2010

Considerations for Test Selection: How Do Validity and Reliability Impact Diagnostic

Jennifer C. Friberg
Illinois State University, jfribe@ilstu.edu

First published in Child Language Teaching and Therapy, vol. 26, no. 1. https://doi.org/10.1177/0265659009349

Considerations for test selection: How do validity and reliability impact diagnostic decisions?

Jennifer C. Friberg
Illinois State University

ABSTRACT: Nine preschool and school-age language assessment tools found to have acceptable levels of identification accuracy were evaluated to determine their overall levels of psychometric validity for use in diagnosing the presence/absence of language impairment. Eleven specific criteria based on those initially devised by McCauley and Swisher (1984) were applied to each of the selected tests in order to determine each test's overall level of psychometric validity. Results indicated that each of the selected assessment tools met at least eight of the 11 criteria used to evaluate each assessment tool. Five tests met 10 out of 11 criteria. Findings are discussed to assist clinicians in applying psychometric criteria to these selected tests, as well as those not reviewed as part of this current review of standardized assessment tools. A decision tree is included within the discussion of this study's findings to aid clinicians in the selection of standardized assessment tools that are most appropriate for clinical use, based on their psychometric characteristics.

I Introduction

Language assessment is a complex endeavor, one that demands that speech–language pathologists (SLPs) synthesize information collected from a variety of sources in order to make clinical deci-sions based on sound evidence. Optimally, SLPs should incorporate data from both quantitative and qualitative sources to fully examine the language abilities of any client undergoing a language-based assessment. Thus, data from case histories, naturalistic environmental observations, and informal assessments (e.g. classroom/teacher checklists, spontaneous language sampling, and/or

criterion-referenced tests) should be combined with more formal, standardized language testing in order to make balanced, well-developed clinical decisions with regard to eligibility for services and overall service planning (Roseberry-McKibbin, 2007). That said, many clinicians seem to accentuate results collected from standardized assessment tools that provide quantitative information to assist in eligibility determination and intervention planning (Roulstone, Peters, Glogowska, and Enderby, 2008). This emphasis can be problematic, as the widespread use of standardized tests assumes that these protocols correctly measure the presence or absence of language impairments. However, this is not always the case, as threats to the psychometric validity of these tools can compromise the accuracy of any decisions made using data collected from these instruments. Thus, there is a real need for clinicians to understand issues related to validity and reliability that accompany the use of standardized assessment tools as part of their diagnostic battery.

1 Psychometric validity and standardized assessment tools

In an effort to examine the psychometric validity of standardized testing instruments commonly utilized by SLPs, McCauley and Swisher (1984) rated 30 standardized preschool language tests based on the presence or absence of 10 psychometric criteria. Results indicated that no test possessed all 10 criteria, and, alarmingly, only 12 of the 30 assessment tools reviewed met three of the original 10 criteria. These findings served as the first indication that standardized tests frequently used by SLPs were not as valid as had been previously assumed.

Using a similar methodology, Plante and Vance (1994) updated McCauley and Swisher's (1984) study to identify changes in test standardization and to better inform test selection. Twenty-one preschool language tests were evaluated using the same criteria used by McCauley and Swisher (1984). Results indicated that only four tests were found to possess six or more (of a possible 10) psychometric criteria, reflecting a modest overall gain in reliability and validity when compared to previous results (McCauley and Swisher, 1984).

2 Identification accuracy

Beyond simply looking at the 10 criteria identified by McCauley and Swisher (1984), several researchers have also looked at the issue of identification accuracy, which refers to an assessment tool's ability to accurately diagnose the presence or absence of a speech and/or language disorder. Researchers have found that good psychometric properties alone are not sufficient to demonstrate identification accuracy (Plante and Vance, 1994, 1995; Gray, Plante, Vance, and Henrichsen, 1999). In fact, Spaulding, Plante, and Farinella (2006) have suggested that the identification accuracy of tests might be a more important indicator of an instrument's validity than other, previously identified psychometric criteria, as this variable indicates the overall precision of diagnosis made by practicing clinicians. Consequently, it has been suggested that it is a 'poor use of time' to assess the overall psychometric validity of assessment tools for which data related to their identification accuracy aren't reported (Spaulding et al., 2006: 70).

Identification accuracy is measured in a multi-step process. Initially, a discriminant analysis is conducted, which evaluates a test's convergent validity to judge its ability to distinguish typical from atypical language functioning (Plante and Vance, 1994). This discriminant analysis occurs through a statistical calculation that juxtaposes different variables, including variance in test scores, typical/atypical language functioning, and overall accuracy of categorization. It is as a function of this discriminant analysis that values representing the sensitivity and specificity of a standardized

test are computed. It is through the data provided with regard to sensitivity and specificity that clinicians can gauge a test's overall identification accuracy.

Sensitivity is considered the likelihood that a child who has previously been diagnosed as language disordered is identified as such when using a particular language assessment tool (Dollaghan, 2004; Spaulding et al., 2006). In contrast, specificity is defined as the possibility that a child who has previously been found to be typically developing is identified as such when tested using a given assessment tool (Dollaghan, 2004; Spaulding et al., 2006). Sensitivity and specificity are measured by percent values, which indicate the overall precision of a particular assessment tool at making an accurate diagnosis. These sensitivity and specificity values occur in a range from 0 to 1.0, with values closest to 1.0 reflecting the most accurate diagnoses. It is suggested that the threshold values for acceptable levels of sensitivity and specificity should be at least .80 or greater, although accuracy levels of .90 or above are considered optimal (Plante and Vance, 1994).

