annually in early spring. Identification testing usually occurred during April and May of each year, and preintervention measures were completed in the first week of June. A detailed description of the Comparison of Language Intervention Programs (CLIP) research project was recently reported by Gillam and colleagues (2008).

Assessment during the identification phase documented that each participant met the EpiSLI criterion of measured oral language abilities that were at or below a standard score of 81 on two or more composite quotients on a standardized measure of oral language abilities (Tomblin et al., 1997). “EpiSLI” is the abbreviation for the diagnostic criterion that was established by Tomblin and his colleagues through a large-scale epidemiological investigation funded by the National Institutes of Health (NIH); the first purpose of which was to “establish a definition of specific language impairment (SLI) that is consistent with current research and clinical experience, and, in accordance with this definition, develop an explicit criterion for the diagnosis of SLI” (Tomblin, 2010, p. 109). The EpiSLI criterion was adopted for use in the RCT because the longitudinal data resulting from the epidemiological study established this diagnostic criterion as the definition of SLI with the most empirical evidence at that time.

In addition to significantly impaired language performance on the TOLD-P:3, the participants also met the other EpiSLI exclusionary criteria which ruled out sensory impairment, gross neurological or psychological disorders, and linguistic differences due to second language acquisition or limited opportunity to learn language. Specifically, in order to be eligible for participation in the study, each child was documented as having a predominantly English-speaking home, adequate vision and hearing, adequate oral structure and function for the purposes of speech, normal gross neurological function, normal nonverbal intelligence along with an absence of debilitating psycho-social disorders, while also demonstrating significant language deficits that were not attributable to a previously diagnosed disorder, disease or syndrome. Children who had been previously diagnosed with learning disabilities, dyslexias or attention deficits were eligible for participation in the RCT. Data from all of the children who participated in the RCT were included in the current study.

Of the 216 participants, 58 were 6 years of age, 78 were 7 years old, and 80 were 8 years old, yielding a mean age of 7 years; 6 months, and a standard deviation of 0.9 years for the participant group as a whole. There were 136 male participants and 80 female participants, creating a 1.66:1 ratio of males to females. The majority of participants were reported by their parents to be White, not Hispanic, (40%), followed by Black or African-American (30%), Hispanic or Latino (17%), more than one race (5%), American Indian (2%), and Asian (0.4%). Eight percent of parents did not report ethnicity information for their children. Most children who participated in the RCT had one or more parents who had attended at least one year of college (44%), or had graduated from college (37.5%), while the parent(s) of 19.5% of the participants had a high school education or less.

Assessment Data

The current study used only a subset of assessments from the RCT. Identification phase data used for this investigation included the standard scores for each participant on the TOLD-P: 3, including all core subtests and the Spoken Language composite quotient. Although the RCT pre-intervention assessments included a variety of measures that are reported in greater detail by Gillam et al. (2008), the current study concerns only data from the TOLD-P:3 subtests and composites from the identification testing period and the CASL subtest and Core Composite standard scores from pre-intervention data collection. Descriptive statistics for TOLD-P:3 and CASL global standard scores and time between testing are provided in Table 1. Note that all data used in the current investigation were gathered prior to the initiation of the intervention arms of the RCT; therefore, the possibility of a treatment effect does not exist in the analyses being reported.

For the current study, we explored concurrent validity between the TOLD-P:3 and CASL by using the Pearson *r* coefficient metric to examine the degree of association between global performance scores on these two norm-referenced oral language assessment instruments. Although both of these published assessment instruments purport to yield valid and reliable indices of overall oral language abilities, and are commonly used with school age children for diagnostic and eligibility purposes, each was developed from differing theoretical perspectives and uses different tasks to measure various components of oral language.

The TOLD-P:3 was developed for use with children between the ages of 4 years, 0 months and 8 years, 11 months. Although the authors of the TOLD-P:3 describe using a linguistic frame of reference, they claim that the instrument does “not adhere to any specific theoretical perspective” (p. 1). The conceptual model for this instrument is described in a two dimensional matrix that crosses “linguistic features” (semantics, syntax, phonology) with “linguistic systems” (listening, organizing, or speaking skills). Through the combination of six core subtests (Picture Vocabulary, Relational Vocabulary, Oral Vocabulary, Grammatic Understanding, Sentence Imitation, and Grammatic Completion) the instrument yields an overall composite standard score (Spoken Language Quotient), as well as five additional composite standard score quotients (Listening, Organizing, Speaking, Semantics, and Syntax). Table 2 provides descriptions of the TOLD-P:3 subtests. Three supplemental subtests augment the instrument but are not required for calculation of standard scores for the six TOLD-P:3 composite quotients, and were not administered as part of the original RCT. None of the instrument’s subtests attempt to assess the pragmatic system of language. Instead, the focus of the TOLD-P:3 is squarely on the structural aspects of language, specifically form and content.

