LSHSS

Research Article

Concurrent Validity of the Fluharty Preschool Speech and Language Screening Test–Second Edition at Age 3: Comparison With Four Diagnostic Measures

Sarita Eisenberg,[a] Kristen Victorino,[b] and Sarah Murray[c]

Purpose: The aim of this study was to examine the concurrent validity of the Fluharty Preschool Speech and Language Screening Test–Second Edition (Fluharty-2; Fluharty, 2001) for mass screenings of language at age 3 years.
Method: Participants were sixty-two 3-year-old children, 31 who had failed and 31 who had passed the Fluharty-2. Performance on the screening was compared to 4 diagnostic measures: Structured Photographic Expressive Language Test–Preschool, Second Edition; mean length of utterance in morphemes (MLU_m), finite verb morphology composite, and Index of Productive Syntax (IPSyn).
Results: Children who failed the Fluharty-2 scored significantly lower on each of the diagnostic measures than children who passed the Fluharty-2, but the effect size for MLU_m was small. Scores on the Fluharty-2 were significantly correlated with scores on the diagnostic measures. There was significant agreement for pass/fail decisions between the Fluharty-2 and diagnostic measures only for IPSyn. However, even for the IPSyn, the agreement rate for passing was only moderate (80%) and the agreement rate for failing was only fair (68%).
Conclusion: The Fluharty-2 showed limited agreement for pass/fail decisions with all 4 of the diagnostic measures. There was reason to question the validity of 2 of the diagnostic measures—Structured Photographic Expressive Language Test–Preschool, Second Edition and MLU_m—for diagnosing language impairment in 3-year-old children. However, there were no such concerns about finite verb morphology composite or IPSyn to account for the limited agreement. Thus, it seems reasonable to conclude that the Fluharty-2 would refer both too few at-risk children and too many nonrisk children for a follow-up assessment, making it an inefficient tool for mass screenings of language.

Speech and language screenings are designed to separate children with potential deficits from children with typical development so that they can receive a follow-up assessment to either confirm or rule out a disorder (Law, Boyle, Harris, Harkness, & Nye, 1998). Screenings are typically administered to all children in a particular setting, such as a preschool or clinic. Screenings result in dichotomous outcomes—failing a screening results in a referral for further evaluation; passing a screening results in dismissal with no further services. Recent estimates suggest that over 40% of children with language impairment (LI) do not receive intervention services (Black, Vahratian, & Hoffman, 2015). The need for high-quality language screening measures is critical for identifying children with LI so that they can benefit from language intervention before entering school.

An evaluation of accuracy for a screening test reflects the extent that performance on the screening test aligns with performance on the specific assessment procedures used for the follow-up diagnostic evaluation (Law et al., 1998). There is widespread agreement that assessments to diagnose LI should include a norm-referenced standardized test (McCauley, 2001; Merrill & Plante, 1997; Paul, Norbury, & Gosse, 2018) and inclusion of such tests is mandated for school district assessments (Blosser, 2012). Assessment to diagnose LI can also include measures based

[a]Department of Communication Sciences and Disorders, Montclair State University, NJ
[b]Department of Communication Disorders and Sciences, William Paterson University of New Jersey, Wayne
[c]Passaic County Technical Institute, Wayne, NJ

Correspondence to Sarita Eisenberg: eisenbergs@mail.montclair.edu

Editor-in-Chief: Shelley Gray
Editor: Kerry Ebert

Received September 17, 2018
Revision received December 6, 2018
Accepted April 26, 2019
https://doi.org/10.1044/2019_LSHSS-18-0099

Disclosure: The authors have declared that no competing interests existed at the time of publication.

on language sampling analysis (LSA; Kelly & Rice, 1986; McCauley, 2001). The most frequently used LSA measure is mean length of utterance in morphemes (MLU_m; Loeb, Kinsler, & Bookbinder, 2000).