It should be noted that sensitivity and specificity should be calculated using data-based cut-off scores, which are values used to determine the level of performance on a particular test that distinguishes typical from atypical functioning (Plante and Vance, 1994; Spaulding et al., 2006). In order to establish these cut-off scores, a test is administered to two groups of children: those with language impairment and those found to be typically developing. Differences in the scores obtained by these two groups lead to the establishment of data-based cut-off scores for clinical use, which inform clinicians of the score at which a language impairment can be diagnosed. With this in mind, any clinician that uses a test and applies arbitrary cut-off scores (those not specified within a test's examiner's manual) could negatively impact a test's identification accuracy (Plante and Vance, 1994; Spaulding et al., 2006). Thus, all aspects of identification accuracy (sensitivity, specificity, and the use of data-based cut-off scores) are important considerations for clinicians relative to identification accuracy, as misdiagnosis and inaccurate clinical decision-making are real threats in the absence of such understanding.

3 Purpose of study

Spaulding et al. (2006) indicated that it can only be acceptable to critically evaluate the psychometric validity of those tools that have been found to have acceptable levels of identification accuracy. To this end, while over 50 tests are currently available for use as part of a language assessment battery, a smaller cadre of standardized preschool and school-age assessment tools identified as possessing acceptable levels of both sensitivity and specificity will be used for the purposes of this review. Each of the selected assessment tools will be evaluated for the presence or absence of criteria directly related to the test's overall psychometric validity for the purposes of diagnostic decision-making.

II Method

1 Assessment tool selection

With the understanding that the sensitivity and specificity levels of assessment tools need to be acceptable in order to have any level of identification accuracy, only those tests found to possess adequate levels of sensitivity and specificity were selected for initial review in this study. Based upon the procedure utilized by Spaulding et al. (2006), tests were selected for further study if they met each of three criteria:

1. the test's purpose was for identification of language impairment;
2. the test was not classified as a screening tool; and
3. information related to identification accuracy needed to be provided within the test's examiner's manual, as this is the primary resource available to SLPs engaged in diagnostic efforts.

Thus, with these criteria in mind, the following 10 language assessment tools identified through a review of examiner's manuals as possessing acceptable identification accuracy of .80 or better were selected for initial review (Spaulding et al., 2006):

- *Clinical evaluation of language fundamentals*, 4th edition (CELF-4; Semel, Wiig, and Secord, 2003);
- *Clinical evaluation of language fundamentals: Preschool*, 2nd edition (CELF-P2; Wiig, Secord, and Semel, 2004);
- *Patterned elicited syntax test* (PEST; Young and Perachio, 1993);
- *Preschool language scale*, 4th edition (PLS-4; Zimmerman, Steiner, and Pond, 2002);
- *Structured photographic expressive language test*, 3rd edition (SPELT-3; Dawson, Stout, and Eyer, 2003);
- *Structured photographic expressive language test: Preschool*, 2nd edition (SPELT-P2; Dawson, Stout, Eyer, et al., 2005);
- *Test for examining expressive morphology* (TEEM; Shipley, Stone, and Sue, 1983);
- *Test of early grammatical impairment* (TEGI; Rice and Wexler, 2001);
- *Test of language competence*, expanded edition (TLC-E; Wiig and Secord, 1989); and
- *Test of narrative language* (TNL; Gillam and Pearson, 2004).

Efforts were made to secure each of these assessment tools for psychometric review. Those that were not available through the principle investigator's university clinic were borrowed or purchased, with the exception of the PEST (Young and Perachio, 1993), which is currently out of print and could not be located for review after an extensive search. Thus, the remaining nine assessment tools were further reviewed for this study. Table 1 provides a brief overview of the tests selected for review.

Table 1 Description of selected assessment tools

Name of test	Source	Age range of test	Language area(s) assessed by test
CELF-4	Semel et al., 2003	5;0–21	• Global language skills, including expressive/receptive language, and sound awareness skills
CELF-P2	Wiig et al., 2004	3;0–6;11	• Global language skills, including expressive/receptive language. Includes pre-literacy scale.
PLS-4	Zimmerman et al., 2002	0–6;11	• Global language skills, including expressive/receptive language.
SPELT-3	Dawson et al., 2003	4;0–9;11	• Syntax and morphology
SPELT-P2	Dawson et al., 2005	3;0–5;11	• Syntax and morphology
TEEM	Shipley et al., 1983	3;0–7;0	• Expressive morphology
TEGI	Rice & Wexler, 2001	3;0–8;0	• Syntax and morphology
TLC-E	Wiig & Secord, 1989	5;0–18;0	• Semantics, syntax, and pragmatics
TNL	Gillam & Pearson, 2004	5;0–11;11	• Discourse-based narrative language skills

2 *Procedures for review of selected assessment tools*

Each assessment battery was evaluated for its psychometric validity systematically. With few modifications, criteria developed by McCauley and Swisher (1984) were applied to make judgments regarding the validity and reliability of these assessment tools. One new criterion was added and one existing criterion was modified in order to identify threats to validity more completely. These criteria were applied to the nine language assessment tools selected for evaluation in this study in order to determine their overall level of psychometric validity.

In the following section is a review of McCauley and Swisher's original criteria along with the additions and modifications made for this current project. For a more detailed review of these criteria, refer to McCauley and Swisher (1984). A brief description of all criteria utilized within this study is provided in Table 2.

a Purpose of the assessment tool is identified (Criterion 1): This criterion was added for consideration in this current study as a result of recent research reflecting trends in the development of language assessment tools, relative to the need to identify the purpose of a particular test (Plante and Vance, 1995; Merrell and Plante, 1997; Spaulding et al., 2006).

