Published in final edited form as:
Lang Speech Hear Serv Sch. 2009 April ; 40(2): 150–160. doi:10.1044/0161-1461(2008/07-0049).

The Diagnostic Accuracy and Construct Validity of the Structured Photographic Expressive Language Test—Preschool: Second Edition

Kathryn J. Greenslade, **Elena Plante**, and **Rebecca Vance**
University of Arizona, Tempe

Abstract

Purpose—In order to support evidence-based practice, this study served to evaluate the diagnostic accuracy, convergent validity, and divergent validity of the Structured Photographic Expressive Language Test—Preschool: Second Edition (SPELT–P2; J. Dawson, J. A. Eyer, J. Fonkalsrud, 2005) in order to determine whether it can be used as a valid measure for identifying language impairment in preschoolers.

Method—The SPELT–P2 was administered to 54 children with typically developing language and 42 children with specific language impairment.

Results—A discriminant analysis revealed good sensitivity (90.6%), good specificity (100%), and good positive and negative likelihood ratios, with a standard score cutoff point of 87 used to determine group membership. Analyses of convergent and divergent validity also supported use of the SPELT–P2 for identifying language impairment in preschoolers.

Implications—The empirical evidence supports use of the SPELT–P2 as a valid measure for correctly identifying the presence or absence of language impairment in 4- and 5-year-old preschool children.

Keywords

diagnostics; evidence-based practice; language; specific language impairment

Norm-referenced tests are frequently used when diagnosing young children with language impairment. Because of the social, academic, monetary, and ethical consequences of such a diagnosis, it is critical to ensure that the tests that are designed to determine the presence or absence of a language impairment are indeed valid for that purpose (Anastasi, 1988; Messick, 1989; Plante & Vance, 1994; Spaulding, Plante, & Farinella, 2006; Zhang & Tomblin, 2000). Since 1997, the Individuals With Disabilities Education Act (IDEA, 1997, P.L. 105-17) has echoed this necessity, requiring that assessments and evaluation measures be "used for purposes for which [they] are valid and reliable" (118 Stat. 2705). This concept was retained in subsequent versions of this law (e.g., Individuals with Disabilities Education Improvement Act of 2004; P.L. 108-446, Stat. 2705).

As a preliminary step, many test makers consult a panel of experts to evaluate whether the content of the test appears representative of the phenomenon being assessed (content validity).

Contact author: Elena Plante, P.O. Box 210071, Department of Speech, Language, and Hearing Sciences, The University of Arizona, Tucson, AZ 85721-0071. E-mail: E-mail: eplante@email.arizona.edu.

However, this process is insufficient to establish the extent to which a test reflects its underlying construct (construct validity) (Gray, Plante, Vance, & Henrichsen, 1999; Messick, 1989). Instead, Dollaghan (2004) advocated an evidence-based practice framework that relies on credible studies to assess diagnostic assessments and procedures. Thus, before using a test for diagnostic purposes, the test's construct validity should be established through empirical evaluation.

The primary evidence needed to demonstrate that a test can be used to identify cases of language impairment is its diagnostic accuracy. An evaluation of diagnostic accuracy must prove that the test is able to adequately identify children with language impairment as having language impairment (sensitivity) and identify children with typical language development as having typically developing language (specificity). Plante and Vance (1994) recommended a criterion of 90%–100% for sensitivity and/or specificity to be considered "good" and a criterion of 80%–89% for sensitivity and/or specificity to be considered "fair" when assessing the diagnostic accuracy of an identification assessment. Although these levels have been adopted by other investigators (e.g., Gutiérrez-Clellen & Simon-Cereijido, 2007; O'Neill, 2007; Peña et al., 2006; Restrepo, 1998), others have left interpretation of the adequacy of sensitivity and specificity to those choosing to use the test (e.g., de Beaman et al., 2004; Emmons & Alfonso, 2005; Sturner, Heller, Funk, & Layton, 1993).

A recent review of 43 tests of child language confirms that the critical evidence needed to support the clinical use of tests is often missing in test manuals. Spaulding et al. (2006) reported that only nine of the 43 tests that they reviewed contained information on sensitivity and specificity in the manual. Of these nine tests, only five reported sensitivity and specificity levels of at least 80%. Acceptable sensitivity and specificity levels could be documented for an additional five tests by drawing from data reported in published test validation studies. Therefore, the data to support the use of a majority of the current commercially available tests is absent in either the test manual or the professional literature.

Alternately, Dollaghan (2004) recommended evaluating a test's diagnostic accuracy by finding positive and negative likelihood ratios.[1] Like sensitivity and specificity estimates, positive and negative likelihood ratios represent the degree of confidence that test scores yield accurate classifications of disordered individuals as disordered and normal individuals as normal, respectively. Though derived from sensitivity and specificity data, positive and negative likelihood ratios are thought to be less influenced by the characteristics of the sample from which they are calculated. Dollaghan suggested that positive likelihood ratios greater than 10 and negative likelihood ratios less than 0.2 indicate that scores are very likely to correspond with correct diagnoses.

A test's sensitivity and specificity are completely dependent on the cutoff score that is used to classify individuals as normal or impaired. Individuals scoring above the cutoff score are deemed to be typically developing, whereas those scoring below the cutoff are determined to be impaired. Many clinicians have been trained to believe that a single cutoff score that reflects the low end of a normal distribution (e.g., −1.5 *SD*s, −2.0 *SD*s) can be used with any language test to detect cases of language impairment. There is now substantial data demonstrating that this practice does not lead to accurate diagnoses. This is because children with impaired language frequently do not obtain scores that fall below these commonly applied cutoff scores. Spaulding et al. (2006) demonstrated this point by presenting data on mean scores that were obtained by children with impaired language as these were reported in test manuals. This review illustrated that the information that was readily available in the test manuals failed to support

[1]The positive likelihood ratio of an assessment is calculated using the formula: sensitivity/(1 – specificity). The negative likelihood ratio of an assessment is calculated using the formula: (1 – sensitivity)/specificity.

use of these commonly applied cutoff scores for a majority of the commercially available language tests (Spaulding et al., 2006). Instead, the current standard of best practice is use of a cutoff score that has been empirically derived for use with an individual test in order to ensure maximal effectiveness in differentiating individuals with typical language development from those with language impairment (Plante & Vance, 1994). This data-based approach maximizes sensitivity and specificity, as well as positive and negative likelihood ratios, and therefore maximizes overall diagnostic accuracy. However, derived cutoff scores are test specific. Even within similar test domains (e.g., expressive language), the cutoff score derived for one test can differ significantly from that of another test, even when these tests were validated on the same sample of children (e.g., Plante & Vance, 1994, 1995).

