Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission for extensive copying of my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

Psychometric Properties of Multi-domain Language Tests for School-age Children

by

Tabitha R. Syme

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Communication Sciences and Disorders

Idaho State University

Summer 2020

Copyright

Committee Approval

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Tabitha R. Syme find it satisfactory and recommend that it be accepted.

_____

Diane A. Ogiela, PhD, CCC-SLP,

Major Advisor

_____

Kristina M. Blaiser, PhD, CCC-SLP,

Committee Member

_____

Peter J. Flipsen Jr., Ph.D., S-LP(C), CCC-SLP,

Committee Member

_____

Ryan Lindsay, PhD,

Graduate Faculty Representative

Table of Contents

List of Tables

Psychometric Properties of Multi-domain Language Tests for School-age Children

Thesis Abstract--Idaho State University (2020)

Standardized norm-referenced tests can contribute to or detract from accurate diagnoses and treatment of language impairments. This review evaluated the presence and quality of the psychometric characteristics of norm-referenced language tests used for assessment of school-age children with potential language impairment. The goals were to provide a reference resource that clinicians can use to assist in decision making and to identify areas of psychometric quality that are in need of further development and improvement. Two reviewers systematically and independently reviewed 13 omnibus language tests for reliability, validity, normative sample characteristics, and fairness according to a subset of *Standards for Educational and Psychological Testing* (AERA et al., 2014). Results indicated overall improvement in the psychometric quality of omnibus language tests when compared with previous reviews. Ongoing concerns are discussed for each aspect of psychometric quality, especially with regards to the normative sample, validity studies, and diagnostic accuracy.

Key Words: school-age language, language test, language assessment, psychometric properties, review

Introduction

The language assessment process is one by which speech-language pathologists (SLPs) diagnose their clients, and it provides the basis for developing a treatment plan. Officially, there is no "gold standard" in language assessment. However, an assessment needs to provide enough detailed information to do that (Taylor-Goh, 2005). Many SLPs agree that such life-impacting clinical decisions should be made based on a compilation of data from several measures (Spaulding et al., 2012). This is consistent with the Individuals with Disabilities in Education Act (IDEA) which prohibits using a single measure or assessment as the only criterion for determining whether or not a child has a disability and for determining an appropriate educational program (U.S. Department of Education, 2006). The SLP scope of practice document states that the assessment process should include "… culturally and linguistically appropriate behavioral observation and standardized and/or criterion-referenced tools; use of instrumentation; review of records, case history, and prior test results; and interview of the individual and/or family to guide decision making" (American Speech-Language-Hearing Association [ASHA], 2016, p. 11). Multiple sources in the literature report collectively accepted components specific to a comprehensive language assessment for school-age children. These components are a case history, client interviews, other related interviews, standardized norm-referenced tests, non-standardized assessments such as curriculum-based assessments, dynamic assessments, and language samples (Ireland & Conrad, 2016; Paul et al. 2018; Spaulding et al. 2012). When norm-referenced tests are part of the assessment process, those tests must be of high quality, as determined by strong psychometric characteristics.

Standardized norm-referenced language tests play a crucial role in most comprehensive language assessments for children who may have a language impairment, in any setting. In

schools across the country, these tests are a major factor in determining whether or not a child is eligible for services from an SLP. A norm-referenced language test is one designed to evaluate language skills on specific tasks for comparison to same-age peers. These tests can provide a variety of scores such as standard scores, percentile ranks, stanines, etc. Given the importance of norm-referenced language tests as a measure in the diagnosis of children with language impairment and for determining eligibility for services, the corresponding psychometric quality of the tests used in these processes cannot be overemphasized. A test with good psychometric quality consistently measures what it claims to measure across time, between individuals and in different settings.

Background

The psychometric properties of norm-referenced tests need to be evaluated in the profession of speech-language pathology in order to ensure that high quality, relevant, and most appropriate tests are used for a given individual and situation. All tests are not of equal quality, nor are they all designed to test the same aspects of language in the same ways. Disparity in the quality of language tests has been documented in reviews over the last several decades beginning in the early 1980s. McCauley and Swisher (1984) reviewed thirty language and articulation tests to determine how well each test met certain criteria for ten psychometric properties, including a description of tester qualifications and test procedures, demographics and size of the normative sample, item analysis, raw score means and standard deviations, concurrent validity, predictive validity, test-retest reliability and inter-examiner reliability. Overall, they found that, at that time, a majority of tests either did not meet most of the psychometric criteria, or failed to provide evidence that they met the criteria (McCauley & Swisher, 1984). That psychometric review provided much-needed recognition of the need for improvement in tests and served as an impetus

for improving the quality of speech and language test development as it pertained to these ten psychometric properties. To measure improvement over the intervening decade, Plante and Vance (1994) conducted a study examining twenty-one norm-referenced language tests for preschool-age children using the same criteria. They found eight tests met more than half the established criteria. The improvement in the percentage, 38% of tests meeting psychometric criteria since the 20% reported in the review of McCauley and Swisher (1984), was not as great as anticipated. Most recently, the psychometric properties of 15 language assessments for school-age children were systematically reviewed by Denman et al. (2017). These researchers used the Consensus-based Standards for the Selection of Health Status Measurement Instruments (COSMIN) taxonomy and checklist, which was developed within a medical model for the development and evaluation of health outcome measurements. Again, while results indicate overall improvement, they found the methodological quality of the studies which provided the evidence for psychometric properties was, for most assessments, lacking in some combination of the specific analyses, procedures, or sample size according to COSMIN requirements. Based on the results of their review, the authors identified and recommended only four tests as having good supporting evidence for use (Denman et al., 2017).

Because one of the major purposes of many language tests is to discriminate between children who have typical language skills and those who do not, diagnostic accuracy must be carefully considered. Therefore, when the purpose of administering a standardized, norm-referenced test, is to assist diagnostic decisions, that test must have good diagnostic accuracy. Diagnostic accuracy is a term describing a test's ability to distinguish between individuals from different categories (Dollaghan, 2004). In the case of language tests that would be between individuals with language impairment and those without. Sensitivity and specificity are common

measures of diagnostic accuracy. Sensitivity is a measure of how well the test identifies an individual with an impairment as actually having that impairment. Specificity is a measure of how well the test identifies an individual without the impairment as not having the impairment. When sensitivity is low, the test will have a high rate of false negatives. When specificity is low, the test will have a high rate of false positives. Poor sensitivity or poor specificity results in children not receiving needed services or children receiving unnecessary services respectively. A test with good diagnostic accuracy has both good sensitivity and specificity. In their study, Plante and Vance (1994) examined this issue. They administered four tests to 20 preschool children known to have a language impairment and an age-matched control group. In order to assess the strength of each test in discriminating between typically developing children and those with Specific Language Impairment (SLI) as having a language impairment, they examined each test's sensitivity and specificity. They measured diagnostic accuracy as a percent accuracy with which test scores discriminated between children with language impairment and children with typical language development (sensitivity, specificity, and apparent error rates). As a benchmark they suggested that 90% be considered good discriminant accuracy, between 80% and 89% be considered fair and anything below 80% as poor because that would mean 20% of children would be misidentified. Results indicated that even though their review of tests indicated that more tests had met criteria for psychometric properties, only one language test adequately discriminated between those children with language impairment from those without (Plante & Vance, 1994).

The importance of diagnostic accuracy is such that Denman et al. (2017) included it in their review, even though it did not fit neatly into a COSMIN measurement property. It was also considered important enough that in their review of the psychometric quality of tests, Friberg

(2010) used diagnostic accuracy as an inclusion criteria. As a result of requiring a minimum of .80 sensitivity and specificity only nine tests were included. The rest of the criteria were based on the work of McCauley and Swisher (1984) along with two additional criteria. These were clear identification of the test purpose and that a clearly defined standardization sample include age, gender, ethnic background, and parental education or socio-economic status (SES). Such information was critical to determine the relevance of the normative sample when applying the test to an individual child. Most criteria were not judged for quality, only presence, such as a clearly defined normative sample. Certain criteria were considered present only at a specific minimum level, such as normative sample size of 100 or more per subgroup. Friberg (2010) found that all of the included tests met eight or more of the 11 criteria, while five tests met 10 of the criteria and none met all 11 criteria.

While the selection of an appropriate language test is critical to the accurate evaluation and effective treatment of language disorders, SLPs do not routinely choose norm-referenced tests based on psychometric properties. Betz et al. (2013) surveyed 364 practicing SLPs to examine how often certain standardized tests were being used when diagnosing children suspected of having SLI. They compared survey results with the psychometric properties of 55 tests looking for correlations. Findings indicated that SLPs regularly use only a small portion of the available standardized language tests when diagnosing children with SLI, and those tests were primarily omnibus language tests or vocabulary tests. Betz et al. did not find significant correlations between the frequency of use and the test's psychometric properties. Similarly, a separate survey of SLPs found that availability, personal familiarity, and diagnostic accuracy weighed more heavily in the test selection process than the test's psychometric features (Montzka, 2015).

Purpose

The unofficial "gold standard" in language assessment includes several components that are considered necessary to compile a comprehensive language profile, one of which is almost always a norm-referenced test (ASHA, 2016; Ireland & Conrad, 2016; Paul et al. 2018; Spaulding et al. 2012) at least for children capable of participating in such tests. Currently, when choosing norm-referenced tests, many SLPs are relying on what is currently on hand, what is familiar, word of mouth recommendations, or publisher advertising, which may or may not be supported by the evidence (Montzka, 2015). Somewhat accidentally, the tests chosen may also have good psychometric quality.

When psychometrically strong and appropriate tests are used, they can contribute to accurate diagnosis and appropriate treatment for children with language impairment. When a test is psychometrically weak or inappropriate it can lead to misdiagnosis. For a child with typical development this could lead to inappropriate or unnecessary treatment. For a child with an impairment, this could lead to inappropriate treatment or a lack of treatment. Such misdiagnoses can result in financial, ethical, psychological, educational, social, and emotional ramifications for any and all individuals involved.

There are three ways a test may be inappropriate; a test can be technically inadequate, a test can be technically adequate but used for the wrong purpose, or a technically adequate test can be used for the wrong child, such as a child who is not represented in the normative sample or whose specific problems may lie outside the scope of the test (Salvia & Ysseldyke, 1981). Consequently, the decision of which tests to utilize in the assessment process must include an evaluation of psychometric properties such as reliability/precision, validity, fairness, norms, and diagnostic accuracy. When these properties are commensurate with the purpose for which the

test is used, the results can be incorporated meaningfully into the comprehensive communication profile of a child with the potential for having language impairment.

The primary purpose of the current review was to create an accessible and evidence-based resource to assist practicing SLPs in choosing which multi-domain norm-referenced language tests, based on their psychometric properties, will best suit their specific purpose and the needs of individual children on their caseloads. However, it is not intended to replace an SLP's independent review of tests for their purposes and application to their client population. This review evaluated tests with regard to the normative sample, validity, reliability, fairness, and diagnostic accuracy for use as diagnostic tools in the assessment of school-age children with potential language impairment. The review criteria were largely based on the revised *Standards for Educational and Psychological Testing* by American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education NCME, which will hereafter be referred to as the *Standards* (AERA et al., 2014). The secondary purpose of the review was to identify areas of psychometric quality specifically in omnibus language tests for school-age children that need further attention. This review will contribute to the body of research regarding such tests.

Aspects of Psychometric Quality

Multiple factors will impact language test selection in various clinical and educational settings, including the availability of newly published tests and recently published editions. In order to continue to provide clinicians with evidence to inform their decisions, it will be beneficial to review and evaluate various dimensions of language tests based on the Standards (AERA et al., 2014). These standards define each psychometric property and provide detailed explanations of standards to inform application.

The present study focused on the foundational aspects of validity, reliability/precision, fairness, and norms and diagnostic accuracy. Within these four areas, there are 25 standards for validity, 20 standards for reliability, 20 standards for fairness, and four specifically for norms. Within the scope of the present study, it was not feasible to examine all 69 standards for these categories for each included test, nor would it have been beneficial as many are conditional to a specified situation.

The Standards are organized in thematic clusters. They outline the **types of evidence** that need to be addressed in tests (AERA et al., 2014). This review addressed certain clusters by examining those standards most representative of the psychometric properties that have been examined in previous research.

In addition, diagnostic accuracy will be addressed because it is an important characteristic that needs to be considered in test selection. This is particularly true for tests that claim to identify children as language-impaired or not. The Standards suggest that tests be evaluated based on the stated purpose. Diagnostic accuracy is recognized as a critical aspect of norm-referenced tests when used for the purpose of assisting diagnosis of language impairment (Denman et al., 2017; Dollaghan, 2004; Friberg, 2010; Plante & Vance, 1994; Spaulding et al., 2006; Tomblin et al., 1996).

This review presents aspects of psychometric quality beginning with the characteristics of the normative sample. A strong understanding of normative data and how it influences every other psychometric property is needed in order to properly evaluate tests. Validity studies often utilize data collected from the normative sample during the standardization process and therefore is presented second. Because diagnostic accuracy is primarily presented as an aspect of test validity, and is calculated using normative data, it will be the third psychometric property

presented. Test reliability and precision are dependent upon test validity. Studies of reliability and precision often utilize norming data and is presented fourth. The final property will be fairness, which deals primarily with test administration and appropriateness for populations of interest.

The Normative Sample

Norm-referenced language tests provide a snapshot of a child's performance relative to peers. The Standards state that "Norms, if used, should refer to clearly described populations. These populations should include individuals or groups with whom test users will ordinarily wish to compare their own examinees" (AERA et al., 2014, p. 104). The normative sample is a cornerstone in the foundation upon which a test is built. It is important because the information from this sample is often used to determine final test items and scoring procedures. Therefore, all measures based on those items and procedures such as test-retest reliability, inter-rater reliability, standard scores, standard errors of measurement, and often diagnostic accuracy depend upon the quality of the normative sample. The test manual must sufficiently describe the demographics of the normative sample to allow for appropriate peer comparisons. The sample must be of adequate size to approximate a normal distribution.

Demographics

Demographic information for normative samples should include age, gender, ethnic representation, parent education/socioeconomic status (SES), disorder/diagnosis exceptionality status, and geographic location (Entwisle and Astone, 1994; Friberg, 2010; McCauley & Swisher, 1984; Peña et al., 2006; Spaulding et al., 2006). This information is crucial in determining the comparability of the performance of the child in question to the normative data

supplied by test developers. In order for that comparison to be a valid interpretation of the child's standing, that child's personal demographics must be represented in the normative sample. The literature indicates a fairly universal acceptance of this principle (Flipsen & Ogiela, 2015; Friberg, 2010; Denman et al., 2017; Norbury & Gosse, 2018; McCauley & Swisher, 1984; e.g.). Salvia and Ysseldyke (1981) recommend local norms when available because they are most representative of a specific geographic area. Yet, local norms are rare. When local norms are not used, sampling distributions should be similar to the national population distribution. National norms are more relevant for the evaluation of development (Salvia & Ysseldyke, 1981).

Adequate Sample Size

In language tests, the normative sample is divided into scoring subgroups or cells. The age range within each cell may vary across a test in order to best capture the nature of language development. For example, because of the rapid development of language in the early years, tests often divide the total normative sample into age-based subgroups as small as three-month cells for the younger ages, through one-year cells in middle childhood, and multiple-year cells for later adolescence. An individual's score is not always compared to all participants at each age, by year, but rather to a scoring subgroup or cell. Thus, the size of that cell (i.e., how many children are included) is an important aspect of assessing whether or not the norms are adequate. For example, when comparing the standard score of a 4-year, 2-month-old child to the normative data in the scoring appendices, the table may provide standard scores for children ages 4-years to 4-years and 5-months, and 4-years and 6-month to 4-years and 11-months, thus dividing a reported normative sample size of 115 4-year-olds (assuming equal division of participants) into approximately only 57 children per cell between 4-years to 4-years and 5-months.