The purpose of a test is an important component of any assessment tool, as testing is often completed for very different diagnostic reasons (Hutchinson, 1996; Peña, Spaulding, and Plante, 2006). For instance, while an assessment tool might be administered to diagnose the presence or absence of a disorder it might also be used to determine the severity level of a known disorder or to establish treatment goals and/or objectives. This information indicates that clinicians need to be cognizant of the purpose of a given test in order to collect data reflecting their diagnostic needs. Additionally, clinicians need to be aware that assessment tools might purport to serve a specific purpose, but offer no data to substantiate the validity of using a test for that rationale. Further, clinicians should have the awareness that any given standardized testing battery 'may not be able to support multiple diagnostic purposes' (Peña et al., 2006: 252).

Table 2 Psychometric criteria for application to selected assessment tools

Criteria number	Description of criteria
1	Purpose of the assessment tool is identified.[a]
2	Tester qualifications are explicitly stated.
3	Testing procedures are sufficiently explained.
4	Adequate standardization sample size (> 100) is noted.
5	There is a clearly defined standardization sample, including information related to the standardization sample with regard to: geographic representation, socio-economic status / parent education representation, gender distribution,[b] ethnic background,[b] presence/absence of impairment(s), age distribution[b]
6	Evidence of item analysis is given.
7	Measures of central tendency are reported.
8	Concurrent validity is documented.
9	Predictive validity is documented.
10	Test/re-test reliability is reported.
11	Inter-rater reliability is reported.

Notes: All psychometric criteria based on McCauley and Swisher (1984); [a] Addition to the criteria established by McCauley & Swisher (1984);[b] Modification to original criteria established by McCauley & Swisher (1984)

If information related to the purpose of a test is not provided, the validity of the information collected using that tool might well be compromised. Clinical decisions could easily be made after using an assessment tool that was meant for one purpose but used for another. Thus, for the purposes of this review, a test was judged to possess this criterion if the examiner's manual specified the intended purpose(s) of the test for consideration by potential diagnosticians.

b Tester qualifications are explicitly stated (Criterion 2): For an assessment tool to demonstrate this criterion, it needed to specify any special training/qualifications potential necessary to administer and score the test in question. This information is considered to be essential to the validity of a test, as any data collected cannot be considered valid if it is administered and/or interpreted by an unqualified individual.

c Testing procedures are well explained (Criterion 3): For this criterion to be considered present, sufficient detail must have been provided within the examiner's manual to allow for test administration in a manner duplicating the conditions and procedures present at the time the test was standardized. Without this information, clinicians cannot be confident that they are administering the assessment tool in a way that matches the presentation of the test to those in the standardization sample. Any differences in how standardized assessment tools are administered yields scores that cannot be reliably compared to the normative sample. Thus the quality of the data collected can be compromised, rendering test scores unusable for the purpose(s) they intended to fulfill.

d Adequate standardization size (Criterion 4): For an assessment tool to have an adequate sample size, it needed to have a normative sample of 100 or more children per subgroup within the standardization sample. The inclusion of fewer children in a subgroup decreases the validity of test results, as the consistency of test scores is questionable. Test scores that are compared to larger groups of children are more stable, and thus can be used more dependably in the clinical decision-making process. Smaller sample sizes can also be indicative of a less representative sample for comparing scores, as with a small group of children included in the standardization pool it becomes doubtful that all possible subgroups of children (e.g. ethnicity, socio-economic status) have been included in a satisfactory manner, thus rendering the assessment tool in question unusable in many clinical settings.

e *Clearly defined standardization sample (Criterion 5)*: For this criterion to be considered present in a given assessment tool, the examiner's manual needed to provide the following information relative to the normative sample: geographic representation, socioeconomic status, and the language status of those in the normative group (typical vs. atypical language skills). Information related to geographic representation and socioeconomic status is vital for consideration, as tests must be representative of the students undergoing evaluations. Further, the inclusion of information relative to the language status of the standardization group must be reported in light of the fact that the inclusion of language impaired students in the normative group of a given standardized assessment tool has been shown to reduce the identification accuracy of these tests (Peña et al., 2006). Conversely, inclusion of language impaired children in the normative group can be acceptable if the purpose of assessment is to assign a level of severity to a child already reliably diagnosed as language impaired (Peña et al., 2006). Thus, it is critical that clinicians have this piece of information in order to select assessment tools that are appropriate to the purpose(s) they hope to serve throughout the diagnostic process.

Modifications were made to this criterion to further assess the diversity of the normative sample for each test reviewed for this study. Specifically, changes to Criterion 5 included the addition of age and gender distribution as well as ethnic background as normative sample subcategories at the recommendation of Spaulding et al. (2006), who suggest that this information is needed in order to provide sufficient information to determine whether an assessment tool's normative sample is representative of the student(s) to be tested. The second modification to McCauley and Swisher's (1984) criterion was made to an existing normative subcategory for Criterion 5, with parental education level included as an acceptable substitute for socio-economic status (Entwisle and Astone, 1994).

f Evidence of item analysis exists (Criterion 6): Item analysis is used to maximize both the reliability and quality of questions included within a particular test battery by looking at the content of individual questions, screening items for inclusion in the assessment tool, and ensuring that tests target the skills they purport to measure. If an assessment tool lacks item analysis, it is possible that questions might be included that are too difficult or fail to access the skills in question. Thus, use of an assessment tool that fails to report data relative to item analysis could lead to clinical judgments being made on the basis of test questions that were poorly constructed.