The present study was designed to evaluate the construct validity of the Structured Photographic Expressive Language Test—Preschool: Second Edition (SPELT–P2; Dawson, Eyer, & Fonkalsrud, 2005) by examining its diagnostic accuracy as well as aspects of its convergent and divergent validity (these aspects of validity are discussed below). A previous study of the SPELT—Preschool (SPELT–P; Werner & Kresheck, 1983b) found acceptable sensitivity (83.3%) and specificity (95%) at a standard score cutoff point of 79 (Plante & Vance, 1995). Unfortunately, the SPELT–P2 manual contains no data on sensitivity, specificity, or likelihood ratios to inform the clinician concerning the revised test's diagnostic accuracy. The inclusion of different test items and a new normative sample raise the strong possibility that data on diagnostic accuracy derived from the SPELT–P would not apply to the SPELT–P2. Indeed, there have been other examples where a new version of a test has had strikingly different diagnostic accuracy than its predecessor (e.g., Ballantyne, Spilkin, & Trauner, 2007). Thus, an empirical analysis of revised tests is needed to determine their diagnostic effectiveness. Lack of such data for the SPELT–P2 prevents clinicians from evaluating the potential use of this test in an evidence-based manner.

According to its manual, the SPELT–P2 was modeled after the design of the SPELT–II (Werner & Kresheck, 1983a) and SPELT–3 (Dawson, Stout, & Eyer, 2003), which are similar tests that were designed for older children. Both of the latter tests have acceptable diagnostic accuracy. The sensitivity and specificity of the SPELT–3 are 90% and 100%, respectively, at a standard score cutoff of 95 (Perona, Plante, & Vance, 2005). The sensitivity and specificity of the SPELT–II are 90% and 90%, respectively, at a standard score cutoff of 51 (Plante & Vance, 1994, 1995). The SPELT–P2 includes three items modified from, and five items found on, the SPELT–3, and 13 of the picture stimuli used in the SPELT–P2 are also used in the SPELT–3. Therefore, the content of the SPELT–P2 is largely independent of these other tests, even though the format for testing language skills is highly similar.

The SPELT–P2 contains six test items that are identical to the original SPELT–P, an additional nine items are modified from this test, and 25 of the items are new. All of the photographs used as stimuli have been changed from those found in the SPELT–P. Six items that are on the SPELT–P are not included in the revised version. Therefore, there is noteworthy content overlap between these two tests, but they are different enough that independent validation of the revised SPELT–P2 is warranted.

The stated purpose of the current edition of the test is "to identify those children who may have difficulty in their expression of early developing morphological syntactic features" in children ages 3;0 (years; months) to 5;11 (Dawson et al., 2005, p. 1). This assessment uses one or two color photographs per item and verbal prompts to elicit responses with targeted structures. For example, item number 13 uses a single photograph, which shows a boy watching television and his mother pointing to her watch. The examiner says, "It is time to go to bed, but he wants to watch more TV. What does the boy tell his mother?" The child is expected to respond with

a statement containing a subject pronoun. In this case, a response starting with either "I" or "he" receives full credit.

A consideration of the content of the test (content validity) indicates that children must comprehend the verbal prompts that are more elaborate than those that are typically found in "expressive" language tests. In some cases, these prompts involve multiple sentences that the child must process in order to formulate the correct response. This logically suggests that a breakdown in comprehension of the prompts could lead to a failure to produce the target response expressively. However, from an evidence-based perspective, a content analysis in the absence of data is not sufficient to determine whether this test reflects expressive language only, or whether receptive language should be considered a secondary construct that can influence children's performance. This can be empirically evaluated by determining whether there is a measurable overlap between performance on the SPELT–P2 and performance on a more traditional receptive language test (i.e., concurrent validity).

Although content validity reflects subjective judgment, the extent to which this test reflects expressive morphosyntax or broader language skills can also be empirically assessed. Construct validation can include analysis of whether a test yields a similar result to other assessments measuring the same construct (convergent validity). In this case, an analysis must be performed to determine whether the SPELT–P2 results correlate with results from alternative measures of expressive language. Furthermore, test results should be independent of factors that are beyond the scope of the measured construct (divergent validity). For example, participants' race and ethnicity should not affect performance on a language assessment, as these factors are peripheral to the construct of language. Thus, to assess the construct validity of the SPELT–P2, this test's diagnostic accuracy, convergent validity, and divergent validity were examined.

The primary purpose of this study was to determine whether the SPELT–P2 can be used to identify cases of language impairment in 4- and 5-year-old children. The secondary purpose was to examine aspects of convergent validity and divergent validity, each of which could affect interpretation of the test scores obtained.

METHOD

Participants

Ninety-six children participated in the current study, including 68 boys and 28 girls. The participants ranged in age from 48 months to 68 months, with a mean age of 57 months.[2] Recruitment sources included a university clinical program (n = 4), three local public schools that had preschool/kindergarten programs for children with language impairment (n = 4), and 21 preschool programs (n = 88) in the Tucson area. The majority of the children were reported by their parents to be white, and approximately one third were reported to be Hispanic. The mean number of years of maternal or caregiver education was 14.4 for the typically developing group (TD) and 14.0 for the group with specific language impairment (SLI). All children spoke English as their native language, according to caregiver report, and attended schools where English was the only language spoken. No child spoke a nonmainstream dialect. The caregivers of 12 of the children (6 TD, 6 SLI) reported that their child was exposed to a second language at home. One child in each group was exposed to multiple languages at home. These languages included Spanish (n = 8), French (n = 1), Italian (n = 1), Hawaiian (n = 1), Chinese (n = 1), Korean (n = 1) and sign language[3] (n = 1).

[2]Although the SPELT–P2 is normed on children from 3;0 to 5;11, for logistical reasons, children below the age of 4 years were not included in this study. Due to this exclusion of 3-year-old children, the results obtained in this study should not be assumed to generalize to the performance of 3-year-old children.

The study was approved by the local institutional review board. Informed consent was obtained for all children, and children provided verbal assent for their participation.

Norm-referenced language measures—A set of norm-referenced language measures was given to each child to assist with the classification of language status. The Test for Examining Expressive Morphology (TEEM; Shipley, Stone, & Sue, 1983) was selected as a formal, standardized measure because evidence from previous studies has shown the TEEM to have adequate to high levels of discriminant accuracy for children of the age studied here. Specifically, using a cutoff of 75, Merrell and Plante (1997) reported a sensitivity of 90% and a specificity of 95%, and Perona et al. (2005) found a sensitivity of 88.1% and a specificity of 85.4%. The TEEM examines children's morphosyntactic structures through cloze tasks. For example, the prompt for one item states, "Here is one boat. Here are two ___." In this case, the child is expected to complete the statement with the word *boats*. Thus, typical responses on the TEEM consist of single words, whereas the SPELT–P2 generally requires a phrase or sentence response.