When conducting a screening, underreferral—that is, not referring children for further evaluation who would have been subsequently diagnosed as having an LI—is a more serious problem than overreferral, that is, referring too many children for further assessment. It is therefore crucial that few, if any, children pass the screening who would fail the follow-up diagnostic assessment (termed *false negatives*), as this means that those children would not receive needed services. In order to accomplish this, we are willing to have some children fail the screening but pass the follow-up diagnostic evaluation (termed *false positives*). A good screening will have very few, if any, false negatives while minimizing the number of false positives.

There are limited options for standardized language screenings. The current report examined one popular screening test, the Fluharty Preschool Speech and Language Screening Test–Second Edition (Fluharty-2; Fluharty, 2001).

Fluharty-2

The Fluharty-2 includes four subtests: For Following Directives and Answering Questions (FDAQ), the child responds to 10 questions and commands by either answering the question or carrying out the command; for Repeating Sentences (RS), the child repeats 10 sentences produced by the examiner; for Describing Actions, the child produces sentences to describe pictured actions; and for Sequencing Events, the child explains how to do something such as making a sandwich. The four subtests are combined into three language composite scores. All four subtests are combined into a General Language Quotient (GLQ). An Expressive Language Quotient (ELQ) is derived from the Describing Actions and Sequencing Events subtests**.** A Receptive Language Quotient (RLQ) is derived from the FDAQ and RS subtests.

The Fluharty-2 was normed on 705 children, including one hundred thirty-six 3-year-olds. The normative sample was balanced for gender. The racial composition was 78% White and 15% African American. Ten percent of the children were Hispanic. The test manual did not provide information about socioeconomic status. According to the test manual, 94% of the normative sample had either no disability (82%) or a speech impairment (12%); the remainder of the normative sample included 2% with learning disability, 1% with mental retardation, and 3% unspecified.

The test manual reported data for three types of reliability. Internal reliability at age 3 years was in the acceptable or higher range (i.e., at or above .80) for all three composite scores (McCauley, 2003). Test–retest reliability, based on 3- and 5-year-old children, was in the high range for the GLQ and RLQ but only in the fair range for the ELQ (McCauley, 2003). Interexaminer reliability was also in the high range (McCauley, 2003). However, this was based on comparing results for two of the publishing company staff members who independently scored responses from 30 children ranging in age from 3 to 6 years, and it did not take into account possible differences in test administration across examiners (Hurford, 2003). In addition, it was not clear whether this correlation was based on total scores or individual responses (McCauley, 2003).

The test manual evaluated concurrent validity based on the Test of Language Development–Primary: Third Edition (TOLD-P3; Newcomer & Hammill, 1997) for a sample of 23 children. The correlation for receptive language—between the TOLD-P3 Listening Composite and the RLQ—was .82. The correlation for expressive language—between the TOLD-P3 Speaking Composite and the ELQ—was .88. The correlation with the TOLD-P3 Spoken Language Quotient, which combines scores from all six core subtests of the TOLD-P3, was .76 for the GLQ, .86 for the ELQ, and .89 for the RLQ. Correlations were thus mostly moderate to high with the exception of the correlation between the TOLD-P2 Spoken Language Quotient and the GLQ (McCauley, 2003).

Of concern was the use of the TOLD-P3 as the criterion measure for establishing concurrent validity. In their review of normative samples for standardized language tests, Peña, Spaulding, and Plante (2006) noted that the TOLD-P3 includes children with LI in the normative sample and that this might reduce the test's ability to differentiate between children with LI and children with typical language (TL). Consistent with this concern, Spaulding, Plante, and Farinella (2006) reported an insufficient mean difference on the TOLD-P3 between scores for children with LI and children with TL. In addition, the TOLD-P3 was developed for ages 4;0–8;11 (years;months), and so the comparison did not include 3-year-old children. We lack information about concurrent validity of the Fluharty-2 for this younger age group.