For the purposes of this review of assessment tools, a test needed to report evidence that test authors had studied and controlled item difficulty and/or item validity in conducting a thorough item analysis. A variety of methods were deemed acceptable forms of item analysis, including the two most common forms: Classical Test Theory, which looks to improve the reliability of standardized assessment tools, and Item Response Theory, which reflects the probability of performance as a function of a particular level of functioning (Fan, 1998).

g Measures of central tendency are reported (Criterion 7): Assessment tools needed to report the mean and standard deviation of all subtest scores for all groups of the normative sample for this criterion to be present. As these measures are the basis for other scores that are derived for comparison of performance, an assessment tool that fails to report these scores lacks flexibility in the use and interpretation of test its scores, which can impact the validity and reliability of the scores derived from a given testing instrument.

h Concurrent validity is documented (Criterion 8): To demonstrate this criterion, the examiner's manual of each test needed to provide verification of concurrent validity; specifically, evidence demonstrating a correlation between results obtained from the test in question as well as other, similar assessment tools in indicating the presence or absence of a communication disorder. Concurrent validity is important because it demonstrates that results from a given assessment tool are more likely to be valid if a tool that assesses a similar construct has yielded analogous results.

i Predictive validity is documented (Criterion 9): To possess predictive validity, the examiner's manual for each test needed to provide evidence that performance on a given test is predictive of performance observed in a more functional setting through direct observation or other, less formal measures (e.g. student observed using specific language skills within a classroom environment). Absence of predictive validity leads to uncertainty as to how assessment tools and real-life tasks can be compared. Further, decisions related to intervention planning could be compromised as a result of a lack of reliability evident in test scores collected from such instruments.

j Test–retest reliability is reported (Criterion 10): For a test to demonstrate this criterion, values for test–retest reliability must be reported in order to ensure that scores attained on a given test are stable over time. A correlation coefficient of greater than .90 is required for an assessment tool to satisfactorily meet this criterion (McCauley and Swisher, 1984). It should be noted that for tests to possess acceptable levels of test–retest reliability, a short test–retest interval should be observed, as longer intervals between testing could lead to an inflated reliability coefficient that reflects not the actual test–retest reliability of a given test, but spontaneous recovery or maturation that could naturally occur outside a testing situation. A test with that lacks test–retest reliability might yield scores that would fluctuate over time, thus compromising the reliability of reported results.

k Inter-examiner reliability is reported (Criterion 11): Evidence of inter-examiner reliability must be reported in the examiner's manual for this criterion to be present. Inter-examiner reliability ensures that test scores do not fluctuate when different clinicians administer the test battery. A correlation coefficient of .90 is required for a test to meet this criterion (McCauley and Swisher, 1984). A score lower than this cut-off value demonstrates a lack of reliability, as significantly different scores could be observed if the same child was administered the same test by two different clinicians.

3 Data collection and analysis

The examination of each of the nine identified assessment tools was conducted by the principle investigator of this study, as well as five students researchers. A one-time training session was provided to all students in order to familiarize them with each psychometric criterion as well as the process of locating relevant information in assessment tools' examiner's manuals to collect data efficiently. Those examining the assessment tools rated each instrument for the presence or absence of 11 psychometric criteria using a plus (+) or a minus (–) rating. At the conclusion of data collection, each student's data were collected and recorded for subsequent data analysis.

Following the collection of all data, inter-rater agreement across judgments of psychometric validity was calculated and was found to range from 79% (Criterion 6) to 100% (Criteria 1, 4, 5a, 5b, 5c, 5d and 5f). Overall, ratings were recorded for a total percent of agreement across examiners of 97%. Following the methods employed by McCauley and Swisher (1984), the examiner's manual was referenced by the primary investigator and student researcher(s) to resolve any disparity found to exist between examiner ratings. In all, a total of 13 ratings were changed as a result of this process.

III Results

Table 3 presents an account of the presence and absence of each psychometric criterion for each assessment tool. Of the nine tests evaluated to determine their level of psychometric validity, no assessment tool was able to fully meet all 11 criteria applied to them as part of this study. Rather, across the nine instruments that were evaluated, tests were found to fit into a range from eight to 10 criteria met. All told, the CELF-P2 (Wiig et al., 2004), PLS-4 (Zimmerman et al., 2002), SPELT-3 (Dawson et al., 2003), SPELT-P2 (Dawson et al., 2005), and TNL (Gillam and Pearson, 2004) each met 10 criteria while the TLC-E (Wiig and Secord, 1989) successfully met eight psychometric criteria. Examination of each individual psychometric criterion revealed the details shown in Table 3.

Table 3 Presence/absence of psychometric criteria

Criteria description		CELF-4	CELF-P2	PLS-4	SPELT-3	SPELT-P2	TEEM	TEGI	TLC-E	TNL	Percentage that meet criteria
1	Test purpose identified	+	+	+	+	+	+	+	+	+	100
2	Tester qualifications	+	+	+	+	+	+	+	+	+	100
3	Procedures explained	+	+	+	+	+	+	+	+	+	100
4	Adequate sample size	+	+	+	+	+	+	+	+	+	100
5	Sample clearly defined										
	a. geographic representation	+	+	+	+	+	+	+	+	+	100
	b. parent education/SES	+	+	+	+	+	+	+	–	+	89
	c. gender distribution	+	+	+	+	+	+	+	+	+	100
	d. ethnic representation	+	+	+	+	+	–	+	+	+	89
	e. +/– impairment	+	+	+	+	+	+	+	+	+	100
	f. age distribution	+	+	+	+	+	+	+	+	+	100
6	Evidence of item analysis	+	+	+	+	+	–	+	+	+	89
7	Measures of central tendency	+	+	+	+	+	+	+	+	+	100
8	Concurrent validity	+	+	+	+	+	+	+	+	+	100
9	Predictive validity	–	–	–	–	–	+	–	–	+	22
10	Test/retest reliability	–	+	+	+	+	+	–	–	–	56
11	Inter-examiner reliability	+	+	+	+	+	+	+	+	+	100
	Number that met critiria (per assessment tool)	9/11	10/11	10/11	10/11	10/11	9/11	9/11	8/11	10/11	