While the need for adequate sample sizes for norms is universally upheld, researchers have disputed the minimum number that constitutes an adequate sample size. These minimums have been argued for as high as 400 (Charter, 1999) and as few as 50 (Bridges & Holler, 2007; Crawford & Howell, 1998; Mitrushina, 2005). Still, others suggest that 100 is an acceptable minimum (Salvia & Ysseldyke, 1981; Weiner & Hoock, 1973). In many previous reviews of assessments related to speech and language, the minimum sample size considered acceptable per scoring cell has been 100 (Flipsen & Ogiela, 2015; McCauley & Swisher, 1984; Plante & Vance, 1994). Many tests present normative data stratified by age, reporting sample sizes of approximately 100 or more per age. However, they may then have scoring cells of fewer than 100 participants in certain age ranges. Very small sample size in a scoring cell tends to result in positively skewed distribution's that yield inaccurate standard scores (Mitrushina, 2005). Given the rapid development of language during the younger years, providing norms based on smaller age brackets is logical, yet is problematic if dividing a one-year age group into 3 to 6-month cells results in actual sample sizes much lower than the standard of N=100 or more.

There are specific justifications given for specific cell size minimums. The prescription by Charter (1999) that 400 participants were the minimum sample size needed to ensure high reliability coefficients of r = .90 or greater is based on very precise statistical significance. Such numbers are also practically impossible to acquire for normative studies of this nature (language assessment). According to Weiner and Hoock (1973) a sample size consisting of 100 to 200 participants allows test developers obtain adequate reliability and an adequate range of scores (including extremes) in order to allow for accurate interpretation of testing results. In a study of how sample size affects confidence intervals for validity and reliability, Mendoza et al. (2000) found that they were able to calculate estimates of reliability that were accurate and constructed

confidence intervals that were reasonable when using a sample of at least 100. Similarly, a full range of percentiles and standard scores cannot be computed without extrapolation for sample sizes consisting of less than 100 (Salvia & Ysseldyke, 1981). Salvia and Ysseldyke (1981) add that the sample size should be of adequate size that it will include "infrequent elements" and will ensure a relatively small size of interpolations and extrapolations (p. 123). However, research by Bridges and Holler (2007) on the effect of sample size on the stability of the estimate of the normative sample means, provides evidence supporting a smaller acceptable minimum. Utilizing the normative sample data for four frequently used pediatric neuropsychological instruments, the authors determined the confidence intervals for each scoring cell. Using that same normative sample data, Bridges and Holler recalculated confidence intervals using various cell sample sizes ranging from 5 to 500. The resulting confidence intervals were presented as a function of sample size and indicated that a sample size of 50 to 75 narrowed the width of the confidence intervals to approximately one-half of a standard deviation when compared to samples smaller than 50. Increasing the sample size beyond 75 narrowed confidence intervals only slightly. However, any sample size of less than 50 yielded unacceptably wide confidence intervals (Bridges & Holler, 2007).

A similar conclusion can be drawn from Crawford and Howell (1998), who demonstrated the need to use the t-test on normative data when n=<50, implying that an n=50 or greater was sufficient for z-score calculations, which is what is typically seen as a standard score for a normal distribution. Finally, in the *Handbook of normative data for neuropsychological assessment (2nd ed.)*, Mitrushina (2005) stated that because the potential positive skew of the sampling distribution of small samples results in inaccurate standard scores, appropriate sample size is necessary and concurs that n=50 is generally adequate. To summarize, while it seems safe

to conclude that less than 50 participants in each cell of a normative sample are inadequate, especially when a test is used for the purposes of diagnoses, it would seem that the commonly used standard of 100, while preferable, may not be necessary.

Full-range Samples Compared to Truncated Samples

Clinicians need to know what type of sample was used for norming a test and how it is relevant to their particular purpose. There are positive and negative attributes of both full-range and truncated normative samples. Full-range samples are, essentially, a sample selected for the norming procedures that are representative of all possible skill levels in the national population (McFadden, 1996; Peña et al., 2006; Salvia & Ysseldyke, 1981). Salvia and Ysseldyke (1981) suggest that in order to have a point of reference regarding test performance for a particular type of problem, children with that particular problem must be included in the normative sample. Peña et al. (2006) found that the inclusion of children with language impairments resulted in poorer diagnostic accuracy, under-identification, than tests normed on children who were typically developing only (truncated samples). However, this inclusion increased a test's utility in determining the severity of impairment when the presence of that impairment had already been established (McFadden, 1996; Peña et al., 2006). Full-range samples yield better severity ratings but poorer diagnostic accuracy with potentially high rates of false negatives than truncated samples.

Truncated samples exclude individuals diagnosed with impairments or other conditions that are likely to negatively impact performance, effectively cutting off or *truncating* the bottom of the bell curve (McFadden, 1996). While Peña et al. (2006) present a strong case for using a truncated sample when norming a test intended for diagnosis, others argue against it on the basis of over identification of impairment. McFadden (1996) explained that cutting off the bottom of a

normal distribution meant the remaining data had to be statistically forced into a new bell curve, no longer representing a full range of skill. Thus, a child must score lower than the entire "normal" sample in order to demonstrate impairment. If the typical cutoff scores used to denote impairment (-1.5, -1.75, and -2.0 *SD*) are applied, it is likely a typically developing child would be identified as impaired (false positive; McFadden, 1996). With successive re-norming of truncated norms, there is the potential for further movement of the sample distribution away from the lower end of normal. This could result in progressively more false positives, creating the impression of language impairment where none exists because children who do score in the low end of a truncated sample and are then diagnosed as impaired would be excluded from subsequent norms (McFadden, 1996). It can, therefore, be concluded that truncated norms yield better diagnostic accuracy, yet potentially higher rates of false positives than full-range samples.

Test Validity

The question of validity for a test is really a question of whether or not a test actually measures what the authors claim it measures. The Standards, rather than requiring specific types of validity, state that validity evidence must be presented to support the test purpose, defining validity as "the degree to which *accumulated evidence and theory* support a specific interpretation of test scores for a given use of a test" (AERA et al., 2014, p. 225). This means the intended purpose of the test must be clearly stated in order to assess whether or not the evidence provided supports the test's validity. Even when a test's purpose is clearly stated, the evidence provided may not actually support its use for that purpose (Friberg, 2010).

Another element necessary for validity is a clear description of the population(s) for whom a test is intended to be used and for whom it is not intended. For language tests this description should specify elements such as age range, languages, dialects, suspected

impairments, diagnoses, or second language learners. A clinician should be able to determine if the child being evaluated for language impairment fits the description of the intended population.

Test developers conduct validity studies to establish the validity of a test for its purpose with its intended population(s). Certain types of validity studies have been consistently conducted and the results reported as evidence of test validity. Historically, test validity has been established using specific types of validity evidence, namely construct (structural) validity, content validity, concurrent validity, and predictive validity (Denman et al., 2017; Flipsen & Ogiela, 2015; Hoffman et al., 2011; McCauley & Swisher 1984; Plante & Vance, 1994). Such measures allow for continuity in reviews and accuracy in longitudinal comparisons of psychometric quality. The current review examined evidence of validity based on the stated purpose of the test, intended population(s), content validity, concurrent validity, predictive validity, and construct/structural validity.

Even though the Standards do not provide specific bench marks for determining validity, evidence presented for validity should not be accepted as good, merely because it is present. Instead it should instead be examined for quality. The quality of validity studies, like any experimental study, need to be considered according to levels of evidence. Such determining factors include the sampling method (randomized/control), sample size (larger = better), procedures (blinding), and examiner biases (Dollaghan, 2004).

Content Validity

"Content-related validity evidence is evidence based on test content that supports the intended interpretation of test scores *for a given purpose"* (AERA et al. 2014, p. 218). It should describe the degree to which a test measures the behavior it is intended to measure. It should be considered in light of the appropriateness of item types, how completely the items sample the

entire range of skill being tested, and the manner in which items assess the stated content (Salvia & Ysseldyke, 1981).

For the purposes of the current review, content validity is reported as the degree to which the test measures the constructs of language it claims to measure. These constructs are described by test developers as various aspects of language and are based upon the theoretical framework of language to which they subscribe. Statistical procedures, such as differential item analysis, should be part of the development process and guide inclusion or exclusion of items. It is key that the constructs measured link to the test purpose, as stated by test developers. This form of validity is highly relevant to test selection criteria for the assessment of language. Whether the intended purpose is to identify a language impairment, determine patterns of strength and weakness, measures of progress over time, or as a research measure, the test must adequately measure the intended aspect of language in order to be valid for that use.

Concurrent Validity

In order to provide evidence of validity, a test may be compared to either clinical judgment or other tests that are *assumed to be valid* (Salvia & Ysseldyke, 1981). In language tests, this form of validity evidence is often presented as concurrent validity. While the Standards do not require such specific types of evidence, they do describe forms of evidence that support test validity. The Standards consider convergent evidence to be "… based on the relationship between test scores and other measures of the same or related construct" (AERA et al., 2014, p. 218). As there is no single undisputed objective measure of language skill, language tests are often compared to other tests that purport to measure the same or similar constructs. The degree to which both tests measure a construct, as well as the degree to which they diverge, must be considered in order to establish meaningful correlations.

Prior studies have examined concurrent validity according to the degree of correlation between the scores of the test in question and the scores of other valid measures of similar language constructs (Denman et al., 2017; Flipsen & Ogiela, 2015; Hoffman et al., 2011; McCauley & Swisher 1984; Plante & Vance, 1994). Because this type of validity is often reported for language tests, it will be considered evidence of validity. However, concurrent validity must be carefully considered. The test chosen as the basis for comparison may erroneously be assumed to be valid, unless there is evidence that it is indeed valid. Concurrent validity is also sensitive to sample size, with large sampling errors resulting from a small sample size (Boateng et al., 2018).

Predictive Validity

Predictive validity is "… evidence indicating how accurately test data collected at one time can predict criterion scores that are obtained *at a later time*" (AERA et al., 2014, p. 222). Predictive validity evidence can be derived from prediction correlations between test results and results on another similar measure (very similar to concurrent validity). It can also be considered a prediction of future performance in the same area, in related areas such as literacy skills, or academic achievement. Requiring studies of such prediction would mean test developers must wait years before publishing their test, if they are to provide such evidence. Therefore, even though McCauley and Swisher (1984) considered predictive validity relevant to planning treatment and intervention, it is understandable why it is still not commonly reported in test manuals (Friberg, 2010). This may also be due in part to the ambiguity of its definition. The term predictive validity is used to mean different things. When it is reported by test developers, this review classified it based on the meaning they used. If it was used with the meaning to predict performance on another test, this review considered it as concurrent validity, and where it was

used with the meaning predict whether an individual did or did not have a language impairment, this review considered it evidence of diagnostic accuracy.

Structural Validity

Structural validity is evidence of a test's internal structure, which is "In test analysis, the factorial structure of item responses or subscales of a test" (AERA et al., 2014, p. 220). It is often determined by factor analysis and depends on the theoretical foundation of language upon which the test was developed. It should support the rationale of composite scores, difference scores, or profiles when provided. A factor analysis with items loading onto a single factor may support the use of a single composite score for language. A two-factor model would potentially support separate receptive and expressive language composites. Caution should be used in evaluating the quality of studies used in factor analysis, just as in all aspects of psychometrics. Data reported in the manual will not always generalize to other populations or settings. For example, in an independent validity study, Hoffman et al. (2011) compared the *Test of Language Development – Primary: Third Edition* (TOLD–P:3; Newcomer & Hammill, 1997) and the *Comprehensive Assessment of Spoken Language* (CASL; Carrow-Woolfolk, 1999). Their study indicated different factor structures for these tests than the test developers found and reported.

Diagnostic Accuracy

The prevalent use of standardized tests in language assessment and determination of treatment eligibility indicate an assumption that scores on these tests differentiate between children with and children without language impairment. However, this is typically based on an arbitrary cut-off score at a standard deviation level that has not been verified for individual tests (Spaulding, et al., 2006). If test results are used for the purpose of assisting in diagnosis, as is often stated in test manuals, then their accuracy in doing so must be evaluated. Diagnostic

accuracy is dependent upon multiple factors. One such factor is the variability in what test developers and other researchers use as the reference standard for impairment when conducting studies to determine diagnostic accuracy. Some may use multiple tests as a basis of comparison, others may use current participation in therapy, and still others may use whether or not another test indicated language impairment. The latter is troublesome because it is not distinguishable from concurrent validity.

Another factor that is of primary concern to this review is the cutoff score at which a test is most diagnostically accurate for identifying those with and those without a disorder. The concept here is that a single particular score or specific standard deviation from the mean can best separate impaired from typically developing children (Ivnik et al., 2000; Tomblin et al., 1996). Cut off scores are often stated in policy regarding eligibility criteria for school services, usually, as a standard score falling a specified number of standard deviations below the mean on a standardized norm-referenced test (typically -1.5, -1.75, or even -2 SDs). This cutoff is often applied regardless of which test is administered and the characteristics of that test. Thus, this use of standard deviation cutoffs, is arbitrary. It does not account for variation between tests with regards to diagnostic accuracy. Not all tests have acceptable levels of diagnostic accuracy at the same cutoff score (Spaulding et al., 2006). In fact, evidence shows that relying on an arbitrary cutoff score to determine language impairment results in children with language impairment who remain undiagnosed and go without needed services while children with typical language development are diagnosed with language impairment and receive unnecessary treatment, wasting time and money for all involved (Dollaghan, 2007; Spaulding et al., 2006). Despite this evidence, the policy of qualifying a child for services based on the arbitrary cutoff has remained

either a requirement or strong recommendation in many states. Therefore, this review examines diagnostic accuracy in terms of the cutoff score.

Given the importance of diagnostic accuracy, it is equally important that clinicians understand the types of diagnostic accuracy evidence presented in currently available test manuals and how they relate to their specific client and clinical context. As tests more frequently report evidence of diagnostic accuracy, the most commonly reported measures are sensitivity and specificity. Others, less frequently provided, include predictive values, likelihood ratios, receiver operating characteristic (ROC) curves resulting in the area under the curve (AUC), and group differences. What follows is a brief explanation of each type of validity evidence, as seen in this review, and what is generally considered an acceptable value for each of them.

Sensitivity and Specificity

Sensitivity is the percentage of individuals classified as having an impairment who did, in fact, have the impairment, according to previously accepted diagnostic criteria, out of the whole sample. Specificity is the percentage of individuals classified as unimpaired who were indeed unimpaired out of the whole sample. For example, a test with .9 sensitivity correctly identified 90% of the individuals in the sample who had met some diagnostic criteria for language impairment. If the test had a specificity of .88, it correctly identified 88% of the sample participants as having typically developing language. Like all psychometric properties, sensitivity and specificity should be viewed according to the context from which the numbers were derived. Sensitivity and specificity are thought to be stable regardless of prevalence (Šimundić, 2009); however, several studies, when examined through meta-analysis, have found otherwise (Leeflang et al., 2013). The range of skill present in the sample and the range of severity present in the sample has the greatest effect on sensitivity and specificity (Leeflang et

al., 2013). Individuals with milder severity are more likely to go undiagnosed, while potentially typically developing children who happen to fall on the lower end of normal could be diagnosed as disordered. With the previous example, even though the majority of children would be correctly identified, 10% of children with a milder impairment would go unidentified, while 12% of typically developing children would be classified as disordered.