The CASL, on the other hand, is intended for use with a wider age range of children and young adults, covering the range of ages between 3 years, 0 months and 21 years, 11 months. As reported in its examiner’s manual, the CASL was designed to assess language abilities according to the Carrow-Woolfolk Integrative Language Theory in which the use of language is considered in addition to the structural aspects of form and content. From this theoretical perspective, distinctly assessing both “language knowledge” (form and content) and “language performance” (including the internal systems used for both comprehension and expression) across four “linguistic categories” (semantics, syntax, pragmatics, and supralinguistics) is purported to provide a comprehensive analysis of the language elements that may or may not meet the child’s communicative needs. This instrument includes 15 subtests that are used in various combinations, and yields a Core Composite standard score for each of six age-specific groups. For the purposes of the original RCT study, the subtests that generate the CASL Core Composite scores for 5 to 6 year olds, and 7 to 10 year olds, were completed by all of the study participants. Those subtests were: Antonyms, Syntax Construction, Paragraph Comprehension, Nonliteral Language, and Pragmatic Judgment.

These CASL subtests are also described in Table 2. The remaining 10 CASL subtests were not administered, nor were CASL supplemental category or processing scores calculated. It is important to note that although none of the supplemental subtests were administered for the TOLD-P:3 or CASL, all of the core subtests were completed for each child on both measures. The TOLD-P:3 Spoken Language Quotient and CASL Core Composite standard scores constitute the primary overall test performance data for each measure as commonly used for research and clinical purposes. The analysis of these data provides another view of inter-test agreement, separate from the data that have been reported by publishers using different instruments.

Statistical Analyses

Individual performance data for all 216 participants in the RCT were included in the current analyses. Correlation coefficients were generated to determine the between-instrument associations of the TOLD-P:3 Spoken Language Quotient standard scores with CASL Core Composite standard scores, and a paired samples t-test compared mean scores on both measures. Because we wanted to examine the consistency of global language performance as documented across assessment measures, intra-instrument correlation coefficients among subtests and composite scores were not of interest in this investigation. Construct validity of the TOLD-P:3 and CASL was explored separately for each instrument using principle component factor analysis. All statistical analyses were completed with PASW for Windows, version 18 software (SPSS, 2009).

Results

Concurrent validity

Calculation of Pearson’s product moment coefficient revealed a statistically significant association between the overall composite standard scores for the TOLD-P:3 Spoken Language Quotient and the CASL Core Composite standard scores with $r = .596$, $p < .001$ (2-tailed), with a 95% confidence interval of 0.479 to 0.676. Calculation of the corresponding proportion of variance, $r^2 = .355$, revealed that approximately 36% of the variance in either measure could be accounted for by the other measure. Interestingly however, across these two norm-referenced measures of global language ability, that leaves 64% of variance unaccounted.

To further examine how these two assessment instruments operated, a paired samples t-test was conducted to compare performance across both tests. This revealed a statistically significant difference between mean scores on the TOLD:P3 Spoken Language Quotient (M=73.76 SD=8.62) and the CASL Core Composite standard scores (M=78.92, SD=11.42), $t_{(215)} = -8.1$, p=.001. This finding provides supportive evidence that, despite substantial correlation, these tests appear to be measuring aspects of language that are not identically accounted for by the other measure.

Of particular clinical interest is the examination of diagnostic consistency between these two instruments. To explore this issue, the pre-intervention CASL Core Composite score for each child was categorically coded as to whether or not it met the criterion for language deficit. Recall that each participant had met the Episli criterion of two or more composite standard scores at or below 81 on the TOLD-P:3. The CASL subtests used in the RCT study yield only a single composite score, the CASL Core Composite standard score. Two or more composite scores from the CASL would have been necessary to apply the EpiSLI standard for the purpose of categorical coding. However, in the original development of the EpiSLI diagnostic classification system, Tomblin and colleagues (1996) reported that a single composite score cutoff at −1.14 standard deviations (Tomblin, et al., 1996)compared similarly to the EpiSLI standard which required two composite scores. On this basis, the criterion of a standard score at or below 83 on the CASL Core Composite was used for assignment to “performance indicative of language impairment” or “language performance above the level of impairment” categories. Comparing the categorical classifications resulting from administration of each test, the degree of overlap for all 216 participants between performance on the identification measure (TOLD-P:3) and the pre-intervention measure (CASL) revealed that 64% of the children classified as having language impairment on the basis of performance on the TOLD-P:3 were also classified as having a language impairment on the basis of their performance on the CASL. These results are lower than might be expected on the basis of psychometric information contained within the instruments’ examiner manuals.