Each of the reviewed assessment tools met Criteria 1, 2 and 3, indicating that each were found to have provided a clear purpose statement for the assessment tool, explicitly identified tester qualifications, and sufficiently explained test administration procedures.

With regard to the normative sample, seven of the nine assessment tools evaluated demonstrated both adequate sample size (Criterion 4) and the presence of a clearly defined standardization sample (Criterion 5). These tests included the CELF-4 (Semel et al., 2003), CELF-P2 (Wiig et al., 2004), PLS-4 (Zimmerman et al., 2002), SPELT-3 (Dawson et al., 2003), SPELT-P2 (Dawson et al., 2005), TEGI (Rice and Wexler, 2001), and TNL (Gillam and Pearson, 2004). It should be noted that with the exception of the TNL, each of these assessment tools chose to represent socio-economic status with parent education level, rather than as a measure of household income. The remaining assessment tools – the TLC-E (Wiig and Secord, 1989) and the TEEM (Shipley et al., 1983) – met Criterion 4, indicating an adequate sample size, but were not successful in meeting Criterion 5 due to failure to clearly define socio-economic status (TLC-E) and ethnic representation (TEEM).

Eight of the nine assessment tools that were evaluated demonstrated sufficient evidence of item analysis procedures applied as tests were constructed and, thus, met Criteria 6. Only the TEEM (Shipley et al., 1983) lacked such evidence within its examiner's manual and was not found to meet this criterion.

Each of the assessment tools that were evaluated met both Criteria 7 and 8, which required that measures of central tendency and coefficients related to concurrent validity were reported within the examiner's manual. Predictive validity (Criterion 9) was only demonstrated within two of the nine evaluated assessment tools: the TEEM (Shipley et al., 1983) and the TNL (Gillam and Pearson, 2004). Overall, this criterion was the least represented psychometric trait amongst the 11 applied to these assessment tools.

Criterion 10 asked that test–retest reliability be demonstrated with a coefficient of .90 or better. Five tests were able to fully meet this criterion, including the CELF-P2 (Wiig et al., 2004), PLS-4 (Zimmerman et al., 2002), SPELT-3 (Dawson et al., 2005), SPELT-P2 (Dawson et al., 2003), and TEEM (Shipley et al., 1983). Other tests were able to partially meet this criterion, meaning that portions/subtests of the assessment tools demonstrated acceptable test–retest reliability, while other portions/subtests did not.

Inter-examiner reliability was examined (Criterion 11) for each assessment tool, as well, with the requirement that a coefficient of .90 or greater be present for a test to meet this criterion. All assessment tools successfully met this criterion.

IV Discussion

Determining the ideal assessment tool(s) for use in assessing the language of preschool and school-aged children is a difficult endeavor, as each standardized test available for clinical use has distinct strengths and weaknesses. Thus, it is the responsibility of each clinician that engages in diagnostics as part of their clinical practice to become an informed user of these assessment tools through careful examination of their properties. The consequences of not conducting a thorough examination of each test under consideration for clinical use are high: students may be improperly diagnosed, leading to inappropriate provision or denial of clinical services in the absence of other diagnostic data to refute such findings. Fortunately, each commercially available standardized assessment tool contains an examiner's manual, which provides the information SLPs must access and critically review in order to make educated diagnostic choices.

Examination of nine language assessment tools yielded encouraging data related to their overall psychometric validity. Comparatively speaking, results indicate that assessment tools selected for inclusion in this study are more mindful of issues related to validity and reliability than was evident in previous inquiry (McCauley and Swisher, 1984) and seemingly reflected trends indicating that the overall psychometric validity of standardized assessment tools has improved dramatically in the last two decades (Mikucki and Larrivee, 2006). That said, it is important to remember that only nine assessment tools were reviewed in depth here, as the vast majority of the standardized language assessment tools currently available to practicing clinicians lack the 'gold standard' for inclusion in this study, namely adequate levels of identification accuracy. Thus, this lack of information relative to sensitivity and specificity appears to be a major limitation of current test development practices. While sensitivity and specificity levels can be calculated by clinicians prior to assessment (Paul, 2007), if requisite data is provided by test developers within the examiner's manual, this is likely not a standard assessment practice. With this in mind, it would seem that test publishers and authors should make a concerted effort to include this data in all tests intended for diagnostic use by SLPs to allow for choices to be made that are based on sound diagnostic principles.