The utility of sensitivity and specificity depends on both what percentage of accurate classification is considered acceptable and at what cutoff score those percentages are found. Previous research related to the sensitivity and specificity of speech and language assessments have maintained the high standard as 90% for both, with 80% deemed acceptable (Plante & Vance, 1994; Merrell & Plante, 1997; Gray et al., 1999; Gray, 2003; Perona et al., 2005; Spaulding et al., 2006; Greenslade et al., 2009; Dispaldro, Leonard, & Deevy, 2013; Pearson et al., 2014; Denman et al., 2017). The cutoff score that achieves the best diagnostic accuracy varies widely from one test to another, even between tests from the same developers. For example, Greenslade et al. (2009) found that the Structured Photographic Expressive Language Test – Preschool: Second Edition (SPELT–P2; Dawson et al., 2005) had the best diagnostic accuracy (90.6% sensitivity and 100% specificity) at a standard score of 87. In contrast, the Structured Photographic Expressive Language Test: Third Edition (SPELT–3; Dawson et al., 2003) had the best diagnostic accuracy (90% sensitivity and 100% specificity at a standard score of 95 (Perona et al., 2005). There was a notable difference between the cutoff scores at which that level of accuracy was attained. Regardless, both of these tests had the best sensitivity and specificity at scores that are less than one standard deviation below the mean. This is problematic as most school eligibility criteria levels are considerably lower than one standard deviation below the mean.

Predictive Values

Unlike sensitivity and specificity, which is a percentage of accurately classified individuals in a sample, predictive values (PV), as evidence of diagnostic accuracy, are a measure of statistical probability of having or not having an impairment (Mitrushina, 2005). A positive predictive value (PPV) is the probability that an individual who actually has a disorder will be classified by the test results according to a specified cut off score as having a disorder (Mitrushina, 2005). Likewise, a negative predictive value (NPV) is the probability that an individual who does not have a disorder will be classified as such according to a specified cut off score (Mitrushina, 2005). PVs, especially PPV, are highly influenced by the prevalence of the disorder in the study sample (Mitrushina, 2005; Šimundić, 2009). The greater the prevalence of the disorder in the sample, the greater the probability that the test would accurately classify someone with the disorder. Test manuals should present predictive values along with the estimated prevalence of the disorder. Often this is presented as a *base rate*, ranging from 10-20% for a screening of the general population and 50-90% for a referral population (Denman et al. 2017). Thus, a clinician needs to have an understanding of which base rate (screening, referral or other) is most applicable to the assessment situation at hand in order to understand the diagnostic accuracy of a particular test with regards to the child or children they are assessing.

Predictive values are generally presented as percentages and PVs in the 70s are considered to be poor, noting the greater the reported percent accurate, the more precise the classification (Denman et al., 2017; Gray et al., 1999). Because predictive values are a measure of the probability of an accurate classification, it can be concluded that percentages comparable to sensitivity and specificity criteria, the high standard (90%) and acceptable (80%) should be expected.

Likelihood Ratio

A positive likelihood ratio (LR+) represent the confidence a clinician can have that the score on a test (one that indicates impairment) actually came from an individual with the impairment and not from and individual without the impairment. A negative likelihood ratio (LR-) represent the confidence that the score on a test (one does not indicate impairment) actually came from an individual without the impairment and not from an individual with the impairment (Dollaghan, 2007). Likelihood ratios are useful for ruling-in a diagnosis or ruling-out a diagnosis, respectively (Šimundić, 2009). Likelihood ratios are calculated using sensitivity and specificity data (Dollaghan, 2007; Mitrushina, 2005; Šimundić, 2009). Consequently, these ratios depend on the quality of the original study, determining sensitivity and specificity. The larger the LR+ value (good LR+ > 10, excellent LR+ > 20) the better the level of diagnostic accuracy for ruling-in a disorder. The smaller the LR- value (excellent LR- < 0.1, good LR- <0.2) the better the level of diagnostic accuracy for ruling-out a disorder (Dollaghan, 2004; Šimundić, 2009). Intermediate values for LR such as LR+ = 4.0 or LR- = 0.40 are of no diagnostic value (Dollaghan, 2004). While some consider likelihood ratios a preferred measure of diagnostic accuracy because they appear to be more resistant to the influence of certain sample characteristics such as prevalence and severity (Greenslade et al., 2009; Dollaghan, 2004), they are not as commonly reported in language tests as sensitivity and specificity.

ROC curves and AUC

Receiver operating characteristic (ROC) curves resulting in the area under the curve (AUC) have been used to determine diagnostic accuracy for discrimination tasks in many fields of study (Compton et al., 2006; Swets, 1988). ROC curves and AUC are typically paired. These are the result of a graphic representation of sensitivity and specificity for various cutoff scores on

a test. The closer this curve draws to the upper left corner of the graph, the greater the area that exists under that curve. Therefore, the closer that area measurement is to 1.0 the more diagnostically accurate the test. Ratings for AUC are interpreted as follows: values greater than .90 are excellent, values between .80 and .90 are good, values between .70 and .80 are fair, and values less than .70 are poor (Compton et al., 2006).

AUC, like all evidence of diagnostic accuracy, must be examined in context. AUC is of value when comparing one diagnostic test to another test because it gives one value, combining all the possible points of sensitivity and specificity into one global measure (Johnson et al., 2009). However, that measure does not describe the shape of the curve. Šimundić (2009) illustrates this by the example of comparing two tests. One test has high sensitivity and low specificity, while the other test has low sensitivity and high specificity. When the ROC curve is plotted, and AUC calculated, the tests have the exact same AUC because even though the curves are skewed very differently, the numbers end up the same. While the two tests would appear to be equal in diagnostic accuracy, that would not be the case. One would be much better at identifying a disorder and the other at not misidentifying someone without a disorder. Thus, even though AUC values provide a simple number for quick comparison of tests, they may not be the best number on which to base clinical test use and diagnostic decisions.

Group Differences

Evidence of group differences has not typically been part of test reviews. Although Greenslade et al. (2009), in discussing the SPELT-P2, say that the minuscule overlap of the score ranges for the typical group and disordered group indicates good discrimination, this is not consistently viewed as a good measure of diagnostic accuracy. Regarding group differences, Salvia and Ysseldyke (1981) explain that basing an inference about an individual's test score,

when compared to a special population, should only be done if it has been established that only individuals in that special population achieve that particular score. This means, for example, that just because someone scores x, does not necessarily mean they have SLI unless only those with SLI score x on this test. Group differences are typically small group comparisons, which may show significant differences between the clinical and control group's scores (Dollaghan, 2004). Differences of this nature, even when statistically significant, are not sufficient evidence of a test's ability to classify an individual accurately (Dollaghan, 2004; Gray et al. 1999). There are much better measures of diagnostic accuracy, such as, sensitivity and specificity and likelihood ratios, which will be utilized here.

Test Reliability/Precision

An SLP must be able to depend on a test to be both valid, measuring the aspects of language it is intended to measure, and reliable, measuring these aspects consistently. In order for a test to be valid, it must also be reliable (Salvia & Ysseldyke, 1981). A valid inference about an individual's performance cannot be drawn from an unreliable test (Plante & Vance, 1994). The Standards combines test reliability with test precision and defines it as "… the degree to which test scores for test-takers are consistent over repeated applications of a measurement procedure in order to be considered dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group" (AERA et al., pp. 222-223). There are several factors that impact test reliability and precision. They are sample size, the number of test items, type of reliability studied, standard error of measure (*SEM*) or confidence intervals (CI), and what is considered acceptable (Charter, 2003; Charter & Feldt, 2002).

There is a relationship between sample size, number of test items, and test reliability. The larger the sample size, the more reliable the test with fewer items. Sample size is one of the biggest factors impacting test reliability and, consequently, the test's confidence intervals (Charter & Feldt, 2002). Charter (2003) states that the reliability coefficient is not necessarily as precise when calculated from a small sample, as it is from a large sample. Similarly, the larger the number of test items, the greater the test reliability. A test with more items can be found reliable with a smaller sample size if the number of test items is large enough (Charter, 2003; Tang et al., 2014).

Different types of reliability evidence can support the use of a test, such as the comprehensive nature of instructions provided for test administration, test-retest reliability, inter-examiner reliability, internal consistency reliability, and the standard error of measurement (*SEM*).

Of the various types of reliability evidence, test-retest reliability and inter-examiner reliability have been examined most often (McCauley & Swisher, 1984; Plante & Vance, 1994; Flipsen & Ogiela, 2015; Denman et al., 2017). These types of reliability evidence are generally reported as correlation coefficients. Statistically, the closer a correlation coefficient to 1.0, the stronger the measure is considered to be. There appears to be variation in what is considered acceptable levels. Plante and Vance (1994) only accepted reported coefficients of .90 or above that were statistically significant at or below $p = .05$, but they determined that some tests presented strong evidence of reliability even though they did not meet or did not report the probability level. Historically most tests reportedly did not meet this criterion or simply did not report these types of reliability evidence (Denman et al., 2017; Flipsen & Ogiela, 2015; McCauley & Swisher, 1984). Yet, Betz et al. (2013) found that 64% of the language tests in their

study had reliability correlations greater than .90 and 94% had correlations above .80, an acceptable level of reliability. When results of language tests will be used to make decisions of individual placements such as providing SLP services, a criterion of .90 or greater correlation coefficient has been considered the minimum acceptable level of reliability (Flipsen & Ogiela, 2015; McCauley & Swisher, 1984; Salvia & Ysseldyke, 1981; Salvia, Ysseldyke, & Bolt, 2010). For the purposes of the current review, .90 will be considered good and .80 will be considered acceptable

Instructions for Test Administration

Several factors can account for variations between test situations and examiners. While not usually quantified and reported as a measure of reliability, the quality of instructions for test administration is critical for reliability/precision. Sufficiently detailed instructions are necessary to ensure consistent administration of a test between situations (Flipsen & Ogiela, 2015). For test interpretations to be reasonably accurate, clinicians must be able to administer the test in a manner consistent with that used in the norming studies. While inter-examiner or inter-rater reliability provides a measure of this, often, the examiners involved in reliability studies are employees of and trained by the publishers or researchers involved in test development. When the test reaches the hands of the clinician in the field, reliability will depend to a substantial extent on how clearly the administration instructions are presented to the test user.

Test-retest Reliability

Test-retest reliability is defined as an index of stability (Salvia & Ysseldyke, 1981). It is considered to be necessary for determining the reliability of a diagnostic measure (Gray, 2003; Salvia &Ysseldyke, 1981). Test-retest reliability is generally measured by a statistically

significant correlation between scores on a test that has been administered on two separate occasions within relatively close proximity. High correlations indicate that the test is accurately measuring a skill that is stable, and therefore higher correlations can be expected for shorter time intervals between testing administrations (Gray, 2003; Salvia & Ysseldyke, 1981). When conducting reliability studies, Salvia and Ysseldyke (1981) recommended the use of coefficient alpha when alternate forms of the test are not available, and only one administration was possible.

Inter-examiner Reliability

Inter-examiner reliability is a measurement of error due to administration and scoring differences between individual clinicians and is a necessary element in test development and in the use of these tests for independent research (McCauley & Swisher, 1984; Plante & Vance, 1994). Good inter-examiner reliability means there is little variability between test results when the test is administered or scored by different clinicians for the same individual. This is important for all tests and subtests but especially so for subtests where clinician judgment is required to determine the accuracy of responses. Inter-examiner reliability is, therefore, most impacted by the quality of administration directions and the qualifications or training of the examiner (McCauley & Swisher, 1984).

Internal Consistency

An internal-consistency coefficient is "… an index of the reliability of test scores derived from the statistical interrelationships among item responses or scores on separate parts of a test" (AERA et al., 2014, p. 220). Internal consistency is generally considered an estimate of how closely related a tests' components are to each other when comparing test items to each other,

test items to the test as a whole, or subtest scores to the composite score (Anastasi & Urbina, 1997; Mitrushina, 2005). Internal consistency is typically derived from alternate-form reliability or split-half reliability and reported as a *coefficient alpha* (Mitrushina, 2005; Salvia & Ysseldyke, 1981). Some consider internal consistency to be evidence of validity because it could be viewed as a measure of homogeneity, which is relevant to determining the construct(s) the test is intended to measure (Anastasi & Urbina, 1997). Based on the common methods employed in calculating internal consistency and its predominant classification as a measure of reliability (Mitrushina, 2005; Salvia & Ysseldyke, 1981), the current review will consider it to be evidence of reliability.

Standard Error of Measurement

Standard error of measurement (SEM) is "… the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions" (AERA et al., 2014, pp. 223-224). In other words, SEM is the distribution of measurements that represent any departure from the true score (Salvia & Ysseldyke, 1981). A true score is what an individual would achieve if there were no measurement errors involved in test development or administration (Feldt & Brennan, 1989). However, because there is always some measurement error when attempting to measure human behavior, a true score can only be estimated. Therefore, SEMs are used to calculate confidence intervals (CIs), which specify a range of values within which the true score is expected to lie, given a certain probability (Dollaghan, 2004). This probability is commonly presented as a percentage.

Confidence intervals aid in interpreting test results by allowing for the possibility of an individual's true score being higher or lower than what was achieved (Charter & Feldt, 2002). The narrower CI means a greater probability that the individual's true score is very near the

actual score (Dollaghan, 2004). Narrow CIs mean the test administrator reasonably certain the individual's performance truly represents his or her ability. The wider CI means a greater probability that the individual's true score is much farther from the actual score (Dollaghan, 2004). Wide CIs are used when the test administrator is less certain the individual's performance truly represents his or her ability. As previously discussed, the larger the sample size, the smaller the *SEM*, which results in narrower CIs, which in turn yields a more precise estimate of the true score. Logically, the opposite holds as well; the smaller the sample size, the larger the SEM, which results in wider CIs, which reduces the precision of the actual score about where the true score actually falls. Typically, CIs are reported at probabilities of 68%, 90%, and 95%.

Test Fairness

The Standards say that "test developers are responsible for developing a test that measures the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics" (AERA et al., 2014, p. 64). Fairness in a language assessment means all children are administered the test most appropriate to their individual needs. The most appropriate test will account for differences in their experiences and backgrounds, reducing, if not eliminating, potential test bias. For example, bias may exist in language tests against speakers of non-mainstream dialects of American English. Dialects have their own rule-governed patterns, and what follows the rules for one dialect may violate the rules of another. Certain common characteristics of SLI, such as particular patterns of morphological errors, maybe a salient feature of SLI in speakers of Mainstream American English (MAE) dialects but may not be errors in other dialects, such as African American English (AAE), Southern White English (SWE), and many others. Therefore, it is important that a test not consider aspects of a speakers'

dialect to be errors. There should be a way to account for these errors within the standardization that does not invalidate the norm-referenced scoring. This may be done through modified procedures, providing alternative norms, correct alternative responses according to dialect, alternate cut scores, or other means. Clearly, there must be an accounting for relevant subgroups in test development that includes instructions for test administration with children from various backgrounds and varying abilities that will permit fair application of test results to the norm-reference scores, or indicate if the test is inappropriate for a particular individual.

Regarding fairness, the Standards also state that "those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test" (AERA et al., 2014, p. 64). Because most language tests are used to assist in the identification of language impairment, test developers should determine how the test may be used with children of various backgrounds, experiences, and linguistic differences. Campbell et al. (1997) examined the performance of minority children and non-majority children on two different types of language measures: processing-dependent and knowledge-dependent. They found that the two groups of children performed the same on the processing-dependent measures. In contrast, the group of minority children performed at a level slightly more than one standard deviation below the group of majority children on the knowledge-dependent measure. Similarly, Paraskevopoulos and Kirk, (1969) state that psychometric quality is dependent on both the test itself and the group of individuals with whom it is used. This means that reliability and validity data should be collected from studies performed on subgroups composed of individuals likely to be referred for evaluation. This would indicate that relevant subgroups must be included in reliability and validity studies, or these groups must be included in the normative sample.