Note that the children were also given the Grammatic Under-standing subtest of the Test of Language Development—Primary: Third Edition (TOLD–GU; Newcomer & Hammill, 1997). Although there is no evidence to support use of this subtest for the purpose of determining language status (Spaulding et al., 2006), the test was used in this study to evaluate concurrent validity. This purpose will be discussed further in the Materials and Procedure section.

Nonverbal measure—In order to rule out mental retardation, the children had to receive a score of 75 or greater on the nonverbal scale of the Kaufman Assessment Battery for Children—Second Edition (K-ABC–II; Kaufman & Kaufman, 2004). The cutoff score was selected in accordance with the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders—Fourth Edition* (1994), which defined "significantly subaverage intellectual functioning" as an IQ score falling two or more standard deviations below the mean (i.e., IQ ≤ 70). As the standard measurement error for this intelligence test is 5 points, a score of 75 or better (70 + 1 SEM) ensured that participants truly fell within normal limits in terms of cognitive functioning.

Hearing screening—To be included as participants, all children were also required to pass a hearing screening at 500, 1000, 2000, and 4000 Hz to rule out hearing loss as a cause of language deficits. At some testing sites, ambient noise interfered with testing at low frequencies, and as a consequence, the acceptable intensity level at 500 Hz was raised at these sites to 30 dB HL but remained at 20 dB HL for the three remaining frequencies.

Demographic and informal measures—Informal measures were used to obtain information about the participants. Participants' care-givers were provided with a questionnaire that asked for information on the child's race and ethnicity, the caregivers' education level, the child's native language, and any other languages spoken in the home. In addition to the demographic information gathered from parent questionnaires, both teachers and parents were asked to answer specific questions regarding concerns they might have about the child's hearing, speech, language, motor, and thinking skills. Parent questionnaires also inquired about the presence of handicapping conditions (e.g., neurological or developmental disorders). The parent and teacher questionnaires were highly similar in content. For the subset of children who were receiving therapy for a language impairment, the clinicians were also provided with

[3]Insufficient information was available to know whether the report of "sign language" represented use of a signed second language or use of individual signs.

a questionnaire that asked about the child's current speech and language functioning. In this case, clinicians were asked to rate the degree to which the child's language impairment affected his or her performance in 10 areas of language, such as "following the classroom routine," "maintaining age-appropriate attention," and "asking simple questions (who, what, where, yes/no)." The child's level of difficulty in each of these areas was rated as *severe, moderate, mild*, or *none*. The general content of the questionnaires can be found in Perona et al. (2005).

In addition to the information concerning language functioning that was provided through the questionnaires, the speech-language pathologists (SLPs) attached to this study obtained information concerning language skills though a brief and informal conversation with each child. This conversational sample was not intended to support a formalized analysis of language skills, but rather to provide a general impression of whether the conversational language deviated from normal age expectations in the judgment of the SLP. This sample also permitted a quick assessment of whether speech sound errors or omissions were likely to interfere with language testing.

Classification of participants—According to Dollaghan (2004), the validity of a new diagnostic assessment should be determined through comparison to a *gold standard*, which is reliable, valid, and reasonable. In the present study, the standard for determining each participant's status as TD or SLI involved a combination of clinical judgment and formal testing. Converging evidence from both sources increases confidence in the standard's diagnostic accuracy (refer to the avoidance of *mono-operation bias*, Cook & Campbell, 1979, p. 65).

For this study, clinical judgment concerning the child's language status was based on converging evidence from several sources of information. In order for a child to meet the study standard for either TD or SLI, the qualitative judgment of impaired language had to agree with an independent test-based classification. In addition, the presence of conditions that would indicate other reasons for poor language (i.e., hearing loss, mental retardation, report of other handicapping conditions, or nonnative speakers of English) would also exclude a child from the study sample.

Language was formally tested using the TEEM. Based on the validation study by Merrell and Plante (1997), a score of 75 or greater on this test was taken to indicate typical language status, and lower scores were interpreted as signaling the presence of a language impairment.

Test-based classification was complemented by an independent classification based on the clinical judgment of a clinically certified SLP. This judgment was based on impressions of the child's language that were formed during the conversation between each child and one of the two SLPs associated with this study. In addition, each clinician considered information that was gathered through the parent and teacher questionnaires that were available for all children, as well as the clinician questionnaire for those children who were receiving services. Several studies have noted that parent questionnaires can be used to provide accurate indications of current skills (e.g., Glascoe, 2000; Marchman & Martínez-Sussman, 2002; O'Neill, 2007). In particular, Glascoe reported that when parent questionnaires ask about each developmental area, they have similar accuracy as many screening tests in identifying delays. Each of these pieces of information was considered, and a holistic judgment of language status was made by the clinician who interacted with that child. It is important to note that this qualitative judgment was made independently from consideration of the standardized test results.

Of the 96 children who participated in the study, 51 met the study standard for TD. Three additional children were judged to have typical language skills based on spontaneous conversations with the study's SLPs and according to the parent and teacher report of speech

and language development. However, each of these children received a score of less than 75 on the TEEM, and thus did not meet the study standard for TD. Because these children reflect abilities that are present in the general population, the data from these children were retained to determine whether they would be identified as TD by the SPELT–P2.

Forty-two children were classified as having SLI. It should be noted that of the study's 42 children with SLI, 11 were receiving speech and language therapy at the time of participation; 31 additional children were identified as having SLI through participation in the study. This rate of identification is consistent with the epidemiologic data of Tomblin et al. (1997), which showed that a majority of kindergarten-age children with SLI had not been identified before that prevalence study. The comparable numbers of children in our study relative to the earlier epidemiologic study suggests that the present sample is closer in composition to a population sample of the disorder than to a referred sample. This probably occurred because the data collection involved the recruitment and testing of large numbers of children at multiple sites that were selected to provide racial, ethnic, and socioeconomic diversity to the sample. Thus, the recruitment simulated population sampling to some extent. In contrast, recruiting a clinical sample solely through a single, university-based or school-based program would likely yield somewhat different characteristics. These types of limited-source clinical samples can be biased toward the inclusion of more severely impaired children (cf. Records & Tomblin, 1994) or toward children whose impairments are most likely to have already provoked a referral from a parent, teacher, or physician (e.g., phonological disorders, telegraphic language, or severely limited expressive language).