Of the psychometric criteria evaluated, seven were met within each of the nine tools examined (Criteria 1, 2, 3, 4, 7, 8, and 11). Certainly this information is valuable to potential diagnosticians and represents a positive step from previously obtained results (McCauley and Swisher, 1984). Notably absent from most tests in this study was information associated with the predictive validity of these selected assessment tools. Of the nine tests reviewed, only the TNL (Gillam and Pearson, 2004) and the TEEM (Shipley et al., 1983) reported this data within their examiner's manuals. Thus, a large majority of tests that were evaluated for this study failed to provide this data for users of their assessment tools. This could be due to inconsistency in the use of predictive validity as a measure of importance, as there seem to be two schools of thought with regard to predictive validity relative to test construction. Some researchers suggest that predictive validity should be eliminated from commonly applied psychometric validity criteria, as concurrent validity can be equally (or even more) useful in the diagnosis of a client (Anastasi and Urbina, 1997; Mikucki and Larrivee, 2006). Conversely, other researchers indicated that because predictive validity is commonly used to forecast performance on real-life tasks at an unspecified time in the future, information from this criterion is important for intervention planning and, thus, should be reported by test authors and publishers in the future (McCauley and Swisher, 1984; Paul, 2007).

Caution should be used in the interpretation and application of data reported within this study, particularly with regard to Criterion 5. Clinicians should keep in mind that assessment tools were reviewed only for the presence or absence of information from the examiner's manuals mandating the clear definition of the normative sample. Thus, no rating was applied to judge the acceptability of the data provided. For instance, the TEEM (Shipley et al., 1983) does provide information about the normative sample and, accordingly, earned credit for the presence of that criterion. However, careful examination of the examiner's manual reveals that the TEEM was standardized using 'middle class' children from an isolated geographic region (p. 9). With the knowledge that tests should only be used with children whose personal demographics are represented within the normative sample (Hutchinson, 1996; Dollaghan, 2004; Paul, 2007), clinicians must take care to scrutinize examiner's manuals prior to the use of such assessment tools.

1 Selecting standardized tests for clinical use

Information related to the psychometric validity of the tests selected for review in this study (as well as any others that clinicians might choose to individually review) can and should be applied carefully to assist in the decision-making process that accompanies all diagnostic endeavors. Clinicians must weigh which properties of a given test should be considered as most essential and, thus, more important to focus upon in selecting assessment tools for diagnostic use.

Based on the results of the review undertaken as part of this study, the following guidelines are recommended for clinicians to consider in selecting assessment tools for diagnostic use. These suggestions are represented graphically in Figure 1.

a Identification accuracy: The first and most important consideration for clinicians in selecting a test for use must be the identification accuracy of the test in question. Clinicians who fail to take this information into consideration face the threat of misdiagnosis when making determinations related to the presence or absence of a language disorder. Thus, clinicians must determine whether information is provided to inform users of the overall identification accuracy of the test in question. At times, test authors/publishers will provide data reflecting the sensitivity and specificity of a particular assessment tool within the examiner's manual. Other times, tests might report information that will allow for sensitivity and specificity to be calculated, allowing clinicians to compute these identification accuracy values using a simple pair of formulas that juxtapose data provided to reflect testing precision (Dollaghan, 2007; Paul, 2007). Formulas for computing identification accuracy values can be found in Table 4.

Overall, if information related to the identification accuracy of a test is not provided (or cannot be calculated), clinicians should question whether the assessment tool should be used at all, for if a test is not able to dependably identify the presence or absence of language impairment, clinicians cannot reliably interpret scores for use in determining eligibility for services. Spaulding et al. (2006) provide guidelines for clinicians who choose to administer assessment tools that lack identification accuracy information. These guidelines suggest that clinicians use group differences in scores obtained by language impaired vs. normative samples as a method of judging the level of confidence clinicians can have in scores obtained using these tests. In presenting this option, however, the following caveat must be considered: 'the weight that a clinician gives to a test score in making his or her final diagnostic decision must be modulated by the strength of the data available to support that decision' (Spaulding et al., 2006: 70).

Table 4 Formulas for calculating identification accuracy values

Sensitivity	Specificity
# true positives[a]	# true negatives[c]
# true positives + # false negatives[b]	# false positives[d] + # true negatives

Notes: Formulas based on information provided within Dollaghan (2007). [a] Number within sample that are language disordered. [b] Number within sample that have a language disorder but are classified as non-disordered using the test in question. [c] Number within sample that do not have a language disorder. [d] Number within sample that are non-language disordered, but are classified as language disordered using the test in question.