Another reason why relevant subgroups should be included in preliminary test development is the potentially different performance of certain subgroups such as English language learners (ELLs). For example, in an examination of screening instruments, Johnson et al. (2009) compared screening results for children who were ELLs, children on a free or reduced-cost lunch (FRL) program, children who were non-ELL, and children who were non-FRL). They utilized screening results from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) and the Peabody Picture Vocabulary Test—Third Edition (PPVT; Dunn & Dunn, 1997). Statistical analysis using ROC curves provided data on the cut scores necessary to achieve 90% sensitivity for each subgroup. Results demonstrated that different cut scores were needed for different subgroups. Across screeners, the lowest cut scores were needed for ELLs (Johnson et al., 2009).

The process of distinguishing between language impairment and a dialectal difference or ELL is complex and multi-leveled. The more fairness has been addressed in test development, the more effective the test's role in language assessment.

Methods

Test Selection

Surveys have indicated that SLPs most frequently choose to administer omnibus/multi-domain tests to school-age children who are suspected of having a language-impairment (Betz et al., 2013; Montzka, 2015). The current test review was therefore limited to omnibus/multi domain language tests for school-age children that are focused on oral language. Potential language tests were identified through literature review, online publisher searches, and online searches for assessment lists. Available information for each potential test was examined to

determine inclusion or exclusion. Inclusion and Exclusion criteria are listed in Table 1. Applying these criteria during preliminary review identified the tests listed in Table 2 for detailed review. Even though the TACL-4 and TEXL assess only receptive and expressive language, respectively, they were developed so that they could be used in conjunction with each other and were standardized on the same normative sample. Therefore, these tests were also included.

Table 1
Test Inclusion and Exclusion Criteria

Inclusion	Exclusion
Norm-referenced, standardized language assessment	Screening tests Tests of Academic achievement Checklists, interview-based assessments, and questionnaires Tests primarily assessing articulatory/phonological production Tests primarily assessing phonological processing and/or phonological awareness Tests primarily assessing reading Tests primarily assessing writing Tests of written language only
Targets school-age children ages 5-12	Most of the targeted age range is above or below the 5- to 12 age range
Currently in print	More than five years out of print
Currently available (online by the publisher, digital copy, etc.) Most current edition	No longer available Previous editions if more recent publication is available.
For monolingual English speakers	
Both receptive and expressive oral language	Tests of only receptive, or only expressive language, without a comparable counterpart.
Includes the domains of morphology, syntax, and semantics	Single domain only
	Tests primarily assessing pragmatic skills Pragmatics tests

Procedures for review of selected language tests

Each test and its manuals were reviewed systematically for evidence of psychometric properties by an SLP graduate student (the current author) and a certified speech-language pathologist. After an initial trial review and meeting to finalize the process of locating and documenting evidence, each reviewer examined the tests independently. Findings were then compared for agreement. Disagreements between results were discussed and reexamined by both reviewers with re-evaluation of information in the test manuals when needed, in order to reach consensus.

Table 2
Tests Identified for Review

Test Acronym	Test	Intended Population
ALL	*Assessment of Literacy and Language* (Lombardino et al., 2005)	PK - 1st grade
CASL-2	*Comprehensive Assessment of Spoken Language–Second Edition* (Carrow-Woolfolk, 2017)	3 - 21;11[a,b]
CELF-5	*Clinical Evaluation of Language Fundamentals--Fifth Edition* (Wiig et al., 2013)	5 - 12
DELV-NR	*Diagnostic Evaluation of Language Variation-Norm-Referenced* (Seymour et al., 2005)	4 – 9[d]
ITPA-3	*Illinois Test of Psycholinguistic Abilities-Third Edition* (Hammill et al., 2001)	5;0 - 11;11[a,b,c]
OWLS II	*Oral and Written Language Scales, Second Edition* - Listening Comprehension and Oral Expression (Carrow-Woolfolk, 2011)	3;0 - 20;11[b]
RESCA-E	*Receptive, Expressive & Social Communication Assessment–Elementary* (Hamaguchi & Ross-Swain, 2015)	5 - 12[a,b,c]
TACL-4	*Test for Auditory Comprehension of Language-4th Edition* (Carrow-Woolfolk, 2014)	3;0 - 8;11[c]
TEXL	*Test of Expressive Language* (Carrow-Woolfolk, 2014)	3 - 12;11[c]
TELD-4	*Test of Early Language Development-4* (Hresko et al., 2017)	3 - 7;11[b,c]
TILLS	*Test of Integrated Language & Literacy Skills* (Nelson et al., 2016)	6 - 18;11[d]
TOLD-I:4	*Test of Language Development-Intermediate, Fourth Edition* (Hammill & Newcomer, 2008)	8;0 - 16;11[c]
TOLD-P:5	*Test of Language Development-Primary, Fourth Edition* (Newcomer & Hammill, 2019)	4;0 - 8;11[e]

Note. Exclusions are either specifically stated or implied by test author's description.
[a] Excludes individuals with severe disabilities (sensory, cognitive, motor, or developmental). [b] Excludes individuals who are not proficient English speakers. [c] Excludes individuals who are unable to understand the directions or formulate oral responses. [d] Excludes individuals who are not native speakers of English. [e] Excludes individuals who speak dialects other than Standard American English.

Procedures

The current review examined the extent to which currently available multi-domain norm-referenced language tests for school-age children either meets or reports on the targeted Standards (AERA et al., 2014), as shown in Appendix A, according to the evaluation criteria described below. Each of these psychometric properties has an over-arching standard and numerous subsequent standards, all of which provide detailed explanations to inform appropriate application. This review focused on the foundational aspects of norms, validity, diagnostic accuracy, reliability/precision, and fairness.

Evaluation Criteria

Criteria used in this review were determined for five aspects of psychometric quality based on the Standards, research, and historical precedent. The first aspect examined the size and characteristics of the normative sample.

Normative Sample. Due to the foundational importance of the normative sample, three criteria for norms were examined first in order to facilitate reviewer understanding of the evidence presented for the other psychometric properties. The first criterion was whether the demographics of normative sample were described in sufficient detail to allow for appropriate comparison to allow for peer comparisons. This meant that demographic information was examined specifically for age, gender, ethnic representation, parent education/socioeconomic status (SES), disorder/diagnosis exceptionality status, and geographic location. The second criterion was regarding the normative sample size, specifically the cell size used to determine the scoring subgroups. Cells consisting of 49 participants or less were considered inadequate. While cells of 50 to74 participants were considered acceptable, and 75 to 99 participants considered good, 100 or more participants has historically been preferred and was considered to be the best. These numbers, when not specifically provided in terms of scoring cells, were calculated using simple division of the total number of participants in a given age subgroup by the number of scoring cell subgroups presented in the scoring indexes of the test manual. Thus, in some cases they represent estimates based on the information provided. For example, the normative sample data might present a sample of n=125 for 5-year-olds and present a scoring index with scoring cells in four-month intervals. Therefore, 125 would be divided by 3 to arrive at an estimated cell size of 41.6 which provides an approximate sample size per scoring cell. It should be noted that this number is not necessarily the exact number in each cell. Following the above example, it is possible, but unlikely, that there could have been 100 children within the first four-month age interval and much smaller numbers in the other two age intervals. Due to the relevance of this criterion to diagnostic accuracy or determining the severity of impairment, the third criterion was whether a full-range or truncated normative sample was used.

Validity. There is little consistency in what and how evidence of validity is determined by test developers and how it is presented across the tests included in the current review. Because the Standards do not require specific types of validity, nor specify levels of acceptable evidence, this review reported whether or not certain evidence was present in the test manuals. One set of criteria concerning test validity pertained to the clarity of information that would inform appropriate test use. Specifically, this review reported whether or not the test clearly stated its purpose(s), whether or not diagnosis was one of those purposes, whether or not the test provided the theoretical foundation upon which it was developed, and what constructs the test claimed to assess. A second set of criteria was examined regarding concerning validity studies, reporting whether the samples used to determine validity were pulled from the normative sample, from an independent group of individuals who were not included in the normative sample, or if the origin of the sample was either unspecified or not clearly explained. Also, regarding validity study samples, the tests were examined to determine whether or not the demographics were adequately described, and if so, was the sample comparable to the normative sample or was it limited in the range of demographic characteristics represented. This information is relevant to establishing test validity for specific populations. The third set of criteria was what types of validity evidence were presented in support of content-related validity and in support of structural validity. The specific types of evidence for content validity that were most consistently presented and allow for comparison across tests, were item analysis, concurrent validity and predictive validity. Likewise, the specific types of evidence most consistently reported for structural validity were factor analysis and inter-correlations of the subtests with each other and the subtests with any composites or the test as a whole.

Diagnostic Accuracy. Diagnostic accuracy was evaluated according to criteria based on the type evidence presented in the test manuals. Based on earlier reviews and historical precedent, sensitivity and specificity were considered good at 90%, with 80% deemed acceptable, and the cut-off score at which the best paired percentages were achieved was reported. Predictive values, which measure the probability of an accurate classification were also considered good at 90%, with 80% deemed acceptable, and the base rate at which these percentages were achieved was reported. Valuable measures for ruling-in or ruling-out a diagnosis, likelihood ratios were considered good when LR+ > 10, good when LR- < 0.1, excellent when LR+ > 20, and excellent when LR- <0.2. Values for LR such as LR+ = 4.0 or LR- = 0.40 or less were considered of no diagnostic value. Another measure, one that looks at the overall diagnostic accuracy of a test is the relative operating characteristic (ROC) curves resulting in the area under the curve (AUC). AUC values > .90 were considered excellent, values between .80 and .90 good, values between .70 and .80 fair, and values less than .70 poor.

Reliability. The first criterion examined for reliability was whether or not sufficient detail, examples, and acceptable variables were provided in the test's instructions to ensure that the test could be administered in a manner consistent with the norming studies. The second criterion was the reliability coefficients that were presented for test-retest reliability, inter-examiner reliability, and internal consistency reliability. These coefficients were considered excellent at >.90, acceptable between .80 and .90, poor when <.80. The last reliability criterion was the reported standard error of measurement (SEM) for the whole test, subtests, or reported CIs. This criterion was reported without judgement in order to allow for comparison between tests by the reader according to relative size of SEM or width of CI from one test to another.

Fairness. The first criterion in reviewing fairness was whether or not the test's instructions accounted for its use with children of various backgrounds, experiences, and linguistic differences so that test results could be applied fairly to the norm-reference scores. Also, the tests were examined to determine whether or not the manual indicated if the test was inappropriate for certain individuals. A second criterion was whether or not relevant subgroups, such as clinical populations had been included in reliability and validity studies or in the normative sample for the test.

Results

Normative Sample

All thirteen tests reviewed included the minimum demographic information for the normative sample; age or grade, gender, SES or parent education or income, race or ethnicity, and geographic region. Uniquely, the TELD-4 normative data were weighted to improve the representativeness of the normative sample and to increase sample size for under-represented groups that the developers had found difficult to obtain during standardization. In doing this, they accounted for approximately 2% of the normative sample, a relatively small portion overall. However, this does raise the concern of a norming procedure based on a too small sample size, weighting not-withstanding.

The predominant approach to normative samples appears to be the full-range sample. All but two tests (CASL-2 and TILLS) attempted full-range normative samples. There is, however, a great deal of variety in who was actually included in normative samples. For example, the DELV included individuals with mild impairments who were primarily in general education classrooms that are age/grade appropriate (implies exclusion of more severe disorders), while the CELF-5 included individuals with a variety of disorders, individuals with languages other than English

(spoken in home or reportedly bilingual) and/or speakers of non-MAE dialects, individuals with language impairment, and individuals with sensory impairments such as visual impairment or hearing impairment. Most tests using a full-range sample included individuals with a variety of disorders (some include gifted and talented) and individuals with language impairment. By contrast, while the CASL-2 did include individuals with mild impairments who were primarily in general education classrooms that are age/grade appropriate (implies exclusion of more severe disorders), they excluded individuals who are not primarily in general education classrooms (implying the exclusion of more severe disorders, those in special education classrooms, as well as gifted and talented) and they excluded individuals with intellectual disability, creating a truncated sample. Similarly the TILLS excluded individuals with language impairment or any who were likely to have moderate to severe language impairment, and individuals who are not primarily in general education classrooms (also implying the exclusion of more severe disorders, those in special education classrooms, as well as gifted and talented).

Most tests reported samples of near or above 100 per age within the manuals. However, only three tests, the ALL, CELF-5, and DELV-NR presented norms based on sample sizes of at least 100 per scoring cell consistently across scoring subgroups. However, the TILLS met the 100 participant per cell benchmark, with the exception of 1 group that had 98 individuals. While most tests had one or two scoring cells with 100 or more participants, the majority of tests had scoring cells with 74 or fewer participants. The fact that this is predominantly the case prompted the need to revisit the reasons behind this "rule of thumb" when evaluating test norms. As indicated earlier, some research has suggested that scoring groups of over 100, may not be necessary (Salvia & Ysseldyke, 1981; Weiner & Hoock, 1973). Other research suggests that as few as 50 participants may be an acceptable number for determining normative data (Bridges &

Holler, 2007; Crawford & Howell, 1998; Mitrushina, 2005). See Table 3 for the cell size for scoring group size by age. Although groups of at least 50 may be considered adequate, it is concerning that seven of the tests are presenting normative data based on samples that are fewer than 50, at least as some ages, when the reported sample size is divided by the number of actual scoring cells. For example, the RESCA-E presents scoring cells for 5- and 6-year-olds as n=95. However, when divided by four based on the three-month interval for scoring subgroups the actual scoring cell for these two age groups is only n=23.75. Twenty-four individuals cannot reasonably provide an adequate sampling of the full range of what is normal (Bridges & Holler, 2007).

Two additional findings from this review include observations regarding the reporting of means and SD, and gender-based scoring. First, interestingly the only two tests that used a truncated normative sample, CASL-2 and TILLS, were also the only two that did not report the means and SD for subgroups by age. Approximately half the tests reported the means and SD for subgroups by demographic subgroups. The only test that did not report the means and SD according to clinical subgroups was the ALL. If clinical subgroups and typically developing matched samples have similar means and SD that could potentially indicate a lack of discriminatory power in the test. If there are apparent differences in means and SD for demographic subgroups it could potentially indicate bias. Therefore, it is worth recording which tests provide this relevant information. Second, while it had historically been thought that there were developmental differences between genders regarding language development, research has not supported this. None of the tests provided gender-specific scoring. The complete absence of gender-specific scoring indicates that test developers acknowledge this.