After children were classified with regard to language, the study's 96 participants were divided into an exploratory and a confirmatory group. The *exploratory* group included only children who met the study standard for either TD or SLI. It was used to determine the classification accuracy for the SPELT–P2 and to derive the cutoff score that maximally differentiates between children with and without language impairment. The *confirmatory* group was designed to serve two purposes in validating the SPELT–P2. In particular, the statistical procedure used to determine classification accuracy may sometimes overestimate diagnostic accuracy (Katchigan, 1986, p. 373). Thus, it is essential to confirm that the cutoff score obtained with the exploratory sample is equally effective in differentiating between children who meet the study standard for TD or SLI status in a second, confirmatory group. In addition, the children who met only part of the study standard for TD were included in the confirmatory sample to determine the SPELT–P2's utility with children who scored poorly on another language test but whose conversational language was unremarkable. Table 1 contains the demographic information from the children who were assigned to the exploratory and confirmatory groups.

Exploratory group—The exploratory group consisted of children who met the study standard for either TD or SLI status. Thirty-two children (24m, 8f) who met the study standard for a diagnosis of SLI were assigned to the exploratory group. The SLI children included in the exploratory group included all of the 11 children who were enrolled in therapy at the time of the study along with 21 children who were identified through the study procedures. An additional 32 children who met the study standard for the TD group were selected as controls. These TD children were matched to the SLI children based on age in months (range: ±3 months) and sex, as both of these factors are known to affect language performance. When more than 1 TD child could potentially serve as a match for an SLI child, the TD match was determined through random selection.

Table 2 summarizes the assessment results for the exploratory group participants and provides the group differences in test scores (in units of standard deviation, or *d*). As Table 2 indicates, the mean standard scores of the TD group closely approximated the normative mean for the TEEM but were slightly higher than the normative means on the K-ABC–II and TOLD–GU.

On the other hand, the mean standard scores of the SLI group fell more than 1 *SD* below the normative mean on the TOLD–GU and more than 3 *SD*s below the normative mean on the TEEM. Although approximately eight points lower than the TD group mean, the SLI group's mean standard score on the K-ABC–II was within two standard score points of the normative mean, and the majority of the children (28 of 32) scored at or above 85 (1 *SD* below the normative mean). The remaining 4 children scored between 1 and 2 *SD*s below the normative mean.

Confirmatory group—Nineteen children (8m, 11f) who met the study standard for TD and 10 children (8m, 2f) who met the study standard for SLI were assigned to the confirmatory group. The confirmatory group also included the 3 children (2m, 1f) who were judged to have normal language skills by means of qualitative assessment but who scored below 75 on the TEEM. These children enabled an assessment of the SPELT–P2's identification accuracy in cases where the TEEM classification was incongruent with more ecologically oriented information on language status. Their results were considered together with those of the other TD children in the confirmatory group and separately to determine if they differed as a subgroup of TD children. The demographic composition of this group is provided in Table 1.

Table 2 summarizes the performance of the confirmatory group on the study's standardized assessments and provides the group differences in test scores (in units of standard deviation, or *d*). As Table 2 indicates, the mean standard scores for the TD group fell within 1 *SD* of the normative mean, although a few participants scored outside this range of performance. On the other hand, the mean standard scores for the SLI group were greater than 2 *SD*s below the normative mean on the TEEM and within 1 *SD* of the mean on the TOLD–GU. The TD confirmatory group had a mean TEEM score that was approximately five points below that of the TD exploratory group, which is to be expected given the inclusion of equivocal cases of TD status. Similar to results on the TEEM, the mean score on the K-ABC–II was slightly depressed in the TD confirmatory group as compared to the TD exploratory group, though both groups were well within normal limits on this measure.

The mean standard score on the K-ABC–II for the SLI confirmatory group was slightly depressed relative to that of the TD confirmatory group mean, yet the majority of scores (8 of 10 children) were within 1 *SD* below the normative mean, with 1 child scoring within 1 and 2 *SD*s below the mean and another child scoring between 1 and 2 *SD*s above the mean. This is not unusual for samples of children with SLI (Plante, 1998). Because both the SLI exploratory group and the SLI confirmatory group consisted entirely of participants meeting the study standard for SLI status, it is logical that the mean standard scores for these two groups were similar.

Children excluded as participants—The degree to which a study sample can be considered to represent a broader population depends, in part, on the characteristics of the children who are included in it. Conversely, it is important not to exclude children from the sample capriciously. For this study, 25 potential participants were excluded from the study. Most of these participants were disqualified based on the exclusionary criteria for the study: 9 were unable to pass the hearing screening at the time of study; 6 did not speak English as their native language; 6 who were judged to have typically developing language were excluded due to impaired articulation and/or phonology in spontaneous speech, which could have led to poor performance on certain test items; 1 who was judged to have impaired language was excluded due to a history of seizures; and 1 received a K-ABC–II score below 75. Two additional children who were referred to the study did not meet either the test-based criterion or clinical judgment for either TD or SLI. These 2 children who were receiving speech-language therapy services at the time of the study failed to meet the inclusion criteria for the study. The clinicians associated with the study judged these children's language skills to be within the normal range

based on informal conversation with the child and the information provided on the teacher, caregiver, and clinician questionnaires. In addition, these children scored above the cutoff score of 75 on the TEEM. Therefore, neither the formal nor informal measures used indicated that these children were currently language impaired. For this reason, they were excluded from the study.

Materials and Procedure

The SPELT–P2 was administered to all children in the study. For this test, targeted morphosyntactic structures are elicited through presentation of 44 photographs and corresponding verbal prompts. Up to three verbal prompts may be provided. However, examiners may only prompt if the targeted structure has not been attempted, and modeling of the targeted structure is not permitted. Suggestions for additional prompts, which were used by examiners in this study, are included in the SPELT–P2 manual. Responses can be recorded by circling or modifying the scoring sheet's preprinted common responses or by transcribing a child's individual response. Accuracy of responses was determined through comparison with acceptable responses described in the manual.

In addition to the SPELT–P2 and other measures used for participant selection purposes, all children were given the TOLD–GU. This evaluation was used as a measure of receptive language for the purpose of assessing the SPELT–P2's convergent validity. The TOLD–GU was selected for this purpose because its content taps language form that is similar to that contained on the SPELT–P2, but in the receptive domain. Its statistical properties also lend themselves to tests of concurrent validity. Specifically, the subtest lacks either a basement or ceiling effect for the age norms relevant to this study, which prevents score truncation. The normative distribution reflects an adequate range to detect intertest correlations, and the items were selected to show a range of difficulty levels (Newcomer & Hammill, 1997).

All children were assessed at their preschool by a research team that included two certified SLPs and seven research assistants. All examiners were trained to administer and score the tests for this study. Testing for each child was completed within 2 to 23 days. The order of test administration was randomized across children. In a few cases, delivery of a test was carried over from one day to another due to difficulty maintaining a given child's attention to the test or limitations in time. In addition, 1 child had received the K-ABC–II within 12 months of this study, and these prior scores were used instead of readministering the assessment. Interrater reliability data were obtained from pairs of examiners independently scoring a live testing session 26.0%, 48.4%, 21.3%, and 27.8% of the time for the TEEM, K-ABC–II, TOLD–GU, and SPELT–P2, respectively. Point-by-point interrater reliability revealed 97.3%, 98.5%, 99.4%, and 97.7% agreement in scoring these respective tests.