A potentially influencing factor could be the time elapsed between measures. Recruitment efforts for the RCT began annually in the early spring. Recall that identification testing (TOLD-P:3) usually occurred during April and May, but was earlier in some cases, while pre-intervention testing always occurred at the beginning of June. The mean number of calendar days between identification and pre-intervention testing was 79.97, with a standard deviation of 44.67, and a range of 13 to 265 days. The wide range was largely due to seven participants for whom the TOLD-P:3 had been completed within the six months prior to their referral for participation in the RCT study. For these seven children, rather than re-administering the TOLD-P:3, their previous TOLD-P:3 scores were accepted for the purposes of eligibility for the RCT. The net result of this recruitment decision was that the length of time between TOLD-P:3 and CASL testing was extended for these seven children and the inter-measure time range increased for the group as a whole. If these seven outliers were to be removed from the dataset for the current study, the mean number of days between identification and pre-intervention testing would have been 74.84, with a standard deviation of 34.78, and a range of 13 to 142 days; however, the diagnostic overlap between TOLD-P:3 and CASL performance would remain essentially unchanged, with 63% of children being classified as having a language impairment on both measures. Because of this negligible difference, data from all 216 RCT participants were included for the current study.

Construct validity

Unrotated principle components factor analysis of performance data for young school age children with language impairment on the two decontextualized measures of language abilities revealed differing factor structures from those reported by the authors of the published tests on the basis of analyses completed with data from the tests’ respective norming samples. In our data set, TOLD-P:3 subtest data converged to reveal two factors with eigenvalues above 1, as shown in Table 3. Examination of the corresponding component matrix, Table 4, revealed that while all of the TOLD-P:3 subtests loaded on the first factor, two subtests (Oral Vocabulary, and Relational Vocabulary) loaded heavily on a second factor. This finding stands in contrast to the factor analysis results reported by the TOLD-P:3 publishers in which factor reduction analysis of the instrument’s norming sample data resulted in a single significant factor. Our two factor solution would be a tidy finding if the six core subtests of the TOLD-P:3 loaded on the linguistic features (semantics and syntax) of the authors’ two-dimensional model of the language structure that was used to generate the TOLD-P:3 subtests (p. 6). However, our data did not yield this loading pattern. While most subtests loaded as expected according to the syntax/semantics dichotomy, the Picture Vocabulary subtest loaded more heavily with the syntax subtests than the other two semantic subtests. We will describe our interpretation of these two factors further in the Discussion section.

Principle component factor analysis of the CASL core subtest data from children who have language impairments, however, yielded a single significant factor with an eigenvalue of

3.32, which accounted for 66.5% of the performance variance in this measure. Recall that the author of this measure reported a three factor structure model using data from the normative sample.

Discussion

To date, the psychometric evidence to support the selection and use of tests of oral language skills in school age children has been predominately supplied by test publishers, and has been largely limited to publishers' investigations that use their own instruments during test development. This information is often based on comparisons with other instruments developed by the same author or distributed through the same publisher. This study presented the first known, large-scale, independent investigation into the concurrent validity of two measures of global oral language abilities with school age children who have SLI. These measures, the TOLD-P:3 and the CASL, were developed by different authors, reflect different theoretical constructs, and use differing elicitation procedures. In addition, this study empirically examined the construct validity of these norm-referenced assessment instruments. Of particular interest was establishing independent evidence about how these language measures operate within a sample of young school age children with language impairment, as this is a diagnostic group particularly relevant to many speech language pathologists.

Concurrent Validity

Concurrent validity is usually established by analyzing the robustness of associations across measures through correlation. In a seminal volume on statistical analyses in behavioral sciences, Cohen (1988) offers an operational guideline for interpreting correlation coefficients when a critical mass of earlier investigations is not yet available for use in establishing conventional expectations for a particular field or area of investigation. These interpretative guidelines were developed using datasets in a variety of behavioral sciences, including measures of intelligence, achievement, personality, and psychological disorders. On the basis of these datasets, he asserts that a correlation value of $r = .30$ might be interpreted as a medium effect size because "many of the correlation coefficients encountered in behavioral science are of this magnitude, and, indeed, this degree of relationship would be perceptible to the naked eye of a reasonably sensitive observer" (p. 80). He cites the field of psychology, where measures of personality correlate with "comparable real life criteria" (p. 81) around the $r = .30$ level as offering support for this interpretive standard. Cohen considers $r = .10$ to indicate a small but realistic effect size, presumably on basis that the complexity of human dimensions and measurement limitations attenuate correlation coefficients when measuring behavioral manifestations of theoretical constructs. He further asserts that $r = .50$ could be considered to be a large effect within behavioral sciences because correlations between measurements of constructs, such as IQ, and real life performance as indicated by school grades, consistently cluster around this value. He notes, however, that a possible exception to $r = .50$ as a realistically expected maximum correlation could be coefficients that are generated by the analysis of test form equivalency, where higher values would be expected.