Two reviews of the Fluharty-2 in the Buros Mental Measurement Yearbook were generally positive. Hurford (2003) concluded that the Fluharty-2 showed acceptable psychometric properties with the exception of interexaminer reliability. McCauley (2003) noted some reliability coefficients that were below an acceptable level and that additional data about validity would be desirable but concluded that the Fluharty-2 would be a reasonable choice for screening language. We found no empirical studies about the Fluharty-2, but several studies investigated the original version, which we will refer to as the Fluharty-1 (Fluharty, 1978). These studies evaluated the extent of agreement between passing and failing the Fluharty-1 and performance on either a standardized test or an LSA measure, Developmental Sentence Scoring (Lee, 1974). Across the studies, the Fluharty-1 showed low agreement for failing with the diagnostic measures (i.e., too many children who passed the Fluharty-1 subsequently failed the diagnostic measures) while showing good agreement for passing (Blaxley, Clinker, & Warr-Leeper, 1983; Goldstein, 1994; Sturner, Heller, Funk, & Layton, 1993). This means that screening with the Fluharty-1 would have resulted in underreferrals for follow-up evaluation relative to those specific diagnostic measures. We wanted to know whether the revised version would perform better.

The Fluharty-1 was designed to test 4- to 6-year-old children. However, school districts are now testing 3-year-old children in order to determine eligibility for special education services at the preschool level, and the Fluharty-2 was modified to include this younger age group. The current study focused on use of the revised version for screening language at age 3 years.

Purpose

The current report examined the concurrent validity of the Fluharty-2 by comparing performance on the Fluharty-2 to performance on four diagnostic measures. In her review of screening tests, Goldstein (1994) described several methods for establishing the validity of a screening test: (a) comparing the screening test's referral rate with the prevalence rate of the disorder, (b) examining the correlation between performance on the screening test and performance on the diagnostic measures, and (c) examining the agreement for pass/fail decisions between the screening test and the diagnostic measures. Because we did not have a random population sample, we could not make a valid comparison with the prevalence rate. We used the latter two methods. In addition, we examined whether children who failed the Fluharty-2 scored lower on the diagnostic measures than the children who passed the screening. We asked the following questions:

1. Did children who failed the Fluharty-2 score significantly lower on each of the diagnostic measures than children who passed?
2. Was performance on the Fluharty-2 significantly correlated with performance on each of the diagnostic measures?
3. To what extent did the pass/fail decisions for the Fluharty-2 agree with pass/fail decisions for each of the diagnostic measures and for the overall diagnostic assessment?

Method

Participants

Participants included 62 children drawn from a larger sample of 3-year-old children who had participated in a study about language production. Participants had been recruited through preschool programs, pediatricians, and speech-language pathologists (SLPs) in the suburban New Jersey area, as well as through online announcements. Approval for this research was granted by the Montclair State University Institutional Review Board, and parents gave consent for their child's data to be used in additional studies.

Thirty-one of the participants scored below the cutoff for one or more of the language quotients of the Fluharty-2 (FAIL group), per the criteria in the test manual. These participants were matched for age and gender with 31 other participants who had passed all three quotients of the Fluharty-2 (PASS group). The sample included 36 boys and 26 girls ranging in age from 3;0 to 3;11 (M_{age} = 3;6). Both groups were balanced for gender, with 13 girls and 18 boys. The mean age in months was 40.42 (*SD* = 3.54) for the FAIL group and 40.55 (*SD* = 3.49) for the PASS group. A one-way analysis of variance (ANOVA) revealed no difference in age between groups, $F(1, 60) = 0.021$, $p = .886$. In contrast, the PASS group obtained a significantly higher mean score on the Fluharty-2 GLQ ($M = 103.03$, $SD = 7.41$) than the FAIL group ($M = 85.77$, $SD = 6.23$), $F(1, 60) = 98.35$, $p < .001$, $\eta_p^2 = .621$. Based on Cohen (1992), we interpreted the magnitude of effect size (i.e., η_p^2, partial eta squared) with the following criteria: $.01 \leq \eta_p^2 < .09$, small effect size; $.09 \leq \eta_p^2 < .25$, medium effect size; and $\eta_p^2 \geq .25$, large effect size. Thus, the difference between the PASS and FAIL groups on the GLQ had a large effect size. The age and GLQ data are provided in Table 1.