RESULTS

Diagnostic Accuracy

Exploratory group results—The mean standard score on the SPELT–P2 was 108.84 (*SD* = 9.00) for the TD group and 71.5 (*SD* = 12.36) for the SLI group (see Table 2). The TD group mean was somewhat higher than the mean of the test's normative sample, and the SLI group mean was nearly 2 *SD*s below the normative group mean. There was alsominimal overlap in the range of SPELT–P2 scores for the TD and SLI groups, suggesting good discrimination between the groups. This can be observed in Figure 1, which depicts the number of children in each group who received each standard score on the SPELT–P2.

A discriminant analysis was performed to determine classification accuracy for the exploratory group. The discriminant analysis indicated that group classification based on SPELT–P2 scores

was statistically significant, $F(1, 62) = 190.86$, $p < .0001$, $R^2 = .7548$. Discriminant analysis yielded a cutoff score of 87 for differentiating between TD and SLI children. The resulting sensitivity and specificity data for the SPELT–P2 are summarized in Table 3. Sensitivity was 90.6%, and specificity was 100%. These results yield a negative likelihood ratio of 0.094. The positive likelihood ratio cannot be specified because specificity of 100% results in division by 0 (+LR = 0.906/0) in the calculation of the positive likelihood ratio. Therefore, as specificity approaches 100%, the positive likelihood ratio approaches infinity.

Twenty-nine of the 32 children who were originally classified as SLI were correctly classified, with 3 misclassified, resulting in an error rate of 9.4%. The characteristics of the few children who were misclassified as TD are provided in Table 4. As this table indicates, these children had relatively high SPELT–P2 and K-ABC–II test scores compared to their group averages, although the TEEM scores were clearly indicative of impairment. The children were all Hispanic, but other demographic characteristics varied across children.

The statistical analysis provided information not only on the classification of all children, but on the probability that each child was correctly classified into his or her assigned group. This is known as the *posterior probability for classification* (see Perona et al., 2005, for further explanation). If children have a posterior probability for classification that is close to 50%, it is equally likely that they should belong to either the normal or impaired group. The discriminant analysis revealed that 1 of these children had a 64% probability of belonging in the TD group, and conversely a 36% probability of belonging in the SLI group. Thus, this child could be considered bordering on a chance classification, as she had a relatively good probability of being placed in either group. On the other hand, the remaining 2 children with SLI had a 77%–82% probability of belonging to the TD group, and therefore only an 18%–23% probability of belonging to their a priori group. Therefore, these 2 children were strongly misclassified as being TD.

Confirmatory group results—A confirmatory group was used to ensure that the classification accuracy that was obtained with the exploratory group's scores was not based on measurement error or chance classifications. In addition, it allows us to assess classification accuracy for a small subset of children who are found in the TD population but would not meet a strict research standard for language status.

For the confirmatory group, the cutoff score derived from the exploratory group was applied to the scores of confirmatory group participants in order to determine sensitivity, specificity, and positive and negative likelihood ratios. Figure 2 displays the number of confirmatory group participants receiving particular SPELT–P2 standard scores. As this figure indicates, all but 1 child in the confirmatory SLI group scored below the cutoff score. The characteristics of the single misclassified child are provided in Table 4. Conversely, all of the children in the TD confirmatory group scored above the empirically derived cutoff score. It is especially important to note that all 3 children who failed to meet the study standard for TD language due to low scores on the TEEM were nonetheless correctly identified as TD by the SPELT–P2. The SPELT–P2's sensitivity and specificity were calculated based on the scores obtained by confirmatory group participants, and the results are reported in Table 3. Thus, the test's sensitivity and specificity were 90% and 100%, respectively. Calculation of positive and negative likelihood ratios revealed a negative likelihood ratio of .1 and a positive likelihood ratio that approaches infinity.

Convergent Validity—After establishing that an assessment intended for the purpose of identifying language impairment has adequate classification accuracy, further information can be gathered to evaluate the test's construct validity. The data from all of the children who met

the study standard (n = 93) were used to establish convergent validity (and later divergent validity).

To examine evidence of convergent validity, results from the assessment can be compared to scores from tests with the same or related underlying constructs. In this case, scores on the SPELT–P2 were compared with scores on tests that assessed the same underlying construct, namely expressive morphology (TEEM), and a related construct, receptive grammar (TOLD–GU). A Pearson product–moment correlation (r) was calculated between the SPELT–P2 and each of these assessments. Table 5 reports the correlations between the SPELT–P2 and these standardized measures. Analysis revealed that scores on the TEEM accounted for 73.92% of the variance in SPELT–P2 scores (r = .859, p < .0001, r^2 = .7392). Given this strong and significant correlation between performance on the SPELT–P2 and the TEEM, there is support to suggest that the SPELT–P2 scores reflect the construct of expressive morphology.

The content of the SPELT–P2 also appears to require receptive language skills to some extent, given that children must comprehend verbal prompts to produce correct responses. To assess the potential contribution of receptive language skills, performance on the SPELT–P2 and TOLD–GU was also correlated. The TOLD–GU accounted for 19.18% of the variance in SPELT–P2 scores (r = .438, p = .0001, r^2 = .1918), revealing a positive correlation between these two tests as well. This can be contrasted with the correlation between the SPELT–P2 and the TEEM (r = .859), which was used to test convergent validity for expressive language skills. This comparison suggests that receptive language skills play a statistically significant but lesser role than expressive language skills in children's performance on the SPELT–P2.

Divergent Validity—To further explore the construct validity of the SPELT–P2, its divergent validity was assessed to determine whether performance on this test is independent of factors that lie outside the constructs of expressive and receptive language. In particular, SPELT–P2 scores were analyzed with respect to the children's performance on the nonverbal subtests of the K-ABC–II as well as demographic variables. A Pearson product–moment correlation (r) was calculated to determine whether performance on the K-ABC–II accounted for variance on the SPELT–P2. As indicated in Table 5, performance on the SPELT–P2 was positively correlated with K-ABC–II scores, accounting for 16.24% of the variance on the SPELT–P2 (r = .403, p < .0003, r^2 = .1624). This statistically significant correlation suggests that the constructs that are tested by the SPELT–P2 may overlap with that of nonverbal intelligence, despite their apparent differences in content. This relation provides evidence against diagnostic practice known as cognitive referencing in which scores on a language test must be discrepant from a measure of IQ (which assumes that the two measures are statistically independent). In this case, the SPELT–P2 is not statistically independent from at least one measure of nonverbal IQ, the K-ABC–II.