Interpreting the findings of this study using the guidelines proposed by Cohen reveals that the statistically significant TOLD-P:3 and CASL inter-test correlation value of $r = .596$ would indicate a large degree of association for behavioral science measures that reflect complex abilities. Given that these two oral language assessment instruments were developed using different theoretical frameworks, the overall composite scores of these tests appear to be well-correlated as independently documented within a sample of young school age children with SLI. The current investigation did not replicate the publishers' procedures by administering the same instruments used during the TOLD-P:3 or CASL norming procedures. Instead, the present study explored the inter-test consistency between these two common clinical assessment instruments. Consistency between these two measures was not reported in the information contained in the publishers' manuals; therefore this investigation extends the psychometric evidence and did not attempt to replicate the publisher's data. However, the two instruments in this study (TOLD-P:3 and CASL) were not as highly correlated as inter-test values reported by either publisher in their respective examiner's manuals.

In addition, it is important to note that the TOLD-P:3 has recently been re-normed and subsequently replaced by the *Test of Language Development-Primary, 4th Edition* (TOLD-P:4: Newcomer & Hammill, 2008). Although the TOLD-P:4 maintains the same format and many of the same test items as the previous version, further investigation to independently establish the psychometric properties of the TOLD-P:4 Spoken Language Quotient would be warranted.

Given that published data from other large scale, independent studies of concurrent validity of norm-referenced language measures with school age children who have SLI are lacking, the application of Cohen's interpretive guideline would seem appropriate. With these standards in mind, the results from this study indicate that for school age children who meet the EpiSLI criterion, the overall standard scores of the TOLD-P:3 and the CASL perform as consistently as two independent measures of complex human abilities can reasonably be expected to perform. Although these two tests have different authors, reflect different theoretical views, and measure different aspects of language, they appear to be highly associated for clinical and research purposes. Yet, the correlation values that were observed in this study are smaller than the concurrent and predictive values that are frequently reported by publishers in examiner's manuals, and performance means for each instrument were significantly different statistically. These findings provide empirical evidence regarding diagnostic variability between norm-referenced measures. Both of the published language tests used in this investigation measured multiple aspects of oral language, but from differing theoretical frameworks. It appears that the diagnostic consistency between them may reflect those differences. Additional sources of variance between test scores might be attributable to time between test administrations and testing error.

This study investigated the degree of association among measures of complex abilities within a highly select population. All of the participants in this study demonstrated significant language impairments and met the EpiSLI criterion during identification testing. Finding correlation values between the TOLD-P:3 and the CASL as high as Cohen suggested to be reasonable within the behavioral sciences is a strong testament to the robustness of these assessment instruments for use with a clinical population of children with language impairments. Until such time as additional research into the concurrent validity of oral language assessment instruments for use with school age children who have language impairments can provide substantively different evidence, the findings of this study establish precedence that it is reasonable to expect that decontextualized, norm-referenced measures of multi-faceted oral language abilities should be moderately to highly correlated with one another as reflected in correlation coefficients of at least .30 to .50 when verified via independent investigation.

Construct Validity

Findings from this study document differing factor structures resulting from a sample of children with language impairment than were reported by authors as having been obtained from the original norming data of two decontextualized measures of language. Our analysis revealed two significant factors for the TOLD-P:3, while its publishers found only one factor by examining the norming sample data. As stated earlier, a two factor solution would have been an expected finding if the six core subtests loaded on the linguistic features (semantics and syntax) of the authors' two-dimensional model of the language structure that was used to generate the TOLD-P:3 subtests (p. 6). An alternative, but equally elegant, two factor loading pattern would have been present if the subtests separated into receptive and expressive tasks. Our data did not yield either of these loading patterns. Recall that, while most subtests loaded as expected according to the syntax/semantics dichotomy, the Picture Vocabulary subtest loaded more heavily with the syntax subtests than the other two semantic subtests.

As researchers with extensive clinical practice backgrounds, these factor findings are consistent with our clinical experiences when administering these subtests. Children who have language learning deficits have much more difficulty completing the subtests that loaded on the second factor (Oral Vocabulary and Relational Vocabulary) than the subtests that loaded on the first factor. Consequently, we interpret the factor loadings represented in the TOLD-P:3 component matrix, Table 4, as indicative of two different levels of language processing based on task requirements. The subtests that load on the first factor at levels above .4 (Grammatic Completion, Sentence Imitation, Grammatic Understanding, and Picture Vocabulary) required only circumscribed responses from children. Specifically, these tasks required children to point to pictures, fill in the last word of sentence, or imitate an utterance. However, the two subtests that loaded heavily on the second factor (Oral Vocabulary and Relational Vocabulary) required children to actively formulate utterances that transmit meaningful messages to the examiner in order to define words or describe similarities and differences. From this perspective, children with language impairment may be activating different elements of language ability during simplistic responses, or minimally communicative tasks, versus responses that require the organization and communication of more complex ideas through the construction and coordination of multiple utterances. Consequently, our factor analysis of the TOLD-P:3 performance data yielded a first factor that could be named "Basic Processing" and the second factor that could be named "Complex Processing."