All children were monolingual English speaking. Participants were excluded if they were exposed to African American English or another language in the home, based on parent report, because of concerns about the validity of the criterion measures for nonmainstream speakers. All participants passed a hearing screening at 25 dB for the frequencies of 500, 1000, 2000, and 4000 Hz. There were no concerns about cognition reported by any of the parents, and all children passed the odd-item-out task of the Reynolds Intellectual Screening Test (Kamphaus & Reynolds, 2003). Eighty-nine percent of the participants were from families in which both parents had a college degree, and 11% (four in the FAIL group and three in the PASS group) had one parent with a college degree and one with a high school degree. The racial and ethnic distribution based on self-identification by the parent was 66% White, 13% African American, 10% Asian, and 11% Hispanic. The racial and ethnic distribution of the PASS and FAIL groups is shown in Table 2.

Data Collection

Each child completed the Fluharty-2 and a diagnostic language battery that included the Structured Photographic Expressive Language Test–Preschool, Second

Table 1. Mean (*SD*) age and Fluharty Preschool Speech and Language Screening Test–Second Edition (Fluharty-2) scores.

Fluharty-2 rating	Age in months	Fluharty-2 GLQ
PASS (*N* = 31)	40.55 (3.49)	103.19 (7.58)
FAIL (*N* = 31)	40.42 (3.54)	85.77 (6.24)

Note. GLQ = General Language Quotient.

Table 2. Race/ethnic composition of the PASS and FAIL groups.

Race/ethnicity	PASS (*N* = 31)	FAIL (*N*= 31)
White	24 (77%)	18 (58%)
African American	2	5
Asian	2	4
Hispanic White	3	4

Edition (SPELT-P2; Dawson et al., 2005) and a 30-min language sample. The Fluharty-2 was always administered first to reflect the order in which testing would be administered in clinical practice. However, all participants were administered the diagnostic language battery, regardless of their performance on the screening test. Administration order for the SPELT-P2 and language sample was randomized.

Language samples were elicited during 30 min of play with a parent. Parents were instructed to follow their child's lead and to avoid asking questions as much as possible. Each child was provided with five different sets of toys. Each participant selected the first set of toys, and an additional toy set was introduced every 6 min. The order of presentation for the toy sets was randomized, and all prior toys remained available for the child to play with when new toys were introduced. The entire session was audio- and video-recorded. Language samples were analyzed for $\mathrm{MLU_m}$, finite verb morphology composite (FVMC; Bedore & Leonard, 1998), and Index of Productive Syntax (IPSyn; Scarborough, 1990).

Transcription

Utterances were transcribed by trained research assistants (RAs). Morpheme coding followed the conventions of Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2003–2007). However, utterance segmentation was in phonological units (p-units) to allow comparison of $\mathrm{MLU_m}$ to the normative data from Rice et al. (2010) and of IPSyn to the normative data in Scarborough (1990), both of which were based on p-unit segmentation. p-units are based on intonation and pausing and allow up to two independent clauses per utterance. Utterances that could not be fully transcribed after three listenings were marked as unintelligible. Transcription accuracy for all samples was checked using a consensus procedure based on Shriberg, Kwiatkowski, and Hoffman (1984). Samples were transcribed by one RA, checked by a second RA, and then rechecked by the first author. Disagreements were resolved, or the utterance was marked as unintelligible and excluded from further analysis. The same consensus procedure was used for segmentation and morpheme coding. We also calculated interrater agreement between the RAs and the first author for transcription, segmentation, and morpheme coding based on 10 samples (i.e., approximately 15% of the samples). Interrater agreement was 97%.