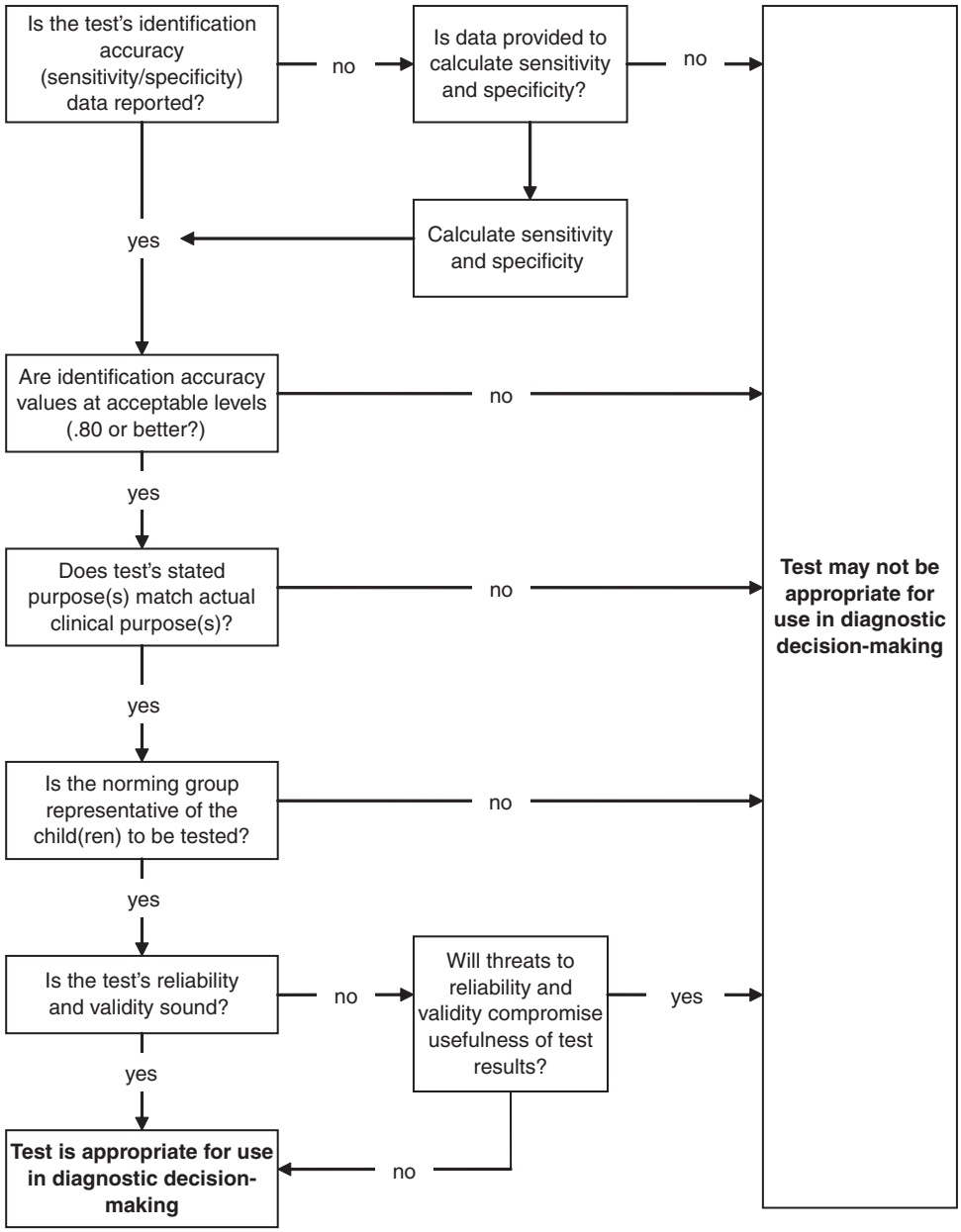

Figure 1 Decision tree for consideration of psychometric validity in selecting tests for clinical use

b Other aspects of psychometric validity: Once information related to identification accuracy has been obtained, clinicians should turn their attention to specific psychometric properties of the tests they plan to utilize. Eleven separate psychometric criteria compromised the core of those used to evaluate the preschool and school-age language tests for this study. Because this amount of information can be confusing – and since identifying the largest threats to validity and reliability is challenging, at best – the question becomes, how can clinicians identify the most important of these psychometric characteristics for their consideration? And, how much evidence is sufficient in making diagnostic decisions?

First and foremost, clinicians must be able to determine if the stated purpose of the assessment tool that they intend to use matches their own clinical purposes. Hutchinson (1996) and Peña et al. (2006) have indicated the importance of identifying the purpose of an assessment tool prior to administration to ensure that information is collected that reflects the actual aim of the assessment being conducted. Ultimately, if clinicians utilize a particular test to serve one clinical purpose when the test is actually designed to address another then the results obtained will not be valid or reliable for diagnostic use.

Once clinicians have ascertained that the stated purpose of a test matches their clinical objectives, they must turn their attention to the normative sample to ensure that it does, in fact, reflect the demographics to which their individual clients belong. A lack of demographic representation in age, gender, socio-economic status, geographic residence, ethnic background and/or presence/absence of language impairment can lead to inaccurate pairings of the child(ren) being tested and the group to which they are compared, which can lead to misdiagnosis. As the normative sample has provided the data that allows for scores to be computed and reported – and, in some circumstances, for the presence or absence of a disorder to be determined – it is the onus of all clinicians engaged in ethical diagnostic activities to ensure that each child is evaluated competently. Thus, if a child's demographics are not represented within a normative sample, that test is not appropriate for use with that child.

The remainder of the psychometric criteria applied in this study exists on a parallel continuum, meaning that they are all of essentially equivalent weight in making diagnostic decisions. Thus, rather than listing these criteria sequentially, clinicians should choose to look at two separate areas as they review these psychometric characteristics: reliability and validity.

Criteria that relate to reliability (test–retest reliability and inter-examiner reliability) exist to ensure that scores obtained for a particular test will be consistent over time. These are doubtless important concerns, as the indication that test scores might vary from one administration to the next or from one tester to another are cause for apprehension. Fortunately, the merit of values reported in examiner's manuals describing these criteria is relatively easy to judge as there is a 'cut-off' value for acceptability: a coefficient of .90 or higher (McCauley and Swisher, 1984).

Remaining criteria related to validity (sample size, item analysis, reporting of measures of central tendency, concurrent validity, and predictive validity) are worthy of scrutiny as they indicate how well a given assessment tool actually tests the knowledge areas it purports to measure. Focusing on the validity of an assessment tool allows clinicians to know that the instrument that they have chosen to address a particular diagnostic purpose has been constructed to do so accurately.