Table 3
Normative Sample Findings

Test Acronym	Minimum Demographic Information	Type of Sample	Scoring Cell Size within the Normative Sample by Age				Included Means and SD for Subgroups		
			N < 50	*N* =50 to 74	*N* = 75 to 99	*N* ≥ 100	Age or Grade	Demo-graphic	Clinical
ALL	Yes	Full-range[a,b,c,f]				all grades*	Yes	NR	NR
CASL-2	Yes	Truncated[d,h,i]		5 - 15	3 - 4	16 - 21	NR	NR	Yes
CELF-5	Yes	Full-range[a,b,c,e]				all ages	Yes	NR	Yes
DELV-NR	Yes	Full-range[a,b,c]				all ages	Yes	NR	Yes
ITPA-3	Yes	Full-range[a,c]	5 - 7, & 9	8 & 10		11 - 12	Yes	Yes	Yes
OWLS II	Yes	Full-range[d]	3, 4, 6, 7, & 9	5, 8, 10, & 11		combined 16 - 18 & combined 19 - 21	Yes	NR	Yes
RESCA-E	Yes	Full-range[a,c]	5 - 6	7 - 10	12	11	Yes	NR	Yes
TACL-4	Yes	Full-range[a,c]	3 - 5	6 - 8		9 - 12	Yes	Yes	Yes
TEXL	Yes	Full-range[a,c,e]	3 - 8	9 - 12			Yes	Yes	Yes
TELD-4	Yes	Full-range[a,c]	3	4 - 7			Yes	NR	Yes
TILLS	Yes	Truncated[g,j]			7;0-7;5	all *except* 7;0 - 7;5	NR	Yes	Yes
TOLD-I:4	Yes	Full-range[a,c]	16 - 17	8 - 9, & 11 - 15	10		Yes	Yes**	Yes**
TOLD-P:5	Yes	Full-range[a,c]			4, 5, & 7	6 & 8	Yes	Yes***	Yes

Note. *Normative sampling was by grade, not by age. ** Reported standard score means only. *** Gender and race/ethnicity only.
NR = Not Reported. Minimum demographic information included age or grade, gender, SES or parent education or income, race or ethnicity, geographic region.
[a] Includes individuals with a variety of disorders (some include gifted and talented). [b] Includes individuals with languages other than English (spoken in home or reportedly bilingual) and/or speakers of non-MAE dialects. [c] Includes individuals with language impairment. [d] Includes individuals with mild impairments who were primarily in general education classrooms that are age/grade appropriate (implies exclusion of more severe disorders). [e] Includes sensory impairments such as visual impairment or hearing impairment. [f] Excludes moderate to severe behavioural or emotional disorders. [g] Excludes individuals with language impairment or who are likely to have moderate to severe language impairment. [h] Excludes individuals who are not primarily in general education classrooms (implies exclusion of more severe disorders, those in special education classrooms, as well as gifted and talented). [i] Excludes individuals with intellectual disability. [j] Excludes individuals with uncorrected sensory impairments such as visual impairment or hearing impairment.

Validity

Due to the variability of validity measures reported in the test manuals, the ambiguity of certain terms (e.g. predictive validity), and the fact that key aspects of validity are qualitative rather than quantitative, it is challenging to report on validity in a way that allows for clear comparison of validity across tests. Thus, the results reported here primarily pertain to whether or not certain types of validity are presented by the publishers rather than whether a certain benchmark was reached for the various types of validity.

While certain elements are generally clearly presented in the manuals, such as demographic information for the samples in validity studies, others, such as the source of those samples, are not always as clear. Many of the validity studies presented in the reviewed test manuals were conducted with subgroups of the normative sample. When this sample included individuals with language impairment, validity studies conducted on independent samples seem to have often been bypassed completely. Across tests, for criterion predictions in studies of concurrent validity, there was considerable variation in the basis for comparison.

For some tests, certain elements regarding the sample demographics of validity studies were either not provided or not clearly described. The Standards state that "(t)he composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics" (AERA et al., 2014, p. 25). The OWLS-II and other tests, presented the demographic information for validity studies as numbers, not percentages (as it is reported in the normative data) which therefore makes it difficult to easily verify or compare the two samples. There was a distinct lack of clarity of information in the validity studies for the

TACL-4 and TEXL. This review found differences in what demographics were reported between validity studies and norming studies. Some demographic charts reported numbers while others reported percentages. It was unclear exactly how many validity studies had been conducted, and it was difficult to distinguish among them. There were also inconsistencies in the categories of exceptionality, the numbers or percentage, and ethnic and gender representation between the normative sample and validity samples. It was unclear in validity study information for the TELD-whether the samples were independent of the normative sample or drawn from it.

For other tests, there was often something about the samples used in the validity studies that may weaken the resulting evidence. For example, the CELF-5 conducted a validity study on a sample that was missing individuals from 17 to 21 years of age, not evaluating the test's validity for the entire age range it claims to cover. For concurrent validity, the RESCA-E studied a sample of children who all had learning disability, but no typically developing and no other disorder subgroups were included. The sample was also limited in the regions it was drawn from, as well as the range in SES. This limits interpretation of their findings.

As evidence of content validity, all tests, except the OWLS II reported conducting either a formal item analysis or described an item response or response process analysis. While all thirteen tests provided evidence claiming either concurrent validity or predictive validity, in reviewing the manuals, the term predictive validity is often used when comparing the test in development to other established tests without any element of predicting future performance. This is more in line with concurrent validity than predictive, and results in a mismatch of data when attempting to report these measures across tests. Most concerning was that some test developers used previous versions of the tests as a basis of comparison such as the CELF-5, which provided comparisons with the Clinical Evaluation of Language Fundamentals - Fourth

Edition (CELF-4; Semel et al., 2003) and Clinical Evaluation of Language Fundamentals Preschool-2 (CELF-P-2; Semel et al., 2006). Other developers used tests with very different content like vocabulary only tests or pragmatics only tests such as the TOLD-I:4, which provided comparisons with the primary age version of itself, the (Test of Language Development-Primary: Fourth Edition (TOLD-P:4; Newcomer & Hammill, 2008), which is not an independent comparison, or the Pragmatic Language Observations Scale (PLOS; Newcomer & Hammill, 2009), Peabody Picture Vocabulary Test- Third Edition (PPVT-3; Dunn & Dunn, 1997), and Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV; Wechsler, 2004). These latter tests do not measure the same aspects of language overall as the TOLD-I4 as they focus only on pragmatics and receptive vocabulary, respectively. Thus, they aren't ideal for validity comparisons with a multi-domain test like the TOLD-I4. Without firm criteria to evaluate the values provided, and given the variations in reporting, this review is limited to reporting whether or not a particular type of evidence was present in the test manual and not their quantitative levels. That will need to be evaluated on an individual basis by SLPs when choosing whether or not to use a test. Additionally, validity evidence, when based on a comparison with other tests and having no element of predicting future performance more closely meets the definition of concurrent validity. Therefore, even though some test developers refer to this evidence as predictive validity, this review classified it as concurrent validity.

There are areas of concern in the reported correlations of these measures of comparison. For some tests the reported correlations seem low when compared to tests that claim to evaluate the same constructs. For example, the correlation between the OWLS II and the CELF-4 was $r = .45$ at $p < .005$ for the receptive language composite, and $r = .59$ at $p < .001$ for the expressive language composite. The CELF-5 reports correlations by subtest with CELF-4 (previous version)

and with PPVT-4 and EVT-2 (tests of vocabulary only). Again, sample size is concerning with the RESCA-E, which used samples of only 13 and 21 individuals to compare the RESCA-E to the OWLS-II. When these were examined closely it was found that some subgroups consisted of only 1 individual and others were not represented at all. Validity for the RESCA-E was studied with only one clinical group, language disordered, and this was done without a typical developing control. The TACL-4 claims criterion-prediction validity. However, since it is comparison to performance on other tests (CELF-4 or P2, OWLS II, Diagnostic Achievement Battery [DAB], and TEXL) it would appear to be concurrent validity. There was a significant difference between the TACL-4 and CELF-4 for receptive language scores which would not be expected. A last example of these concerns is the TELD-4 which provided concurrent validity with other measures the Preschool Language Scale – Fifth Edition (PLS-5; Zimmerman et al., 2011), TACL-4, Bankson Expressive Language Test–Third Edition (BELT 3; Bankson et al., 2018), TEXL, and Young Children's Achievement Test, Second Edition (YCAT-2; Hresko et al., 2018) that had large to very large correlations. However, although the authors claim trivial or small magnitudes, the different tests often had significantly different scores as measured by t-test. This is specifically of concern for tests that seem to measure the same aspects of language (e.g., TELD-4 & TACL receptive language scores, TELD-4 & BELT expressive language scores, TELD-4 & TEXL expressive language, etc.).

The most consistently provided evidence of structural validity were a factor analysis and inter-correlations of the subtests with each other and the subtests with any composites or the test as a whole. The Standards do not provide specific criteria for what these should be, only that evidence for structural validity be provided. Correlations should vary depending on what is being compared as they are a measure of relatedness. The contribution of a factor analysis relies on

how the factor loadings reflect on the theoretical foundation upon which the test is build. Given the variability of such measures and the need to examine each on its basis of comparison, Table 4 reports its presence or absence rather than whether or not it reaches a particular benchmark, as the Standards don't provide such benchmarks. All thirteen tests provided evidence of either factor analysis or inter-correlations. In fact, ten tests provided both.

Factor analysis appears to support the development and use of composite scores, however, it does not necessarily suffice as a rationale for such composites. Factor analysis, as was provided by all the tests except the DELV-NR and TELD-4, appears to support the use of at least some of the composite scores. This is due to factor loading on a single factor such as language, two factors such as expressive and receptive language, or more depending on the language models used. For example, exploratory factor analysis for the TILLS resulted in factors that did not load on an expressive/receptive language dichotomy, nor on an oral/written language dichotomy. Instead, factors loaded in a manner consistent with the TILLS' theoretical model which is based on the premise that these aspects of language have more in common than not.

The Standards do state that the basis and rationale for arriving at the composites should be provided. Even though this is clearly stated in the Standards, except for the CASL-2, CELF-5, and TILLS, which did describe the reasoning behind their respective composite scores, most tests appeared to assume that composites were simply standard procedure and provided the calculations for deriving them. For some tests the rationale behind composites could be inferred from the inter-correlations and factor loadings, but most did not clearly address it. Three tests, the ITPA-3, TELD-4, and TOLD-I:4 claimed composites to be the most reliable or useful, however, considering the fact that the greater the number of items on the test the higher the tests

reliability, this is only to be expected. It does not necessarily follow that these composites are truly the best measures.

Table 4

Test Validity Findings

Test Acronym	Diagnostic Purpose?	Theory?[a]	Samples for Validity Studies			Demographics of Validity Study Samples		Evidence Presented for Content Validity[c]		Evidence Presented for Structural Validity	
			Norms	Indep.	Unspec.	Adeq. Discrip.	Comp - Norms	Item Analysis	Conc.	Factor Analysis	Inter-corr.
ALL	Yes	Yes	Yes	NCS	NCS	Yes	Yes	Yes	Yes	Yes	Yes
CASL-2	Yes	Yes	Yes	Yes	NCS	Yes	No	Yes	Yes	Yes	Yes
CELF-5	Yes	No	NCS	NCS	Yes	Yes	No	Yes[b]	Yes	Yes	Yes
DELV-NR	Yes	Yes	Yes	NCS	Yes	Yes	No	Yes[b]	Yes	NR	Yes
ITPA-3	No[a]	Yes	NCS	NCS	Yes	No	No	Yes	Yes*	Yes	NR
OWLS II	Yes	Yes	NCS	NCS	Yes	No	No	NR	Yes	Yes	Yes
RESCA-E	No[b]	Yes	Yes	NCS	Yes	No	No	Yes	Yes	Yes	NR
TACL-4	Yes	Yes	Yes	Yes	NCS	No	No	Yes	Yes*	Yes	Yes
TEXL	Yes	Yes	Yes	Yes	NCS	No	No	Yes	Yes*	Yes	Yes
TELD-4	Yes	Yes	Yes	NCS	Yes	No	No	Yes	Yes*	NR	Yes
TILLS	Yes	Yes	Yes	Yes	NCS	No	N/A	Yes	Yes	Yes	Yes
TOLD-I:4	Yes	Yes	Yes	NCS	NCS	Yes	No	Yes	Yes*	Yes	Yes
TOLD-P:5	Yes	Yes	Yes	NCS	NCS	No	No	Yes	Yes*	Yes	Yes

Notes. Diagnostic Purpose? = Is diagnosis a purpose for the test? Theory? = Was a Theoretical Foundation Provided? Adeq. Discrip. = Adequate Description. Comp-Norms = Comparable to Normative Sample. Conc. = Concurrent Validity. Inter-Corr. = Inter-correlations. NCS = Not clearly specified in the manual. NR = Not reported. N/A = Not Applicable. Norms = Validity studies utilized samples from the normative (standardization) sample. Indep. = Validity studies utilized samples that were not part of the original normative (standardization) sample. Unspec. = The test did not clearly state the sampling method, whether pulling from normative sample or from an independent group of subjects.
* Validity evidence based on a comparison with other tests and had no element of predicting future performance, meeting the definition of concurrent validity, however, these tests called it predictive validity.
[a] Some of tests only described their model or theory, but others also provided research references and citations. [b] Item response or response process analysis appeared to be equivalent or in lieu of a formal item analysis. [c] Concurrent and predictive validity are reported according to the terminology used in the test's manual, regardless of was actually being measured.

Diagnostic Accuracy

If diagnosis is a purpose for a test, then evidence of diagnostic accuracy should be presented. Diagnosis was considered a purpose for a test if, in the test manual, it was clearly state as a purpose or implied by a purpose of identifying performance significantly below peers.

While it is recognized that a low score does not a diagnosis make, it is an important component in the diagnosis process. The OWLS II claims to help diagnose, based on the stated purposes to "identify strengths and weaknesses in language, help determine the existence of language delays and disabilities, and help guide eligibility for services and intervention planning," yet offers only group differences as evidence of diagnostic accuracy. Group differences are not strong evidence of diagnostic accuracy, and because much stronger measures exist and are widely in use, group differences were not considered evidence of diagnostic accuracy in this review. Interestingly, even though group differences are not strong evidence of diagnostic accuracy, nearly every test reports some sort of group differences.

In Table 5, findings for diagnostic accuracy are reported for overall or composite scores. The diagnostic accuracy for individual subtest may vary. Sensitivity and specificity were reported for all tests that provided evidence of diagnostic accuracy. Most tests had good levels (sensitivity and specificity > .80) at -1 SD or even less. In fact, the CASL-2 reported the highest levels of sensitivity and specificity, which were .86 and .76 respectively, at - 0.7 SD which is a standard score cutoff of 90. Only three tests out of the ten who reported sensitivity and specificity, had good levels at -1.5 SD. While the TILLS did not use SD as a means of providing cutoffs for diagnostic accuracy, it provided cut scores for three age ranges that have good levels of sensitivity and specificity, evidencing the diagnostic accuracy of that test. The TILLS is also the only test to provide likelihood ratios, which at LR+ 4.34 -9.7 and LR- .03-.24 are fair to good for ruling in language impairment, and good to excellent for ruling it out.

Positive and negative predictive values seem to be referred to by some tests as positive and negative predictive powers. These appear to have the same meaning and will be referred to in this review as positive and negative predictive values. The ALL, CASL-2, CELF-5, DELV-

NR, and TOLD-I:4 report predictive values. The first three of these had the highest levels (most were excellent > .90) for both positive and negative predictive value at -1 SD. At -1.5 SD, the negative predictive values are consistently fair to poor, running from .74 all the way down to .43 across all three tests except for the CELF-5 which, at 60% base rate had a PPV = .99/.81 at - 1.5 SD. It is interesting that the TOLD-I:4 only reports positive predictive values, not negative predictive values, nor do they provide the base rate or SD at which this range (.60 to .73 rather poor diagnostic accuracy) was calculated.

The AUC was reported for the CASL-2, TACL-4, TEXL, TELD-4, and TOLD-P:5. The TEXL and TELD-4 reported excellent AUC values > .90 while the CASL-2 and TACL-4 reported values >.80 (good). The TOLD-P:5 based on two types of accuracy, one in predicting other criterion (.90), and one in predicting diagnosis (.78).

The need to critically consider diagnostic accuracy in terms of what exactly these numbers are based on is essential. The two primary reference points test developers used to calculate diagnostic accuracy were existing diagnoses or performance on other tests. Of the two, existing diagnosis seems more relevant to diagnostic accuracy. The latter is more reflective of concurrent validity than the test's ability to truly differentiate impairment from a lack of impairment. Measuring diagnostic accuracy by comparing one test to another assumes that the test being compared to has good diagnostic accuracy. However, this is not necessarily true when there is no clear reference standard for diagnosis. Of the ten tests that presented evidence in support of diagnostic accuracy, the CASL-2, DELV-NR, TELD-4, and TILLS referenced existing diagnoses. The CELF-5 and TOLD-I:4 appear to reference performance on other tests only. The TACL-4, TEXL, and TOLD-P:5 appear to reference both existing diagnoses and

performance on other tests. With rather unique wording, the ALL references an existing diagnosis of SLI based on performance on other tests.