A related, but alternative, way to name these factor loadings would be to use terms that reflect the extent of language information exchanged. Recall that TOLD-P:3 subtests that loaded on the first factor required language use primarily at the utterance level. For each subtest item, children received an utterance level prompt and responded nonverbally, or verbally with a single word or single utterance. The TOLD-P:3 subtests that loaded on the second factor required children to respond with a series of utterances that developed and conveyed coherent meanings across multiple utterances. From this view, the first factor could be named "Utterance Level Communication" while the second could be named "Discourse Level Communication." Although these terms may be slightly more familiar to practitioners than processing-based terms, essentially the distinction between the two factors remains one of language processing requirements and task load.

Regardless of how we name the resulting two factors, the current analysis of TOLD-P:3 performance data from a clinical sample reveals a two factor structure that does not fit the factor structure generated from the test's norming data and presented in the publisher's manual. Nor does it fit the conceptual model that was described by the test's authors. This new factor structure may have implications for our theoretical models as related to language impairment and its measurement. The empirical evidence from this study appears to support language constructs more aligned with language processing load than with the traditional theoretical constructs of receptive versus expressive language skills or linguistic subsystems including semantics, morphology, and syntax. The definition and verification of such theoretical language processing load constructs await future investigation and elaboration.

It is important to consider the possibility that the constitution of the differing samples may have influenced the resulting factor structures. The disparity between our findings and the TOLD-P:3 factor structure described by the publisher may reflect the ease with which children who are developing language within normal expectations, and who would comprise the majority of the norming sample, are capable of performing language-dependent tasks, which resulted in a single factor structure. While at the same time, children who have language learning difficulties may be more susceptible to processing overload than children whose language skills are typically developing (Edwards & Lahey, 1998; Hoffman & Gillam, 2004) or may have greater sensitivity to task demands (Gillam, Hoffman, Marler, & Wynn-Dancy, 2002), thereby yielding a two factor structure reflecting the differing task demands of the subtests. If this is the case, both factor structure models may be true for their respective samples, with the current study's two factor structure reflecting the inherent limitations of language impairment.

Alternate explanations for our two factor finding could include the possibility that the Picture Vocabulary subtest is measuring syntax rather than semantics, and therefore loads with the grammar subtests, or that children in our sample potentially document an emerging separation of vocabulary and grammar skills earlier than the trend toward that separation during early adolescence found by Tomblin and Zhang (2006).

With respect to the CASL, our finding of a single factor solution is divergent from the publisher's evidence for a three factor model of language ability. Our findings indicated that the performance data on the CASL core subtests for young school age children who have SLI do not separate into lexical, syntactic, and supralinguistic (i.e., form, content, and use) groupings. Instead, our CASL performance data with a clinical sample are consistent with Tomblin and Zhang's (2006) proposition of the unidimensionality of language data during the early elementary years.

The possibility exists that using only the core subtests of the CASL does not capture enough information about children's language abilities to support the differentiation of multiple facets of language ability in a sample of children who have known language learning deficits. Perhaps if the CASL supplemental subtests had been completed with the children in our study, different factor structure findings would have resulted. Alternatively, language may indeed be a three dimensional ability but testing constructs have not yet clearly delineated and captured those dimensions.

It is with interest that we note that our data reduction analyses of two published global language measures revealed not only differing factor structures than were reported by the tests' publishers, but also did not entirely support Tomblin and Zhang's (2006) assertion that language abilities, as documented via current assessment strategies, appear to be unidimensional during the early elementary years. Importantly, however, our findings are consistent with Tomblin and Zhang's (2006) in that the results of factor analyses for the TOLD-P:3 and the CASL did not yield support for the comprehension versus production dichotomy of language abilities as demonstrated by performance on published metrics. Clearly, additional research is needed to resolve these language construct and measurement issues.

Clinical implications

The findings of this study indicate that there is substantial overlap in the measurement of oral language abilities via two norm-referenced tests, TOLD-P:3 and CASL. However, each of these measures also accounts for considerably different aspects of oral language performance. This independent analysis of concurrent validity between two decontextualized norm-referenced tests of language ability within a sample of school age children with SLI confirmed that these measures are robustly associated while also substantially different from each other. Further, our study establishes an empirical baseline for the degree of association that can be expected for multifaceted measures of language abilities within a clinical sample. This empirical evidence extends our understanding of these measures beyond the information that is reported in publishers' examiner manuals using standardization data.