Other aspects of divergent validity were supported by additional analyses. Demographic variables were also assessed to determine whether these factors might bias performance on the SPELT–P2. In particular, t tests were conducted to determine whether the child's sex, exposure to a second language, ethnicity, race, or prior participation in speech and/or language therapy (for children with SLI) resulted in significant differences in SPELT–P2 performance. Table 6 summarizes the results of these t tests as well as the effect size (d), where d refers to the statistical effect's magnitude in units of standard deviation. Therefore, the effect sizes reflect the degree to which score discrepancies reflect differences in demographic variables. As Table 6 indicates, there were no significant score differences on the SPELT–P2 related to any demographic variables. It should also be noted that the d values associated with potential bias factors (d = .03 to .35) are considerably smaller than the d value associated with the difference in performance based on language status (d = 3.02).

Finally, a Pearson product–moment correlation (*r*) was calculated to determine the correlation between primary caregiver education level and performance on the SPELT–P2. Results revealed that this correlation was not statistically significant, accounting for only 4% of the variance ($r = .209$, $p = .06$, $r^2 = .04$).

DISCUSSION

Based on empirical evidence of construct validity, including discriminant analysis and an evaluation of convergent and divergent validity, the SPELT–P2 can be used as a valid measure for classifying 4- and 5-year-old children as having or not having a language impairment. Data indicate that when a cutoff score of 87 is used, this measure has good sensitivity and specificity as well as good positive and negative likelihood ratios. These results were stable across two independent samples of children of this age. The results also revealed that performance on the SPELT–P2 was correlated with performance on the TEEM, providing robust support for the idea that scores on the SPELT–P2 reflect the construct of expressive morphology. Weaker, but statistically significant, correlations with the TOLD–GU suggest that receptive grammar is also represented by the test's construct, although to a lesser degree than expressive language skills. In addition, SPELT–P2 scores were correlated with performance on the K-ABC–II, suggesting that the construct of nonverbal intelligence may overlap with the constructs that are tested by the SPELT–P2. This outcome failed to support one aspect of divergent validity; however, other aspects of divergent validity were supported. SPELT–P2 scores did not differ significantly based on any of the demographic variables examined.

Discriminant analysis yielded a cutoff score of 87 that maximized accurate classification. However, it is important to note that this cutoff score is unique to this version of the SPELT (see Perona et al., 2005; Plante & Vance, 1994, 1995, for cutoff scores for different versions of the SPELT tests). Likewise, the cutoff applies only to use of this test with children of the ages studied here. An additional study focused on other age ranges covered by the test norms would be required to determine the optimal cutoff score for use with children of other ages. Furthermore, the use of alternative, and therefore non-evidence-based cutoff scores, will change the sensitivity of an identification measure (Spaulding et al., 2006). Evidence from the present study confirms the truth of this statement as it applies to the SPELT–P2. Although using a slightly lowered cutoff score of 85 (−1 *SD*) would not result in a decrease in the SPELT–P2's sensitivity for the exploratory group, applying a cutoff score of 78 (−1.5 *SD*s) would misclassify 9 additional children, yielding a sensitivity of 71.9%. Table 7 summarizes changes in sensitivity using arbitrarily defined cutoff scores. Thus, when arbitrary cutoff scores are used instead of the empirically derived cutoff, the SPELT–P2's sensitivity is lowered, which may result in false negatives or misclassifications of children with SLI as being TD. This finding supports the current standard of best practice, which requires the use of empirically derived cutoff scores to maximize the correct classification of language impairment based on children's performance on identification measures (Plante & Vance, 1994; Spaulding et al., 2006). Given that school districts frequently establish eligibility criteria requiring the use of arbitrary cutoff scores with language tests (Spaulding et al., 2006), we strongly encourage clinicians to advocate for a change in such policies to enable accurate classification of children with language impairment based on this standard of best practice.

Even though use of the empirically derived cutoff score of 87 maximizes sensitivity and specificity, a small number of children with SLI who participated in the present study were misclassified by the SPELT–P2. However, an analysis of potential bias did not indicate systemic score differences for demographic factors that showed any degree of similarity across these misidentified children. These included ethnicity (all were Hispanic), sex (2 were female), lack of prior SLP services (true for all 3), or primary caregiver education level (all had college-educated parents). On the other hand, SPELT–P2 scores were positively correlated with

performance on the K-ABC–II; thus, these children's high nonverbal intelligence may account, in part, for their relatively high scores on the SPELT–P2. Consistent with these high scores, the clinicians' qualitative characterization of these children's spontaneous language indicated only mild impairment. Therefore, one simple explanation for these children's misclassification is that they represent the mild end of the SLI continuum.

Further examination of the results reveals that the SPELT–P2 correctly classified 3 TD children as TD, despite their low scores on the TEEM. This difference between test results may simply represent regression to the mean, a phenomenon that describes the tendency of extreme scores to become less extreme when the skill is measured a second time. Still, the results of this and prior studies (Merrell & Plante, 1997; Perona et al., 2005) indicate that the TEEM had good sensitivity and specificity overall. Sensitivity was 100% and specificity was 96.7% for the present sample when test-based classification is compared against a standard of clinical judgment of TD or SLI status.

According to IDEA requirements, identification measures must be valid and reliable for their intended purpose. The present study indicated that the SPELT–P2 meets both of these requirements in terms of identification accuracy and interrater reliability, and therefore provides needed evidence that the SPELT–P2 can be used to determine the presence or absence of a language impairment. Thus, the SPELT–P2 can be added to the increasing list of identification measures that have been empirically validated (see Spaulding et al., 2006). This form of validation provides the evidence base needed to ensure that clinicians can have confidence that they are correctly interpreting test performance when attempting to identify language impairment in children.

ACKNOWLEDGMENT

This work was completed as a master's thesis by the first author. It was supported by National Institute of Deafness and Other Communication Disorders Grant R01 DC04726.

REFERENCES

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Vol. 4th ed.. Washington, DC: Author; 1994.

Anastasi, A. Psychological testing. Vol. 6th ed.. New York: Mcmillan; 1988.

Ballantyne AO, Spilkin AM, Trauner DA. The revision decision: Is change always good? A comparison of CELF–R and CELF–3 test scores in children with language impairment, focal brain damage, and typical development. Language, Speech, and Hearing Services in Schools 2007;38:182–189.

Cook, TD.; Campbell, DT. Quasi-experimentation design and analysis issues for field settings. Boston: Houghton Mifflin; 1979.

Dawson, J.; Eyer, JA.; Fonkalsrud, J. Structured Photographic Expressive Language Test—Preschool: Second Edition. DeKalb, IL: Janelle Publications; 2005.