A final, but disconcerting note regarding diagnostic accuracy is the lack of clarity in reporting for the TOLD-P:5. The authors of the manual seem to be using the term SLI (Specific Language Impairment) and speech and language impairment (SLI), an acronym used by IDEA 2004 and many school districts to designate a primary diagnostic category of a speech or language impairment (In the absence of some other causal impairment), interchangeably (see Table 6.15 on pg. 88 – 91 of the TOLD-P:5 manual). This is confusing and may also indicate that the group they are referring to in the manual is not actually a group of children diagnosed with specific language impairment. Also, criterion Prediction for the TOLD-P:5 is listed for multiple tests but they did not describe how this comparison was conducted. It is unclear if all participants took all tests or if tests are differentiated.

Reliability

Table 6 reports findings for test reliability/precision related to standardization, test administration, SEM, and confidence intervals. With very few exceptions, all the tests provided clear administration instructions with adequate scoring parameters and examples that these reviewers judged to allow for test administration consistent with the test's standardizations procedures. Exceptions were determined to be somewhat limited in guidance, but not entirely lacking. For the ITPA-3, the Sight Decoding and Sound Decoding subtests do not provide instructions for severe articulation impairment or irregular/multiple patterns as it requires correct pronunciation. Omitting these subtests would prohibit the use of 5 of their eleven composites, including the General Language composite which looks at all twelve subtests combined.

Table 5
Diagnostic Accuracy Findings

Test Acronym	Diagnostic Purpose?*	Was Evidence Provided?	Types of Diagnostic Accuracy Evidence: Sensitivity / Specificity at Highest Balanced Levels	Positive/Negative Predictive Values at Referral Base Rates	AUC	L+/L-	Reference for Language Impairment[g]
ALL	Yes	Yes	-1 SD (SS 85) .98/.89 -1.5 SD (SS 77) .86/.96	70% BR .96/ .96 at -1 SD, .98/.74 at -1.5 SD 80% BR .97/.93 at -1 SD, .99/.63 at -1.5 SD 90% BR .99/.85 at -1 SD, 1.00/.43 at -1.5 SD	NR	NR	ECD CT
CASL-2	Yes	Yes	-0.7 SD (SS 90) .86/.76) -1 SD (SS 85) .74/.84)	NR	0.89	NR	ECD & RC
CELF-5	Yes	Yes	-1 SD (SS 85) 1.0/.91 -1.3 SD (SS 80) .97/.97 -1.5 SD (SS 77) .85/.99	60% BR .94/1.00 at -1 SD, .98/.96 at -1.3 SD, .99/.81 at -1.5 SD 70% BR .96/1.00 at -1 SD, .99/.93 at -1.3 SD, .99/.74 at -1.5 SD 80% BR .98/1.00 at -1 SD, .99/.89 at -1.3 SD, 1.00/.62 at -1.5 SD	NR	NR	CT
DELV-NR	Yes	Yes	-1 SD (SS 85) .95/.93 -1.5 SD (SS 77) .69/.99	60% BR .95/.93 at all -1 SD, .99/.68 at -1.5 SD 70% BR .97/.90 at -1 SD, .99/.58 at -1.5 SD 80% BR .98/.84 at -1 SD, 1.00/.45 at -1.5 SD	NR	NR	ECD
ITPA-3	No[a]	No	NR	NR	NR	NR	NR
OWLS II	Yes	No	NR	NR	NR	NR	NR
RESCA-E	No[b]	No	NR	NR	NR	NR	NR
TACL-4	Yes	Yes	-0.5 SD (SS 92) .71/.84 -0.7 SD (SS 90) .68/.87	NR	0.86	NR	SD & CT
TEXL	Yes	Yes	-0.5 SD (SS 92) .83/.81 -0.7 SD (SS 90) .80/.31	NR	0.90	NR	SD & CT
TELD-4	Yes	Yes	-0.5 SD (SS 92) .87/.85 -0.7 SD (SS 90) .86/.86 -1 SD (SS 85) .81/.95	NR	0.93	NR	ECD
TILLS	Yes	Yes	6-7;11[c] (CS 24) .84/.84 8-11;11[c] (CS 34) .88/.85 12-18;11[c] (CS 42) .86/.90	NR	NR	4.34 - 9.7/.03-.24	ECD & ES
TOLD-I:4	Yes	Yes	-0.7 SD (SS 90) range .71/.92 to .80/.95[d]	ranged from .60 to .73/NR[d]	NR	NR	CT
TOLD-P:5	Yes	Yes	-1 SD (SS 85) .94/.84[e] -1.3 SD (SS 81) .91/.85[e] -1.5 SD (SS 78) .88/.88[e] -0.5 SD (SS 93 .96/.71[f] -0.65 SD (SS 90) .86/.81[f] -0.85 SD (SS 98) .76/.92[f]	NR	0.90[e] .78[f]	NR	CT &/or ECD

Note. *Yes, if specified diagnosis or implied by identifying significantly below peers. ECD = Existing clinical diagnosis. SD = School Diagnosis. ES or RS = Eligible for, or Receiving Services. CT = Based on comparison with other test(s). Sensitivity/Specificity is the percentage of accurately classified individuals in a sample (good .90, acceptable .80). Positive/Negative Predictive Values are the statistical probability of having or not having an impairment (good .90, acceptable .80). Relative operating characteristic (ROC) curves result in the area under the curve (AUC) (excellent > .90, good .80 - .90). L+ L- Likelihood Ratios represent the confidence that the tests score representing impairment or unimpaired came from someone with or without impairment, respectively (excellent LR+ > 20, good LR+ > 10) (excellent LR- < 0.1, good LR- <0.2). BR = Base Rate. SS = Standard Score cutoff. CS = Cut Score. [a] Not directly. Stated purpose is to identify risk for school failure & strengths & weaknesses in linguistic abilities. [b] Stated purpose is to provide information about a child's language development & social communication behaviors. [c] For ages. [d] Reported sensitivity/specificity & predictive values (positive only) a rage for each of the four reference tests & a combined (No base rate for PPV). [e] Based on predicting criterion measures. [f] Based on differentiating children with prior diagnosis. [g] Reference for language impairment is discussed in detail under the Discussion heading.

Due to variability in how SEM is presented between tests, SEM is reported here as an approximate. Values are rounded to the nearest whole number by composites and subtests, ranging from less than 1 to 6 for both. Confidence intervals were provided at various percent confidence levels, most commonly at 68%, 90%, and 95%. The only test that did not provide confidence intervals, the ITPA-3, provided information based on 1 SD across ages that would allow for the calculation of SEM at different confidence intervals.

Most reliability studies appear to have utilized the data from the normative sample. Table 7 reports findings for test reliability/precision coefficients. Test-retest reliability coefficients reported for composites was above .80 for all tests, considered acceptable, while .90 would have been considered good. Only three tests reported test-retest reliability coefficients above .80 for subtests. Most tests reported test-retest reliability coefficients by age and subtest. Those with subtests having unacceptable reliability coefficients (< .80) are reported in Appendix C. The DELV-NR reported the largest number of age groups and subtests with unacceptable test-retest reliability coefficients.

Internal consistency reliability coefficients paralleled test-retest, in that composites were all reported above .80. Internal consistency reliability was more commonly reported by age and subtest than test-retest. Five tests reported all ages and subtests with an internal consistency coefficient above .80. Of the tests with poor internal consistency reliability for subtests, the DELV-NR and the RESCA-E reported the largest amount of ages and subtests with poor test-retest reliability.

Inter-examiner reliability was above .80 for all subtests or composites as reported in the manuals, except the TILLS which had reliabilities ranging from .77 to .99. It should be noted that most of the TILLS' subtests and story versions were above .90, with only one below .80 (story

C, word score was .77). Interestingly, the TOLD-I:4 based their reported inter-examiner reliability (ranging from .90 to .99) on a sample of ages 12 – 17 (the test covers ages 8-16;11), 10 males and 40 females, which are from only the south or mid-west regions. Although these participants were selected from the normative sample, their sampling method is unclear. It is possible, as well that there may be less variability in older children than in younger children, perhaps artificially inflating this measure of reliability.

Table 6
Test Reliability/Precision Findings

Test Acronym	Clear Instructions Provided	Adequate Scoring Parameters/ Examples Provided	Test-retest & Internal Consistency Provided by Age	Standard Error of Measure[e]		Confidence Intervals
				Composites	Subtests	
ALL	Yes	Yes		1 to 2	3 to 4	90% & 95%
CASL-2	Yes	Yes	Yes [c]	1 to 3	1 to 6	90% & 95%
CELF-5	Yes	Yes	Yes	3	1	68%, 90%, & 95%
DELV-NR	Yes	Yes	Yes	5	1	90% & 95%
ITPA-3	Yes	No	Yes	2 to 5	1	NR**
OWLS II	Yes	Yes	Yes [c]	2 to 3	2 to 4	90% & 95%
RESCA-E	Yes	Yes	Yes [c]	< 1 to 6	1 to 2	90% & 95%
TACL-4	Yes	Yes	Yes	2 to 3	1	68%[d]
TEXL	Yes [a]	Yes [a]	Yes	2 to 3	1	68%[d]
TELD-4	Yes	Yes	Yes	2 to 4	3	90% & 95%
TILLS	Yes	Yes	No	N/A	1 to 4	68% & 90%
TOLD-I:4	Yes	Yes	Yes	2 to 3	1	68%[d]
TOLD-P:5	Yes	Yes [a]	Yes	3 to 5	1	90% & 95%

Note. * Determined with interclass correlation. ** Based on 1 SD across ages; confidence intervals & SEM can be calculated. NR = Not Reported. N-R: = Norm-referenced. C-R: = Criterion-referenced. N/A = Not Applicable.
[a] Exception: somewhat limited guidance for one or two subtests. [b] limited or combined age groups. [c] Reported for internal consistency only. [d] Provided formulas to calculate the true score range for 95% and 99% probability. [e] Due to variability in how standard error of measure (SEM) is presented between tests, SEM is reported here as an approximate, values are rounded to the nearest whole number.

Table 7
Test Reliability/Precision Findings; Types of Reliability

Test Acronym	Test-retest reliability Coefficient Range		Internal-consistency Reliability Coefficient Range		Inter-Examiner Reliability Coefficient Range
	Composites	Subtests	Composites	Subtests	
ALL	.87-.96	.75-.93	. 92-.96	.72-.93	.97-.99
CASL-2	.88-.96	.73-.94	.95-.99	.85-.99	.86-.97
CELF-5	.83-.90 [a]	.56-.93 [a]	.92-.97	.60-.99	.91-.99
DELV-NR	.87-.90	.71-.89	.81-.92	.59-.96	.92-1.00
ITPA-3	.90-.99	.86-.99	.87-.99	.75-.96	.95-.99
OWLS II	.89-.90	.73-.91	.96-.99	.92-.98	.93-.96
RESCA-E	.93-.99	.63-.95	.85-.96	.61-.91	.83-.97
TACL-4	.81-.89	.66-.93	.94-.99	.91-.98	.99
TEXL	.88-.90	.72-.90	.96-.99	.87-.98	.99
TELD-4	.81-.92 [a]	.80-.90 [a]	.97-.99	.93-.99	.99
TILLS	N/A	.71-.99*	N/A	.97-.99**	.77-.99
TOLD-I:4	.80-.98 [a]	.80-.96 [a]	.92-.99	.85-.98	.90-.99***
TOLD-P:5	.86-.97 [a]	.75-.99	.87-.98	.77-.97	.97-.99

Note. * Determined with interclass correlation. ** Coefficient Omega *** Based on sample with limitations in ages, gender, & regions. NR = Not Reported. N/A = Not Applicable. [a] Limited or combined age groups.

Fairness

Of the thirteen tests reviewed, six tests; the ALL, CASL-2, CELF-5, DELV-NR, OWLS II, and TILLS provided some special instructions for administration to children of various linguistic or cultural backgrounds. These included such things as alternative correct responses for various dialects, some accommodations, and cultural considerations for aspects of pragmatics. The DELV provided scoring adjustments according to parent education level, unique among

what the other tests provided. The seven other tests provided no instructions pertaining to any variation of linguistic or cultural background.

Even fewer tests, only three, provided special instructions for administration to children with various cognitive, sensory, or other developmental disabilities. These included primarily accommodations such as extra practice with trial items, modified start points, or breaks in testing. Of these three tests, the TILLS special instructions, when followed, allow for use of normative data. For the CELF-5, some of the instructions allow for use of normative data, but not all. However, the ALL's instructions for these children do not allow for use of normative data.

Discussion

In conducting this review, it has become increasing clear that the development of a norm-referenced language task is a challenging undertaking. One goal of this review is to support test developers, as they continue to develop and improve norm-referenced tests. However, because these tests are used to make very important decisions about individuals, careful evaluation, recognition, and consideration of potential psychometric issues is necessary for improvement. This review led to several observations that are worth discussion. There were points of curiosity that led the reviewers to reexamine some established ideas. Specific to norming procedures, in looking for the requisite minimum normative sample of 100, a question arose regarding the origin of this number. Research into this indicated that it may in fact not be necessary and that a minimum of 50 may be acceptable, although more would still be better. In examining the demographic information for normative samples to determine full-range or truncated sampling, the impact of certain inclusions became a concern. The CELF-5 is the only test that reported

including children who had languages other than English spoken in their homes, or who were reportedly bilingual.

Table 8
Test Fairness Findings

Test Acronym	What Special Instructions for Various Linguistic or Cultural Backgrounds were Provided?	What Special Instructions for Various Cognitive, Sensory or other Developmental Disabilities were Provided
ALL	Dialectal variations for speakers of AAE[a]	Modifications for non-standard administrations:[b] Rewording test questions, continuing to test beyond the ceiling, asking a child to explain incorrect responses, and/or using alternative scoring procedure.
CASL-2	Provides alternative correct responses and scoring for speakers of AAE or a similar dialect (such as Southern English) for expressive tests.[a] Not to be given to those judged to not have sufficient English (based on examiner judgement).	None
CELF-5	Extra time for responses, increasing number of trial items, continue testing past ceiling without points, supplement results with language sample, observations, interviews and/or dynamic assessment.[a] Provides examples of dialectal variation, dialectal patterns, and common contrasts between dialects and MAE. For the Word Structure Subtest, there are alternate responses provided for speakers of AAE, SE, S-IE, A-IE and C-IE. Additional instructions considering cultural background on Pragmatics Profile.	Special testing considerations include: motor, sensory, or cognitive impairments. Extra time for responses, increasing number of trial items, continue testing past ceiling without points, supplement results with language sample, observations, interviews and/or dynamic assessment.[a]
DELV-NR	Scoring adjustment for parent education level.[a]	None
ITPA-3	None	None
OWLS II	Alternative correct responses are provided for speakers of AAE or similar dialects.[a]	None
RESCA-E	None	None
TACL-4	None	None
TEXL	None	None
TELD-4	None	None
TILLS	Adjusted cut scores which should be used for identifying language impairment in students from low SES backgrounds[d]	Considerations for three previously identified special populations (ASD, deaf or HH, or ID). Specifies functioning at no less than 6 years of age, appropriate hearing technology, and language learning primarily through auditory-oral means. Modified start and stop rules, steps to ensure maximum auditory access, and breaks or stopping testing.[a]
TOLD-I:4	None	None
TOLD-P:5	None	None

Note: * May not be appropriate for non-MAE speakers, given the stated purpose of testing standard American English-speaking children. MAE = mainstream American English. AAE = African American English. SE = Southern English. S-IE = Spanish-Influenced English. C-IE = Chinese-Influenced English. A-IE = Asian-Influenced English. SES - Socioeconomic status. ASD = Autism Spectrum Disorder. HH = Hard of Hearing. ID = Intellectual Disability.
[a] Allows for comparison to normative data. [b] Does not allow for comparison to normative data. [c] Some modifications allow for comparison to normative data. [d] Alternative basis of comparison regarding normative data based on subgroup of normative sample

Including children who may be English language learners (ELL) in the normative sample, even if they are otherwise typically developing could be artificially lowering the normal mean. Because children who do not speak the English language at native-like proficiency may make errors similar to children with language impairment, their inclusion would contribute to normalizing these errors in the normative sample. Some tests may simply include children who are ELL as a bi-product of simply including a nationally representative sample of race/ethnicity. Because these children may be included in the normative sample for some tests but not others, it may be an area to consider when evaluating language tests. There is a distinct possibility that including children who are ELL in the normative sample may contribute to misdiagnosis by increasing the number of false negatives of test results. Although representative diversity is important and desirable, this has to be attained cautiously in order to avoid unintended consequences. Future test development should probably include some benchmark for English language proficiency before including English language learners in the normative sample and also some clearer definition of bilingual children versus English language learners.