Our findings underscore the practical importance of selecting assessment measures on the basis of each test's theoretical framework (i.e., matching a test's theoretical constructs to areas of suspected disability), as well as selecting measurement formats that are clear and purposeful. While the TOLD-P:3 and the CASL met an acceptable level of inter-test consistency, they also measured considerably different subsets of language abilities. Although these tests yielded standard scores that were intended to document global language abilities, they did not demonstrate enough shared variance to allow them to be used interchangeably. Relying on test performance to inform clinical decisions, particularly with regard to eligibility for services, requires consideration of these important measurement limitations. In addition, it is important for practitioners to note that these tests may yield potentially conflicting information when completed with the same child. In all cases, interpreting test results must be based on the theoretical constructs of each assessment instrument. Differences in scores across tests may be more reflective of the construction of the assessment instruments, or task formats, than the inherent abilities of a child. The empirical evidence regarding how these tests operated in a clinical sample suggests that practitioners should cautiously select and interpret assessment measures on the basis of psychometric properties. Different tests may document each child's language abilities differently. Care should also be taken to avoid excessive reliance on the results of a single test or measure.

Cumulatively, this study's findings provide evidence to support the proposition that assessment of language abilities with school age children requires systematic collection of data from a variety of sources (Gillam & Hoffman, 2001). In addition, analysis of performance data from a large clinical sample yielded factor structures that did not fit the theoretical models posited by publishers, which may have implications for our conceptual models of language abilities and measurement. Future research is needed to further investigate the concurrent and construct validity of additional language assessment instruments, particularly those that attempt to measure multiple dimensions of language in school age children.

References

Allen DV, Bliss LS. Concurrent validity of two language screening tests. Journal of Communication Disorders. 1987; 20(4):305–317. [PubMed: 3624526]

ASHA. Highlights and trends: ASHA counts for year end 2009. American Speech-Language-Hearing Association; 2009.

Ballantyne AO, Spilkin AM, Trauner DA. 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 & Hearing Services in Schools. 2007; 38(3): 182–189. doi: 10.1044/0161-1461(2007/019).

Bankson, NW. Bankson Language Test. 2nd. ProEd. Inc.; Austin, TX: 1990.

Bruckner C, Yoder P, Stone W, Saylor M. Construct validity of the MCDI-I receptive vocabulary scale can be improved: Differential item functioning between toddlers with autism spectrum disorders and typically developing infants. Journal of Speech, Language and Hearing Research. 2007; 50(6): 1631–1638. doi: 10.1044/1092-4388(2007/110).

Carrow-Woolfolk, E. Test for Auditory Comprehension of Language -Revised. DLM Teaching Resources; Allen, TX: 1985.

Carrow-Woolfolk, E. Oral and Written Language Scales. Pearson; Minneapolis, MN: 1995.

Carrow-Woolfolk, E. Comprehensive Assessment of Spoken Language. American Guidance Service, Inc.; Circle Pines, MN: 1999.

Chiat S, Roy P. The Preschool Repetition Test: An evaluation of performance in typically developing and clinically referred children. Journal of Speech, Language and Hearing Research. 2007; 50(2): 429–443. doi: 10.1044/1092-4388(2007/030).

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Lawrence Erlbaum Associates; Hillsdale, NJ: 1988.

Dale PS. The validity of a parent report measure of vocabulary and syntax at 24 months. Journal of Speech, Language and Hearing Research. 1991; 34(3):565–571.

Edwards J, Lahey M. Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. Applied Psycholinguistics. 1998; 19(2):279–309.

Feldman HM, Dollaghan CA, Campbell TF, Kurs-Lasky M, Janosky JE, Paradise JL. Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. Child Development. 2000; 71(2):310–322. [PubMed: 10834466]

Friberg JC. Considerations for test selection: How do validity and reliability impact diagnostic decisions? Child Language Teaching and Therapy. 2010; 26(1):77–92. doi: 10.1177/0265659009349972.

Gillam, RB.; Hoffman, LM.; Rosello, D. Tests and measurements in speech-language pathology. Butterworth Heinemann; Newton, MA: 2001. Language assessment during childhood; p. 77-118.

Gillam RB, Hoffman LM, Marler JA, Wynn-Dancy ML. Sensitivity to increased task demands: Contribution from data-driven and conceptually driven information processing deficits. Topics in Language Disorders. 2002; 22(3):30–48.

Gillam RB, Loeb D, Hoffman LM, Bohman T, Champlin CA, Thibodeau L, Friel-Patti S. The efficacy of Fast ForWord language intervention in school-age children with language impairment: A randomized controlled trial. Journal of Speech Language and Hearing Research. 51(1):97–119. doi: 10.1044/1092-4388(2008/007).