Dawson, JI.; Stout, CE.; Eyer, JA. Structured Photographic Expressive Language Test: Third Edition. DeKalb, IL: Janelle Publications; 2003.

de Beaman S, Beaman P, Garcia-Peña C, Villa M, Heres J, Córdova A, et al. Validation of a modified version of the Mini-Mental State Examination (MMSE) in Spanish. Aging, Neuropsychology, and Cognition 2004;11(1):1–11.

Dollaghan C. Evidence-based practice in communication disorders: What do we know, and when do we know it? Journal of Communication Disorders 2004;37:391–400. [PubMed: 15231419]

Emmons MR, Alfonso VC. A critical review of the technical characteristics of current preschool screening batteries. Journal of Psychoeducational Assessment 2005;23:111–127.

Glascoe FP. Evidence-based approach to developmental and behavioral surveillance using parents' concerns. Child: Care, Health and Development 2000;26(2):137–149.

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

Individuals With Disabilities Education Act of 1997. Pub. L. No. 105-17, Section SEC. 614. 1997. Retrieved February 22, 2008, from http://www.ed.gov/policy/speced/leg/idea/idea.pdf

Individuals With Disabilities Education Act Amendments of 2004. Pub. L. No. 108–446, 118 Stat. 2705. 2004. Retrieved July 17, 2006, from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=108_cong_public_laws&docid=f:publ446.108

Gutiérrez-Clellen V, Simon-Cereijido G. The discriminant accuracy of a grammatical measure with Latino English-speaking children. Journal of Speech, Language, and Hearing Research 2007;50:968–981.

Katchigan, SK. Statistical analysis. New York: Radius Press; 1986.

Kaufman, AS.; Kaufman, NL. Kaufman Assessment Battery for Children. Vol. 2nd ed.. Circle Pines, MN: AGS; 2004.

Marchman VA, Martínez-Sussman C. Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish. Journal of Speech, Language, and Hearing Research 2002;45:983–997.

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

Messick S. Meaning and values in test validation: The science and ethics of assessment. Educational Researcher 1989;18:5–11.

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

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

Peña E, Gillam R, Malek M, Ruiz-Felter R, Resendiz M, Fiestas C, et al. Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. Journal of Speech, Language, and Hearing Research 2006;49:1037–1057.

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

Plante E. Criteria for SLI: The Stark and Tallal legacy and beyond. Journal of Speech, Language, and Hearing Research 1998;41:951–957.

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

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

Records NL, Tomblin JB. Clinical decision making: Describing the decision rules of practicing speech language pathologists. Journal of Speech and Hearing Research 1994;37:144–156. [PubMed: 8170120]

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

Shipley, KG.; Stone, TA.; Sue, MB. Test for Examining Expressive Morphology. Tucson, AZ: Communication Skill Builders; 1983.

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

Sturner R, Heller J, Funk S, Layton T. The Fluharty Preschool Speech and Language Screening Test: A population-based validation study using sample-independent decision rules. Journal of Speech and Hearing Research 1993;36:738–745. [PubMed: 8377486]

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

Werner, EO.; Kresheck, JD. Structured Photographic Expressive Language Test—II. Sandwich, IL: Janelle Publications; 1983a.

Werner, EO.; Kresheck, JD. Structured Photographic Expressive Language Test—Preschool. Sandwich, IL: Janelle Publications; 1983b.

Zhang X, Tomblin JB. The association of intervention receipt with speech-language profiles and social-demographic variables. American Journal of Speech-Language Pathology 2000;9:345–357.

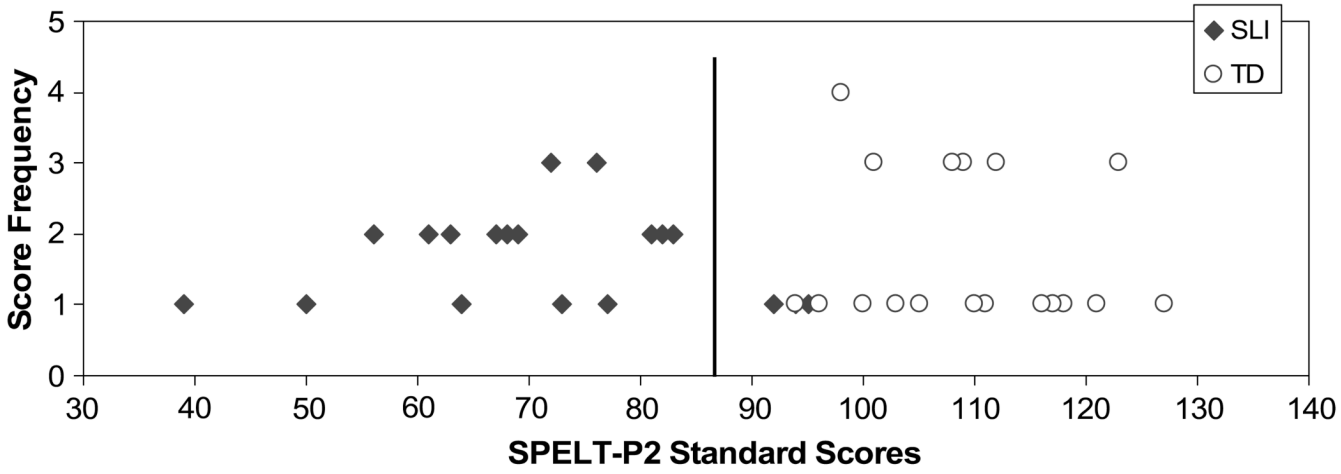

Figure 1.
Scatterplot of SPELT–P2 standard scores obtained by exploratory group participants. The empirically derived cutoff score of 87 maximally differentiated between children in the SLI and TD groups.