Evidence of validity was the most difficult to evaluate due to several factors, including the variability of measures used, how values were reported, and ambiguity in terminology across tests. Overall, many tests' validity studies appeared to lack quality either in methodology, sampling, or reporting of findings. Another aspect of validity that needs attention by test developers in the future is the rationale behind composite scores. While a majority of tests provide a basis for arriving at composite scores, few provide a rationale for their use. Some composites are supported by factor loadings, but many tests manuals only state that composites are the most reliable scores. However, given that reliability increases with the number of items on a test, it makes sense that the composites would demonstrate greater reliability because they

incorporate multiple subtests into one score. Therefore, a higher reported reliability is not an adequate rationale or justification for using a composite score as the basis for determining a child's language ability.

There has been a marked improvement in reporting diagnostic accuracy by test developers. Spaulding et al. (2006) found only 9 out of 43 tests provided sensitivity and specificity data, and only 5 of those 9 reported sensitivity and specificity at or above .80. The current review found that all 13 tests reported either sensitivity or specificity, if not both, at .80 or above. However, this aspect of psychometric quality still needs a considerable improvement. Most of the tests had their highest levels of sensitivity and specificity at SSs and SDs above the typical requirements for eligibility for special services in most educational settings. The issue that arises as a result of so many tests having their highest levels of sensitivity and specificity at cutoff scores higher than many state eligibility requirements for therapy services needs to be addressed. This is, however, beyond the scope of the current review.

Another aspect of diagnostic accuracy that warrants close scrutiny is the classification criteria used to determine the status of individuals included in studies to determine diagnostic accuracy. For many tests included in this review, the reference for determining impairment boiled down to performance on another test. For some of the tests in this review those reference tests were published over twenty years ago such as the Test of Language Development – Intermediate: Third Edition (TOLD-I:3; Hammill & Newcomer, 1997), Preschool Language Scale – Third Edition (PLS-3; Zimmerman et al. 1992), Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1995), or Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 1999). Other reference tests, such as the Diagnostic Achievement Battery - Fourth Edition (DAB-4; Newcomer, 2014), Pragmatic Language Observations Scale (PLOS;

Newcomer & Hammill, 2009), Peabody Picture Vocabulary Test- Third Edition (PPVT-3; Dunn & Dunn, 1997), Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV; Wechsler, 2003), or Young Children's Achievement Test, Second Edition (YCAT-2; Hresko et al., 2018) either test only one domain of language, are tests of intelligence, or test of achievement. This does not appear to support a diagnosis of language impairment upon which to base a study of diagnostic accuracy. The basis of comparison, when it is another test is suspect. As a last example, the OWLS II is also used as a reference standard for other test's diagnostic accuracy evidence. Current findings for the OWLS II suggest it does not provide evidence of diagnostic accuracy (group differences are not sufficient), so it is not clear how it could be considered an acceptable reference for identification of language impairment.

Clearly, if children participating in diagnostic accuracy studies are incorrectly classified as either language impaired or not, or are classified by varying criteria, then the validity of the results would be negatively affected. If results based on inaccurate classification or widely diverse bases of classification are used as evidence of diagnostic accuracy, then the diagnostic accuracy estimate would be questionable. As decisions are made for children based on using tests with questionable diagnostic accuracy, the likelihood of misdiagnosis increases. It follows then that those misdiagnosed children will then be included in future samples classified as language impaired in future studies. Thereby continuing to muddy the waters for clinicians, rather than enhancing the accuracy of their diagnostic decisions. Therefore, the classification criteria currently employed by test developers in studies of diagnostic accuracy warrants careful scrutiny by potential test users. It should also serve as an impetus for establishing clearer reference standards that are not based only on test scores. It should not be acceptable to adopt the classification of participants as either impaired or unimpaired without independent verification

such as that employed by Tomblin et al. (1997), Merrell and Plante (1997), Greenslade et al. (2009), or Pearson et al. (2014). These researchers confirmed the diagnosis of participants with measures independent of the original diagnosis. Ideally, Dollaghan (2007) states that all participants should receive both the measure under examination and the reference standard. This would serve to confirm the status of both impaired and controls and allow for more meaningful comparisons.

One final observation about diagnostic accuracy is concerned with AUC. That some tests were reporting AUC values as evidence of diagnostic accuracy led to questioning its practical utility as a measure of diagnostic accuracy. Clinicians need to be aware that while AUC may be a quick single number from which to gauge overall diagnostic accuracy, it does not indicate when either sensitivity or specificity is greater. Therefore, it should only be used as a means of ruling in or ruling out, and not as the sole basis of choosing which test to use in their particular setting or situation.

Reliability is an area of strength, across all thirteen tests, out of the five aspects of psychometric quality reviewed. Each test, with very few exceptions provided adequate instruction, examples, and parameters to ensure standardization in test administration. Adequate information regarding SEMs and CIs was also provided to ensure accurate estimates of where an individual's true score would lie. Reliability coefficients, while acceptable for all tests for composites, were the main area still in need of improvement for the subtests. Several tests reported reliability coefficients below .80 on subtests for various ages as reported in Appendix C. Again, this serves as a caution. Clinicians should examine test reliability beyond the composite scores.

Fairness, while new in name, is not new to the Standards. The need for tests to address appropriate administration for non-native English speakers or individuals with various handicapping conditions has been part of the Standards for decades (AERA et al., 1985). This review found that less than half the tests addressed fair administration for those of various linguistic or cultural backgrounds, and only three did so for children with various cognitive, sensory, or other developmental disabilities. These factors should always be considered when deciding which test to use for an individual. It is encouraging though, to find that some tests are incorporating components such as alternative correct responses for various dialects, accommodations like extra practice with trial items, modified start points, or breaks in testing, and cultural considerations for aspects of pragmatics into test development that allow for use of normative data.

Conclusion

There appears to have been a great deal of improvement in the psychometric quality of language tests over since McCauley and Swisher (1984). Tests with a purpose of diagnosis are providing evidence of diagnostic accuracy. Reliability coefficients are improving, with most tests reporting at lease acceptable reliability and several reporting good reliability. There is still room for growth, improvement and further research. While normative samples are reported in manuals as meeting the historical minimum of 100, when the actual number of individuals in the scoring cells is calculated, there are still tests that do not even meet our new bare minimum of 50. The impact of inclusion of ELL in normative samples of tests for English is a question that warrants research. The reference for language impairment in studies of diagnostic accuracy is in desperate need of critical evaluation. Validity studies need to be conducted according to standards expected in any quality research, with clinical and control groups, adequate sampling methods, consistent

terminology and clear reporting. And while tests are generally attempting to eliminate bias based on demographic differences such as race/ethnicity or SES, many tests are simply not appropriate for children of diverse cultural backgrounds or children with a range of existing disabilities. In test development, studies independent of the norming process that investigate aspects of validity, diagnostic accuracy, reliability/precision, and fairness would be valuable contributions to the future of language assessment.

The goal of this review is to contribute to the future development of language tests, and even more, that this review provides a resource that practicing clinicians can use to help guide their assessment decisions based on psychometric evidence. Clinicians should use the information provided here on the various aspects of psychometric properties of particular tests as a starting point for their own reviews of tests and to narrow down their test options. The tables herein should be used to compare the various reported measures for each test, according to each aspect of psychometric quality, and provide a start point for narrowing choices. The appendices provide additional comparisons relevant to specific areas and should be referred to for the Standards this review is based on, the targeted constructs each test claims to assess, and the ages or grades for which certain subtests have unacceptable levels of reliability. In using the information presented here, the clinician can certainly narrow the choice of test. However, the responsibility still lies with the clinician to evaluate the appropriateness of that test to their situation. Even though this is intended to be a valuable resource, it is vital that clinicians remember that they need to understand the various concepts important to psychometric quality in order to make informed decisions about the populations and individuals with whom they work.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Speech-Language-Hearing Association, (2016). *Scope of practice in speech-language pathology* [Scope of Practice]. Retrieved April 6, 2020 from www.asha.org/policy/.

Anastasi, A., & Urbina, S., (1997). *Psychological Testing* (7th Edn.). Upper Saddle River, NJ: Pearson Education/Prentice Hall.

Bankson, N. W., Mentis, M. M., & Jagiellko, J. R. (2018) *Bankson Expressive Language* Test (3rd Edn.). Austin, TX: PRO-ED.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44*(2), 133-146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00149

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology*, *13*(6), 528–538. https://doi.org/10.1080/09297040701233875

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment. *Journal of Speech, Language, and Hearing Research*, *40*(3), 519–525. https://doi.org/10.1044/jslhr.4003.519

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales.* Circle Pines, MN: AGS Publishing.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language.* MN: American Guidance Service.

Carrow-Woolfolk, E. (2011). *Oral and Written Language Scales, 2nd Edn.* Minneapolis, MN: Pearson Psychcorp.

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

Carrow-Woolfolk, E. (2017). *Comprehensive Assessment of Spoken Language- 2nd Edn*. Torrance, CA: WPS.

Carrow-Woolfolk, E. & Allen, E. A. (2014) *Test of Expressive Language.* Austin, TX: Pro-Ed.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1076/jcen.21.4.559.889

Charter, R. A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology*, *130*(2), 117–129. https://doi.org/10.1080/00221300309601280

Charter, R. A., & Feldt, L. S. (2002). The importance of reliability as it relates to true score confidence intervals. *Measurement and Evaluation in Counseling and Development*, *35*(2), 104–112. https://doi.org/:10.1080/07481756.2002.12069053

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394–409. https://doi.org/:10.1037/0022-0663.98.2.394

Crawford, J., & Howell, D. C. (1998). Comparing an individuals' test score against norms derived from small samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. https://doi.org/10.1076/clin.12.4.482.7241

Dawson, J., Eyer, J. A., & Fonkalsrud, J. (2005). *Structured Photographic Expressive Language Test—Preschool* (2nd Edn.). DeKalb, IL: Janelle Publications.

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *Structured Photographic Expressive Language Test* (3rd Edn.). DeKalb, IL: Janelle Publications.

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y., & Cordier, R. (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. *Frontiers in Psychology, 8*, 1-28. https://doi.org/10.3389/fpsyg.2017.01515

Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Real-word and nonword repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, *56*(1), 323–336. https://doi.org/10.1044/1092-4388(2012/11-0304)

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders*, *37*(5), 391–400. https://doi.org/10.1016/j.jcomdis.2004.04.002

Dollaghan, C. A. (2007). *The Handbook for Evidence-based Practice in Communication Disorders*. Baltimore: Paul H. Brookes Pub.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

Dunn, L. M., & Dunn, D. M. (1997). *The Peabody Picture Vocabulary Test* (3rd *Ed.). Bloomington, MN: NCS Pearson, Inc.*

Dunn, L. M., & Dunn, D. M. (2007). *The Peabody Picture Vocabulary Test (4th Ed.). Bloomington, MN: NCS Pearson, Inc.*

Flipsen, P., & Ogiela, D. A. (2015). Psychometric characteristics of single-word tests of children's speech sound production. *Language, Speech, and Hearing Services in Schools, 46*(2), 166-178. https://doi.org/10.1044/2015_lshss-14-0055

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

Gray, S. (2003). Diagnostic accuracy and test–retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders*, *36*(2), 129–151. https://doi.org/10.1016/s0021-9924(03)00003-0

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. https://doi.org/10.1044/0161-1461.3002.196

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

Hammill, D. D., Mather, R., & Roberts, R. (2001). Illinois Test of Psycholinguistic Abilities, (3rd Edn.). Austin, TX: Pro-Ed.

Hammill, D. D., & Newcomer, P. L. (2008). *Test of language development – Intermediate* (4th Edn.), Austin, TX: Pro-Ed.

Hammill, D. D., & Newcomer, P. L. (1997). *Test of Language Development – Intermediate* (3rd Edn.). Austin, TX: PRO-ED.

Hamaguchi, P. & Ross-Swain, D., (2015) *Receptive, Expressive & Social Communication Assessment–Elementary.* Torrance, CA: WPS.

Hoffman, L. M., Loeb, D. F., Brandel, J., & Gillam, R. B. (2011). Concurrent and construct validity of oral language measures with school-age children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *54*(6), 1597–1608. https://doi.org/10.1044/1092-4388(2011/10-0213)

Hresko, W. P., Peak, P. K., Herron, S. R., & Hicks, D. L. (2018). *Young Children's Achievement Test* (2nd Edn.). Austin, TX: PRO-ED.

Hresko, W. P., Reid, K., & Hammill, D. D., (2017) *Test of Early Language Development, 4th Edn*. Austin, TX: Pro

Ireland, M., & Conrad, B. J. (2016). Evaluation and eligibility for speech-language services in schools. *Perspectives of the ASHA Special Interest Groups, 1*(Part 4), SIG 16, 78-90.

Ivnik, R. J., Smith, G. E., Petersen, R. C., Boeve, B. F., Kokmen, E., & Tangalos, E. G. (2000). Diagnostic accuracy of four approaches to interpreting neuropsychological test data. *Neuropsychology*, *14*(2), 163–177. https://doi.org/10.1037/0894-4105.14.2.163

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice*, *24*(4), 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Leeflang, M. M. G., Rutjes, A. W. S., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*, *185*(11), E537–E544. https://doi.org/10.1503/cmaj.121286

Lombardino, L. J., Leiberman, R., & Brown, J. C. (2005*). Assessment of Literacy and Language.* San Antonio, TX: Pearson Psychcorp.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*(1), 34-42. https://doi.org/10.1044/jshd.4901.34

McFadden, T. U. (1996). Creating language impairments in typically achieving children. *Language, Speech, and Hearing Services in Schools*, *27*(1), 3–9. https://doi.org/10.1044/0161-1461.2701.03

Mendoza, J. L., Stafford, K. L., & Stauffer, J. M. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Psychological Methods*, *5*(3), 356–369. https://doi.org/10.1037/1082-989x.5.3.356

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. https://doi.org/10.1044/0161-1461.2801.50

Mitrushina, M. N. (2005). *Handbook of normative data for neuropsychological assessment* (2nd Edn.). New York: Oxford University Press.

Montzka, J. L. (2015). *Factors influencing standardized test selection for children presenting with language difficulties* (Unpublished master's thesis). Idaho State University, Meridian, Idaho.

Nelson, N., Plante, E., Helm-Estabrooks, N., & Hotz, G. (2016). *Test of Integrated Language & Literacy Skills*. Baltimore, MD: Paul H. Brookes Publishing Co.