Gray S, Plante E, Vance R, Hendrichsen M. The diagnostic accuracy of four vocabularly test administered to preschool-age children. Language, Speech, and Hearing Services in Schools. 1999; 30:196–206.

Greenslade KJ, Plante E, Vance R. The diagnostic accuracy and construct validity of the Structured Photographic Expressive Language Test--Preschool: Second edition. Language Speech & Hearing Services in Schools. 2009; 40(2):150–160. doi: 10.1044/0161-1461(2008/07-0049).

Hoffman LM, Gillam RB. Verbal and spatial information processing constraints in children with specific language impairment. Journal of Speech, Language, and Hearing Research. 2004; 47(1): 114–125. doi: 10.1044/1092-4388(2004/011).

Kaufman, AS.; Kaufman, NL. Kaufman Brief Intelligence Test. American Guidance Service; Circle Pines, MN: 1990.

Marchman VA, Martinez-Sussmann C. Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish. Journal of Speech, Language and Hearing Research. 2002; 45(5):983–997. doi: 10.1044/1092-4388(2002/080).

McCauley RJ, Swisher L. Use and misuse of norm-referenced test in clinical assessment: A hypothetical case. Journal of Speech and Hearing Disorders. 1984; 49(4):338–348. [PubMed: 6389982]

Newcomer, PL.; Hammill, DD. Test of Language Development-Primary. 3rd. Pro-Ed, Inc.; Austin, TX: 1997.

Newcomer, PL.; Hammill, DD. Test of Language Development-Primary. 4th. ProEd Inc.; Austin, TX: 2008.

Pankratz ME, Plante E, Vance R, Insalaco DM. The diagnostic and predictive validity of the Renfrew Bus Story. Language, Speech, and Hearing Services in Schools. 2007; 38(4):390–399. doi: 10.1044/0161-1461(2007/040).

Peña ED, Spaulding TJ, Plante E. The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. American Journal of Speech Language Pathology. 2006; 15(3):247–254. doi: 10.1044/1058-0360(2006/023). [PubMed: 16896174]

Perona K, Plante E, Vance R. Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition (SPELT-3). Language, Speech and Hearing Services in Schools. 2005; 36(2):103–115. doi: 10.1044/0161-1461(2005/010).

Plante E, Vance R. Diagnostic accuracy of two tests of preschool language. American Journal of Speech Language Pathology. 1995; 4(2):70–76.

Rescorla L, Ratner NB, Jusczyk P, Jusczyk AM. Concurrent validity of the language development survey: Associations with the MacArthur-Bates Communicative Development Inventories: Words and sentences. American Journal of Speech Language Pathology. 2005; 14(2):156–163. doi: 10.1044/1058-0360(2005/016). [PubMed: 15989390]

Restrepo MA, Schwanenflugel PJ, Blake J, Neuharth-Pritchett S, Cramer SE, Ruston HP. Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. Language, Speech and Hearing Services in Schools. 2006; 37(1):17–27. doi: 10.1044/0161-1461(2006/003).

Roman VP. The relationship between language ages of preschool children derived from a parent informant scale and language ages derived from tests administered directly to the preschool child. *Language*. Speech and Hearing Services in Schools. 1980; 11(1):50–55.

Spaulding TJ, Plante E, Farinella KA. Eligibility criteria for language impairment: Is the low end of normal always appropriate? Language Speech & Hearing Services in Schools. 2006; 37(1):61–72. doi: 10.1044/0161-1461(2006/007).

SPSS. Predictive Analytic SoftWare (Version 18). IBM; 2009.

Sturner RA, Heller JH, Funk SG, Layton TL. The Fluharty Preschool Speech and Language Screening Test: A population-based validation study using sample-independent decision rules. Journal of Speech and Hearing Research. 1993; 36(4):738–745. [PubMed: 8377486]

Thal DJ, O'Hanlon L, Clemmons M, Fralin L. Validity of a parent report measure of vocabulary and syntax for preschool children with language impairment. Journal of Speech, Language and Hearing Research. 1999; 42(2):482–496.

Tomblin, JB.; Norbury, CF.; Tomblin, JB.; Bishop, DVM. Understanding developmental language disorders: From theory to practice. Psychology Press; New York, NY US: 2008. Validating diagnostic standards for specific language impairment using adolescent outcomes; p. 93-114.

Tomblin JB. The EpiSLI database: A publicly available database on speech and language. Language Speech & Hearing Services in Schools. 2010; 41(1):108–117. doi: 10.1044/0161-1461(2009/08-0057).

Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. Journal of Speech Language and Hearing Research. 1997; 40(6):1245–1260.