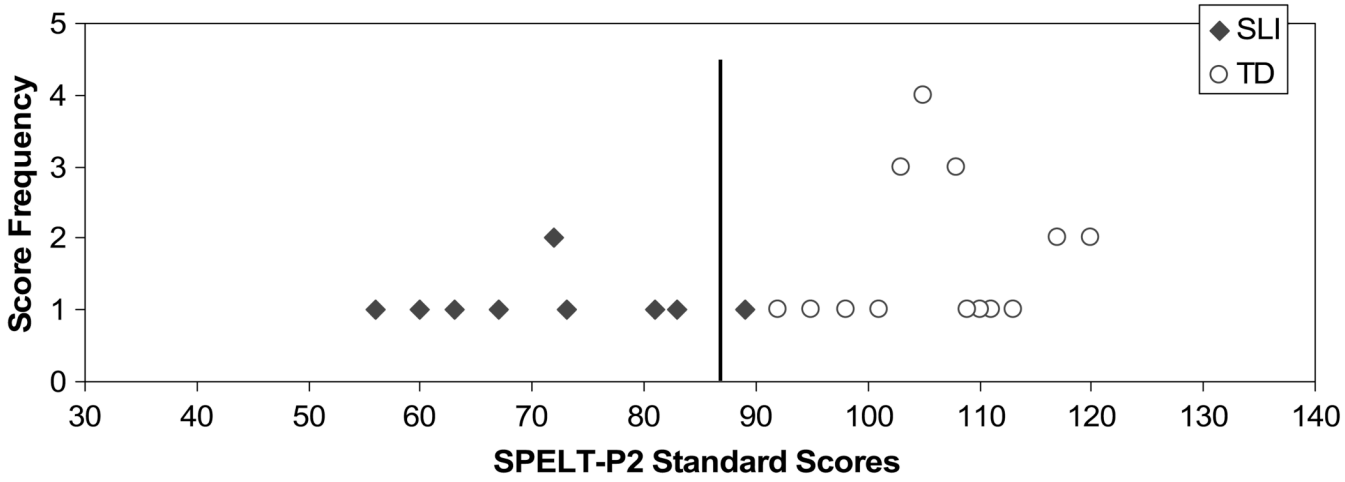

Figure 2.
Scatterplot of SPELT–P2 standard scores obtained by confirmatory group participants. The distribution demonstrates that the cutoff score of 87, derived from the exploratory group data, also accurately differentiated between SLI and TD for a second sample of children.

Table 1

Group demographic information.

	Exploratory TD group	Exploratory SLI group	Confirmatory TD group	Confirmatory SLI group
N	32	32	22	10
Male	24	24	10	8
Female	8	8	12	2
Mean age (in months)	54.7	55.25	60.3	57.8
Age range (in months)	49–65	48–68	50–67	49–64
Race				
White	21	16	14	8
Asian American	1	1	1	0
Black/African American	1	6	0	1
American Indian/Alaska Native	0	0	1	0
Hawaiian/Pacific Islander	0	0	0	0
Multiple races	4	3	2	0
Not reported	5	6	4	1
Ethnicity				
Hispanic or Latino	7	16	7	3
Not Hispanic or Latino	5	3	2	2
Not reported[a]	20	13	13	5
Mean maternal or primary caregiver education level (in years)	14.4	14.0	13.9	11.3
Range of maternal or primary caregiver education level (in years)	11–17	11–17	11–17	10–14

Note. TD = typically developing, SLI = specific language impairment.

[a]The high rate of nonreporting of ethnicity may reflect the lack of a distinction between race and ethnicity among the general public such that individuals who select a race category often do not also select an ethnicity category and vice versa.

Table 2

Exploratory and confirmatory group performance on standardized assessment measures.

	TD group			*SLI group*			
	M	**SD**	***Range***	**M**	**SD**	***Range***	**d**
Exploratory group							
SPELT–P2 ($\bar{X}$ = 100, *SD* = 15)	108.84	9.00	94–127	71.5	12.36	39–95	3.02
TEEM ($\bar{X}$ = 101, *SD* = 15)	100.50	12.35	77–128	44.50	13.78	10–69	4.06
TOLD–GU ($\bar{X}$ = 10, *SD* = 3)	10.84	2.08	8–16	8.00	2.46	4–15	1.15
K-ABC–II ($\bar{X}$ = 100, *SD* = 15)	106.94	10.25	85–128	98.66	13.00	79–130	0.64
Confirmatory group							
SPELT–P2 ($\bar{X}$ = 100, *SD* = 15)	107.09	7.42	92–120	71.6	10.52	56–89	3.37
TEEM ($\bar{X}$ = 101, *SD* = 15)	94.86	18.28	55–130	36.20	14.25	4–66	3.21
TOLD–GU ($\bar{X}$ = 10, *SD* = 3)	9.82	2.82	6–17	8.5	2.46	6–13	0.47
K-ABC–II ($\bar{X}$ = 100, *SD* = 15)	102.00	12.06	79–130	96.5	12.03	77–118	0.46

Note. SPELT–P2 = Structured Photographic Expressive Language Test—Preschool: Second Edition (Dawson, Eyer, & Fonkalsrud, 2005), TEEM = Test for Examining Expressive Morphology (Shipley, Stone, & Sue, 1983), TOLD–GU = Grammatical Understanding subtest of the Test of Language Development—Primary: Third Edition (Newcomer & Hammill, 1997), K-ABC–II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004).

Table 3

SPELT–P2 sensitivity and specificity data for the exploratory and confirmatory groups.

	Group categorization based on clinical judgment and TEEM scores			
Group categorization based on discriminant analysis	***SLI* (n = *32*)**		***TD* (n = *32*)**	
Exploratory group				
SLI	29	90.6%	0	0.0%
TD	3	9.4%	32	100.0%
Confirmatory group				
SLI	9	100.0%	0	0.0%
TD	1	4.3%	22	95.6%

Table 4

Characteristics of the children with SLI who were misidentified as TD on the SPELT–P2.

	Test standard scores				Demographics			
Child	**SPELT–P2**	**TEEM**	**TOLD–GU**	**K-ABC–II**	**Age (yrs;mo)**	**Sex**	**Race/ethnicity**	**Parent education**
Exploratory group								
1	92	41	8	113	4;5	Female	Multiracial/Hispanic	16
2	94	42	10	109	4;3	Male	NR/Hispanic	16
3	95	41	8	108	5;1	Female	White/Hispanic	NR
Confirmatory group								
1	89	66	9	87	5;1	Male	White/NR	12

Note. NR = not reported.

Table 5

Correlation (*r*) between the SPELT–P2 and other standardized measures.

	TEEM	TOLD–GU	K-ABC–II
SPELT–P2	.859*	.438*	.403*

* $p < .0001$.

Table 6

Effect of demographic variables on SPELT–P2 performance.

	N	t	p	*Effect size* d
Sex	64 m, 29f	1.55	.1252	.35
Second language exposure	83 English only 10 other languages[a]	.72	.4746	.19
Ethnicity	33 Hispanic, 12 nonhispanic	.93	.3554	.26
Race	59 White, 21 Nonwhite	.92	.3595	.20
SLI treatment status	11 in therapy, 28 no therapy	.10	.9247	.03
	N	r	p	*Effect size* r^2
Maternal or primary caregiver education level	81	.209	.06	.04

[a]The small *n* for other languages warrants caution in interpreting this result.

Table 7

Changes in sensitivity based on the use of arbitrary cutoffs (n = 32).

Cutoff	Number of exploratory SLI children misclassified as TD	Sensitivity
−1 *SD*	3	90.6%
−1.5 *SD*	9	71.9%
−2 *SD*	17	46.9%