Newcomer, P. L. (2014). *Diagnostic Achievement Battery* (4th Edn.). Austin, TX: PRO-ED.

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development – Primary (3rd Edn.).* Austin, TX: PRO-ED.

Newcomer, P. L., & Hammill, D. D. (2008*). Test of language development - Primary, 4th Edn*. Austin, TX: PRO-ED.

Newcomer, P. L., & Hammill, D. D. (2009) *Pragmatic Language Observations Scale*. Austin, TX: PRO-ED.

Newcomer, P. L., & Hammill, D. D. (2019). *Test of Language Development - Primary, 5th Edn.* Torrance, CA: WPS.

Paul, R., Norbury, C., & Gosse, C. (2018). Assessing students' language for learning. In *Language disorders from infancy through adolescence: Assessment and intervention* (5th ed.) (pp. 440-483). St. Louis, MO: Mosby Elsevier.

Paraskevopoulos, J. N., & Kirk, S.A. (1969). *The development and psychometric characteristics of the Revised Illinois Test of Psycholinguistic Abilities* Urbana: University of Illinois Press

Pearson, B. Z., Jackson, J. E., & Wu, H. (2014). Seeking a valid gold standard for an innovative, dialect-neutral language test. *Journal of Speech, Language, and Hearing Research*, *57*(2), 495–508. https://doi.org/10.1044/2013_jslhr-l-12-0126

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. https://doi.org/10.1044/1058-0360(2006/023)

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

Plante, E., & Vance, R. (1994). Selection of preschool language tests. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Plante, E., & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology*, *4*(2), 70–76. https://doi.org/10.1044/1058-0360.0402.70

Salvia, J., & Ysseldyke, J. E. (1981). *Assessment in special and remedial education* (2nd Edn.). Boston: Houghton Mifflin.

Salvia J, Ysseldyke J. E., & Bolt S. *Assessment: In Special and Inclusive Education* (11 Edn). Boston, MA: Wadsworth/Cengage Publications, 2010.

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

Semel, E. M., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals* (4th Edn.). Toronto, Canada: PsychCorp.

Semel, E. M., Wiig, E. H., & Secord, W. A. (2006) *Clinical Evaluation of Language Fundamentals, Preschool* (2nd Edn.). San Antonio, TX: Pearson Education

Seymour, H. N., Roeper, T. W., de Villiers, J., & de Villiers, P. A. (2005*). Diagnostic Evaluation of Language Variation*. Minneapolis, MN: Pearson Psychcorp.

Šimundić A. M. (2009). Measures of diagnostic accuracy: Basic definitions. *EJIFCC*, *19*(4), 203–211. Published online on PMC US National Library of Medicine National Institutes of Health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975285/

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment. *Language, Speech, and Hearing Services in Schools, 37*(1), 61-72. https://doi.org/10.1044/0161-1461(2006/007)

Spaulding, T. J., Szulga, M. S., & Figueroa, C. (2012). 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, 43*(2), 176-190. https://doi.org/10.1044/0161-1461(2011/10-0103)

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tang, W., Cui, Y., & Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it? *Journal of Psychology and Behavioral Science*, *2*(2), 205–220. Retrieved from https://jpbsnet.com/journals/jpbs/Vol_2_No_2_June_2014/13.pdf

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1284–1294. https://doi.org/ 10.1044/jshr.3906.1284

U.S. Department of Education. (2006). Sec. 300.304 (b). (n.d.). Retrieved May 20, 2019, from https://sites.ed.gov/idea/regs/b/d/300.304/b

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th Edn.). San Antonio, TX: Psychological Corp.

Weiner, P. S., & Hoock, W. C. (1973). The standardization of tests: Criteria and criticisms. *Journal of Speech and Hearing Research*, *16*(4), 616–626. https://doi.org/10.1044/jshr.1604.616

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals, 5th Edn*. Bloomington, MN: Pearson Psychcorp.

Williams, K. T. (2007). *Expressive Vocabulary Test* (2nd Edn.). *Bloomington, MN: NCS Pearson, Inc.*

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale* (3rd Edn.). San Antonio, TX: Psychological Corporation

Zimmerman, I., Steiner, V., & Pond, R. (2011). *Preschool Language Scale (5th Edn.)*. San Antonio, TX: Pearson Assessments.

Appendices

Appendix A
Standards for Educational and Psychological Testing (AERA et al., 2014)

	Validity
Standard 1.0 "Clear articulation of each intended test score interpretation for specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA et al., 2014, p. 23).	**Standard 1.1** "The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly" (AERA et al., 2014, p. 23). **Standard 1.2** "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation" (AERA et al., 2014, p. 23). **Standard 1.8** "The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics" (AERA et al., 2014, p. 25). **Standard 1.11** "A) Content-Oriented Evidence When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria should also be clearly explained and justified" (AERA et al., 2014, p. 26). **Standard 1.13** "C) Evidence Regarding Internal Structure If the rationale for a test score interpretation for a given use depends on premises about the relationships among the test items or among parts of the test, evidence concerning the internal structure of the test should be provide" (AERA et al., 2014, pp. 26-27). **Standard 1.14** "When interpretation of sub-scores, score differences or profiles is suggested, the rational and relevant evidence in support of such interpretations should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given" (AERA et al., 2014, p. 27). **Standard 1.16** "When validity evidence includes empirical analyses of responses to test items together with data on other variable, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent."(AERA et al., 2014, p.27)
	Reliability/Precision
Standard 2.0 "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (AERA et al., 2014, p. 42).	**Standard 2.1** "The range of replications over which reliability/precision is being evaluated should be clearly stated along with a rationale for the choice of this definition, given the testing situation" (AERA et al., 2014, p. 42). **Standard 2.2** "The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores" (AERA et al., 2014, pp. 42-43).

	Standard 2.12 "If a test is proposed for use in several grades or over a range of ages, and if separate norms are provided for each grade or each age range, reliability/precision data should be provided for each age or grade-level subgroup, not just for all grades or ages" (AERA et al., 2014, p. 45). **Standard 2.13** "The standard error of measurement, both overall and conditional (if reported), should be provided in unites of each reported score" (AERA et al., 2014, pp. 45-46). **Standard 2.14** "When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score" (AERA et al., 2014, p. 46).
	Fairness
Standard 3.0 All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population." (AERA et al., 2014, p. 63)	**Standard 3.2** "Test developers are responsible for developing test that measure the intended construct and for minimizing the potential for test' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics" (AERA et al., 2014, p. 64). **Standard 3.3** "Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test" (AERA et al., 2014, p. 64).
	Norms
Standard 5.0 Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use (AERA et al., 2014, p. 102).	**Standard 5.8** "Norms, if used, should refer to clearly described populations. These populations should include individuals or groups with whom tests users will ordinarily wish to compare their own examinees" (AERA et al., 2014, p. 104).
	Diagnostic Accuracy
Definitions in the glossary of *Standards for Educational and Psychological Testing* by (AERA et al., 2014, p. 223)	*"Sensitivity* - in classification, diagnosis, and selection, the proportion of cases that are assessed as meeting or predicted to meet criteria and which, in truth, do meet the criteria" (AERA et al., 2014, p. 223). *"Specificity* - in classification, diagnosis, and selection, the proportion of cases that are assessed as not meeting or predicted to not meet criteria and which, in truth, do not meet the criteria" (AERA et al., 2014, p. 223).

Appendix B

Tests, Subtests, and Targeted Constructs

Test Acronym	Subtests		Targeted Constructs
ALL	Letter Knowledge Rhyme Knowledge Basic Concepts Receptive Vocabulary Parallel Sentence Production Elision, Word Relationships Phonics Knowledge Sound Categorization	Sight Word Recognition Listening Comprehension Book Handling Concept of Word Matching Symbols Word Retrieval Rapid Automatic Naming Invented Spelling	Language (semantics, morphology, and syntax), phonological awareness, print awareness, alphabet knowledge, fluency, and comprehension.
CASL-2	Receptive Vocabulary Antonyms, Synonyms Expressive Vocabulary Idiomatic Language Sentence Expression Grammatical Morphemes Sentence Comprehension	Grammaticality Judgment Nonliteral Language Meaning from Context Inference Double Meaning Pragmatic Language	Language knowledge (lexical/semantic, syntactic, supralinguistic and pragmatic) and Language performance (auditory comprehension and oral expression).
CELF-5	Observation Rating Scale Sentence Comprehension Linguistic Concepts Word Structure Word Classes Following Directions Formulated Sentences Recalling Sentences	Understanding Spoken Paragraphs Word Definitions Sentence Assembly Semantic Relationships Reading Comprehension Structured Writing Pragmatics Profile Pragmatics Activities Checklist	Semantics, morphology and syntax, pragmatics, reading and writing.
DELV-NR	Wh-question Items Passive Items Article Items Communicative Role-Taking Items Narrative Items Question Asking Items Verb and Preposition Contrast Items	Quantifiers Items Fast Mapping: Real Verbs Fast Mapping: Novel Verbs Number of Syllables Number of Consonants in Cluster Medial Cluster Context Manner	Syntax (Wh-questions, passives and articles), pragmatics (communicative role-taking, short narrative and Questions Asking), semantics (verb contrast, preposition contrast, quantifiers, and fast mapping) and phonology (formation of consonant clusters in the initial and medial position of words produced in sentence contexts).
ITPA-3	Spoken Analogies Spoken Vocabulary Morphological Closure Syntactic Sentences Sound Deletion Rhyming Sequences	Sentence Sequencing Written Vocabulary Sight Decoding Sound Decoding Sight Spelling Sound Spelling	Global constructs (general Language, spoken Language, and written language) Specific constructs (spoken language [semantics, grammar, phonology] and written language (comprehension, word identification, spelling, sight-symbol processing, and sound-symbol processing).
OWLS II	Listening Comprehension Oral Expression		Lexical/semantics, syntax, supralinguistic, and pragmatics.

RESCA-E	Comprehension of Vocabulary Comprehension of Oral Directions Comprehension of Stories and Questions Comprehension of Basic Morphology and Syntax Executing Oral Directions Expressive Labeling of Vocabulary Expressive Skills for Describing and Explaining Narrative Skills Expressive Use of Basic Morphology and Syntax Comprehension of Body Language and Vocal Emotion Social and Language Inference Situational Language Use Elicited Body Language Social Communication Inventory		Receptive language (word, sentence, narrative levels): vocabulary, concepts, comprehension, recall, inference, morphology/syntax, execution of oral directions (fine motor, gross motor, or verbal). Expressive language (word, sentence, narrative levels): vocabulary, descriptive language skills, narrative language skills, morphology/syntax. Social Communication (language, social cognition, and social behavior): inferring emotion from facial expression, body language and tone of voice, language inferencing (idioms/slang), situational inferencing (perspective taking, visual comprehension, language comprehension, and social cognitive skills), use of expressive language in specific social contexts (problem-solving, perspective-taking, and expressive language), and outward portrayal of emotion.
TACL-4	Vocabulary Grammatical Morphemes Elaborated Phrases and Sentences		Receptive: vocabulary (semantics), morphology, and syntax.
TEXL	Vocabulary Grammatical Morphemes Elaborated Phrases and Sentences		Expressive: vocabulary (semantics), morphology, and syntax.
TELD-4	Receptive Language Expressive Language		Receptive and Expressive: semantics and syntax/morphology
TILLS	Vocabulary Awareness Phonemic Awareness Story Retelling Nonword Repetition Nonword Spelling Listening Comprehension Reading Comprehension Following Directions Delayed Story Retelling Nonword Reading Reading Fluency Written Expression Social Communication Digit Span Forward Digit Span Backward		Lexical knowledge, awareness of semantic relationships, cognitive-linguistic flexibility, phoneme awareness, ability to listen to, comprehend, and retell a story, speech perception, immediate memory, ability to reproduce phonological sequences accurately, use conventional orthographic patterns to represent phonemic and morphemic components of novel words, comprehension of complex syntax, and inferencing, listen to, understand, hold in short-term memory, and execute directions, retention of narrative information over 20-30 minutes, decoding non-words, automatic word recognition, written expression, pragmatic ability to formulate responses appropriate to social contexts, short-term and verbal working memory, and working memory.
TOLD-I:4	Sentence Combining Picture Vocabulary Word Ordering	Related Vocabulary Morphological Comprehension Multiple Meanings	Semantics, grammar (syntax/morphology) and listening, organizing and speaking (expressive and receptive).
TOLD-P:5	Picture Vocabulary Relational Vocabulary Oral Vocabulary Syntactic Understanding Sentence Imitation	Morphological Completion Word Discrimination Phonemic Analysis Word Articulation	Semantics, syntax/morphology, and phonology.

Appendix C

Subtests with Reliability Coefficients < .80

Test Acronym	Test-retest reliability Coefficient Ages or Grade (Subtest) with Coefficients < .80	Internal-consistency Reliability Coefficient Ages or Grade (Subtest) with Coefficients < .80
ALL	PK (Rec. Vocab) K (Elision) 1st (Rhyme Knowledge, Basic Concepts, Rec. Vocab, Sound Categorization, Listening Comp.)	PK (N-R: Rec. Vocab C-R: Book Handling, Concept of Word, Matching Symbols) K (N-R: Basic Concepts, Rec. Vocab C-R: Book Handling) 1st (N-R: Basic Concepts, Rec. Vocab, Parallel Sentence Production, Elision C-R: Book Handling, Concept of Word)
CASL-2	NR[a]	None
CELF-5	8, 10, 11, 12, & 14 (Structured Writing) 15 (Formulated Sentences)	7, 8, 10, 11, 12, & 14 (Structured Writing) 15 (Formulated Sentences) 7 (Sentence Comp.)
DELV-NR	4-4;11 (pragmatics) 5-5;11 (syntax, pragmatics) 6-6;11 (syntax) 7-7;11 (syntax, pragmatics, semantics) 8-8;11 (pragmatics) 9-9;11 (syntax, pragmatics) all combined (syntax)	4-4;5 (syntax, semantics) 4;6-4;11 (syntax) 5-5;5 (syntax, semantics) 5;6-5;11 (syntax, semantics) 6;6-6;11 (pragmatics, semantics) 7-7;11 (syntax, pragmatics, semantics) 8-8;11 (syntax, pragmatics, semantics, phonology) 9-9;11 (syntax, pragmatics, semantics, phonology) all combined (syntax, pragmatics, semantics)
ITPA-3	None	5, 7, 8 (Rhyming Sequences)
OWLS II	NR[a]	None
RESCA-E	NR[a]	8, 10, & 12 (Comp. of Oral Directions, Comp. of Vocab) 5, 6, & 9 - 12 (Comp. of Stories & Questions) 12 (Comp. of Basic Morphology & Syntax) 5 & 8 - 11 (Exp. Skills for Describing & Explaining) 11 (Exp. Use of Basic Morphology & Syntax) 7 - 12 (Comp. of Body Language & Vocal Emotion) 11 & 12 (Situational Language Use)
TACL-4	6 - 12 (Vocab) 6 - 12 (Elaborated Phrases & Sentences) all combined (Vocab, Grammatical Morphemes, Elaborated Phrases & Sentences)	None
TEXL	6 - 12 (Vocab) 6 - 12 (Elaborated Phrases & Sentences) all combined (Vocab, Grammatical Morphemes, Elaborated Phrases & Sentences)	None
TELD-4	None	None
TILLS	NR	NR
TOLD-I:4	None	None
TOLD-P:5	7 & 8 (Oral Vocab)	7 & 8 (Phonemic Analysis)

Note. NR = Not Reported. N-R: = Norm-referenced. C-R: = Criterion-referenced. N/A = Not Applicable. [a] Reported for internal consistency only. Reliability coefficients were considered excellent at >.90, acceptable between .80 and .90, poor when <.80