Tomblin JB, Records NL, Zhang X. A system for the diagnosis of specific language impairment in kindergarten children. Journal of Speech and Hearing Research. 1996; 39(6):1284–1294. [PubMed: 8959613]

Tomblin JB, Shonrock CM, Hardy JC. The concurrent validity of the Minnesota Child Development Inventory as a measure of young children's language development. Journal of Speech and Hearing Disorders. 1989; 54(1):101–105. [PubMed: 2915520]

Tomblin JB, Zhang X. The dimensionality of language ability in school-age children. Journal of Speech, Language and Hearing Research. 2006; 49(6):1193–1208. doi: 10.1044/1092-4388(2006/086).

Wetherby AM, Allen L, Cleary J, Kublin K, Goldstein H. Validity and reliability of the Communication and Symbolic Behavior Scales Developmental Profile with very young children. Journal of Speech, Language and Hearing Research. 2002; 45(6):1202–1218. doi: 10.1044/1092-4388(2002/097).

Table 1

Descriptive statistics for the norm-referenced assessment measures.

	Kaufman Brief Intelligence Test (KBIT) Matrices Standard Score	Test of Language Development: Primary –Third Edition (TOLD-P:3) Spoken Language Quotient Standard Score	Comprehensive Assessment of Spoken Language (CASL) Core Composite Standard Score	Time Between Administration of TOLD-P:3 and CASL in Calendar Days
Mean	96.10	73.76	78.92	79.97
Standard Deviation	9.1	8.62	11.42	44.67
Range: Minimum-Maximum	77-124	46-96	46-115	13-265
N	216	216	216	216

Table 2

Description of TOLD-P:3 and CASL core subtests and listing of supplemental subtests.

TOLD-P:3 Core Subtests (all were included in this study)		
Subtest Name	**Task Description**	**Expected Response**
Picture Vocabulary	Child is shown 4 pictures on one page and verbally prompted to “Show me ________” (single object)	Closed response format. Nonverbal pointing to a single picture.
Relational Vocabulary	No picture prompts. The child is verbally prompted to describe how two items are alike.	Open response format. Utterance level verbalization summarizing the prominent similarities between the two items.
Oral Vocabulary	No picture prompts. The child is verbally prompted to describe a single vocabulary item.	Open response format. Utterance level verbalization summarizing the critical attributes of the item
Grammatic Understanding	Child is shown 3 pictures on one page and verbally prompted to “Show me _______” (sentences of increasing length and complexity.)	Closed response format. Nonverbal pointing to a single picture.
Sentence Imitation	No picture prompts. The child is verbally prompted to repeat the sentences that the examiner says.	Open response format. Utterance level repetition.
Grammatic Completion	No picture prompts. The child is verbally prompted to supply the missing last word in sentences that the examiner says.	Open response format. Single word sentence completion.
CASL Core Subtests (all were included in this study)		
Subtest Name	Task Description	Expected Response
Antonyms	No picture prompts. The child is verbally prompted to “Tell me a word that means the opposite of _____”	Open response format. Single word production.
Syntax Construction	Picture and verbal prompts.	Open response format. Partial and whole utterance formulation or imitation.
Paragraph Comprehension	Paragraphs of increasing length are read to the child, after which they are shown pictures and asked questions about the paragraph.	Closed response format. Nonverbal pointing to a single picture that would answer the question.
Nonliteral Language	No picture prompts. The examiner reads aloud a sentence that includes figurative language, an indirect request, or sarcasm. The child is verbally prompted to explain its meaning.	Open response format. Utterance level verbalization.
Pragmatic Judgment	Some picture prompts. The examiner reads aloud vignettes that describe everyday life situations. The child is verbally prompted to judge the appropriateness of the language used.	Open response format. Utterance level verbalization.
List of Supplemental Subtests		

TOLD-P:3 Core Subtests (all were included in this study)		
Subtest Name	**Task Description**	**Expected Response**
(not required for global scores and not included in this study)		
TOLD-P:3	Word Discrimination Phonemic Awareness Word Articulation	
CASL	Synonyms Sentence Completion Idiomatic Language Grammatical Morphemes Grammaticality Judgment Meaning from Context Inference Ambiguous Sentences	

Table 3

Eigenvalue and variance results of principle component factor analysis for TOLD-P:3 data.

Component	Eigenvalue	Percent of Variance	Percent of Cumulative Variance
1	1.845	30.76	30.76
2	1.154	19.24	49.997

Table 4

Component matrix of factor loading results from principle component factor analysis for TOLD-P:3 data.

TOLD-P:3 Subtest	Component	
	1	2
Grammatic Completion	.746	−.163
Sentence Imitation	.644	−.353
Grammatic Understanding	.601	−.115
Picture Vocabulary	.566	.023
Oral Vocabulary	.221	.785
Relational Vocabulary	.378	.611