This current review indicated the presence or absence of these validity-related criteria, which is not difficult to ascertain. It is more complex to make judgments regarding whether the information related to a test's validity provided within the manual is sufficient enough to be deemed acceptable. With the exception of sample size and the reporting of measures of central tendency, the remaining validity-related criterion lack cut-and-dried explanations of sufficiency. Overall, it might be more

effective to look for quality rather than quantity. Criteria related to concurrent and predictive validity ask that clinicians compare scores from a given test to other, validated, methods to make suppositions related to a test's ability to provide accurate measurements. The question, then, is what are these other methods that are being used for comparison? Other standardized tests that might lack acceptable levels of identification accuracy or psychometric validity would not be a desirable comparison. Conversely, tests that do possess high levels of precision in discriminating disordered from non-disordered children would be advantageous for use in making comparisons. Non-validated clinician judgments or less formal criterion measures may not be valid, either, although if they reflect how a child will need to use his/her language skills in the 'real world', then they may be preferred.

Overall, information that clinicians collect in considering reliability and validity should be carefully considered to determine whether any identified threats would undermine the usefulness of the data they might collect in using the test. If clinicians judge any threats to reliability and validity as minimal, then the test is likely appropriate for clinical use. On the other hand, if considerable concern is evident in evaluating the reliability and validity of a particular test, then it is likely unsuitable for diagnostic use.

2 Integration of standardized tests within a language assessment battery

Just as clinicians have the responsibility for being informed users of standardized tests, they are accountable for formulating a holistic picture of a child's language strengths and weaknesses as part of the diagnostic process. This current study has focused exclusively upon the evaluation and application of standardized tests, yet it is important that all results garnered from a standardized test be considered in light of other information collected as part of a language evaluation, as 'no single measure can provide sufficient data to create an accurate and comprehensive [language] profile' (American Speech–Language–Hearing Association, 2000: 18). Past research has identified other concerns relative to the application of standardized test scores alone in making eligibility determinations. In fact, there is tremendous support for the use of other forms of testing to substantiate or refute results collected using a standardized assessment (Spaulding et al., 2006; Paul, 2007; Roseberry-McKibbin, 2007).

Doubtless, data collected within this study has demonstrated that most commercially available standardized tests are imperfect. That said, clinicians should operate under the notion that they do need to carefully consider many important factors in choosing a standardized test for administration, as standardized tests are an important component of a language assessment battery. Optimally, the diagnostic ideal would include standardized tests used in conjunction with other information from multiple sources, collected in various environments to inform decisions related to eligibility and treatment planning (Plante and Vance, 1994) since, ultimately, clinicians have the onus of balancing a variety of data to make well-informed clinical decisions.

References

American Speech–Language–Hearing Association (2000) *Guidelines for the roles and responsibilities of the school-based speech–language pathologist*. Available online at www.asha.org/policy (November 2009).

Anastasi A and Urbina S (1997) *Psychological testing*. 7th edition. Upper Saddle River, NJ: Prentice-Hall.

Dawson J, Stout C, and Eyer J (2003) *Structured photographic expressive language test*. 3rd edition. DeKalb, IL: Janelle.

Dawson J, Stout C, Eyer J, Tattersall P, Fonkalsrud J, and Croley K (2005) *Structured photographic expressive language test: Preschool*. DeKalb, IL: Janelle.

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

Dollaghan CA (2007) *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Paul H Brookes.

Entwisle DR and Astone NM (1994) Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Development* 65: 1521–40.

Fan X (1998) Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement* 53(3): 1–17.

Gillam RB and Pearson NA (2004) *Test of narrative language*. Austin, TX: Pro-Ed.

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

Hutchinson TA (1996) What to look for in the technical manual: Twenty questions for users. *Language, Speech, and Hearing Services in Schools* 27: 109–21.

McCauley RJ and Swisher L (1984) Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders* 49: 34–42.

Merrell AW and Plante E (1997) Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools* 28: 50–58.

Mikucki BA and Larrivee L (2006) Validity and reliability of twelve child language tests. Unpublished poster session presented at the American Speech–Language–Hearing Association's national convention, Miami, FL.

Paul R (2007) *Language disorders from infancy through adolescence: Assessment and intervention*. St Louis, MO: Mosby.

Peña ED, Spaulding TJ, and Plante E (2006) The composition of normative groups and diagnostic decision-making: Shooting ourselves in the foot. *American Journal of Speech–Language Pathology* 15: 247–54.

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

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

Rice ML and Wexler K (2001) *Test of early grammatical impairment*. San Antonio, TX: Psychological Corporation.

Roseberry-McKibbin C (2007) *Language disorders in children: A multicultural and case perspective*. Boston, MA: Allyn and Bacon.

Roulstone S, Peters TJ, Glogowska M, and Enderby P (2008) Predictors and outcomes of speech and language therapists' treatment decisions. *International Journal of Speech–Language Pathology* 10(3): 146–55.

Semel E, Wiig EH, and Secord WA (2003) *Clinical evaluation of language fundamentals*. 4th edition. San Antonio, TX: Psychological Corporation.

Shipley KG, Stone TA, and Sue MB (1983) *Test for examining expressive morphology*. Austin, TX: Pro-Ed.

Spaulding TJ, Plante E, and Farinella KA (2006) Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools* 37: 61–72.

Wiig EH and Secord W (1989) *Test of language competence*. Expanded edition. San Antonio, TX: Psychological Corporation.

Wiig EH, Secord WA, and Semel E (2004) *Clinical evaluation of language fundamentals: Preschool*. 2nd edition. San Antonio, TX: Psychological Corporation.

Young EC and Perachio JJ (1993) *Patterned elicited syntax test*. Tucson, AZ: Communication Skill Builders.

Zimmerman IL, Steiner VG, and Pond RE (2002) *Preschool language scale*. 4th edition. San Antonio, TX: Psychological Corporation.