Criterion Measures

The current report compared the Fluharty-2 to four language measures. We included a standardized test because norm-referenced standardized tests are typically required by school districts for determining eligibility for speech-language services (Blosser, 2012). We included $\mathrm{MLU_m}$ because it is the most frequently used LSA procedure (Loeb et al., 2000) and because school districts typically require a second norm-referenced measure (Blosser, 2012). However, the validity of $\mathrm{MLU_m}$ for diagnosing LI in young children has been questioned (Eisenberg, Fersko, & Lundgren, 2001). Because of this concern, we also included two other LSA measures: FVMC, which has been shown to have good diagnostic accuracy at age 3 years (Bedore & Leonard, 1998; Guo & Eisenberg, 2014), and IPSyn, which has been shown to be more sensitive to LI than $\mathrm{MLU_m}$ at age 3 years (Rescorla, Dahlsgaard, & Roberts, 2000).

Our decision to include only expressive measures reflected the predominately expressive content of the Fluharty-2. Although the Fluharty-2 includes an RLQ, one of the two subtests that make up that quotient is the RS task. Sentence repetition tasks, however, are typically classified as expressive tasks (Pawlowska, 2014; Seeff-Gabriel, Chiat, & Dodd, 2010; Semel, Wiig, & Secord, 2004), and low performance on the RLQ by participants in the current study largely reflected performance on this expressive task rather than on the FDAQ subtest. Of the children who failed the Fluharty-2, all but one child earned fewer of the points on the RS task than on FDAQ—nine children earning 0 points, 15 earning 1 point, and four earning 2 points out of the 10 total points on the RS task. It is also noteworthy that a study by Greenslade, Plante, and Vance (2009) reported a significant correlation between the SPELT-P2 and the TOLD-P3 Grammatic Understanding subtest (r = .438). Thus, the SPELT-P2 taps receptive and expressive skills, albeit to a lesser extent.

SPELT-P2

The SPELT-P2 is a 40-item test for production of grammatical morphemes and syntactic structures. Items are elicited by showing the child photographs and providing a verbal prompt (e.g., asking a question or starting a sentence to be completed). The test was normed on 1,747 children, including three hundred eighty-eight 3-year-olds. The racial composition for 3-year-olds included 77.3% White and 9.0% African American, 7% Hispanic, and 6.7% other. Maternal education level was 2% having some high school, 11% high school graduates, 19% having some college, 38% college graduates, and 29% unknown.

Several types of reliability were reported in the test manual. Applying the same criteria as were used by McCauley (2003), internal consistency reliability was in the acceptable range (.882 for the entire standardization sample; .877 and .882 for younger and older 3-year-olds, respectively). Test–retest reliability was not reported for 3-year-olds but was high (.96) based on scores from sixty-two 4-year-old children. Interexaminer reliability was also high (.99 overall and 1.0 at age 3 years) based on comparing two independent ratings of test responses from seventy-four 3-year-old children. However, this measure only evaluated reliability for scoring the test. The authors did not evaluate interexaminer reliability for test administration.

To evaluate content validity, the authors noted the overlap between grammatical forms included on the SPELT-P2 and the IPSyn (Scarborough, 1990): The test includes nine out of 17 forms on the IPSyn Verb Phrase subscale, 10 out of 12 forms on the IPSyn Question/Negation subscale items, and complex sentence forms on the IPSyn Sentence Structure subscale while including fewer of the forms from the IPSyn Noun Phrase subscale. The authors further noted that the inclusion of more verb form items reflects the relative difficulty of these forms for children with LI.

Concurrent validity was evaluated by comparing the SPELT-P2 to the Syntax Construction subtest of the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999) for a sample of 61 children (age not specified). The correlation was acceptable (.86). Construct validity was evaluated by showing that scores significantly increased with age. The test manual did not address diagnostic accuracy.

A review of the SPELT-P2 by Towne (2010) in the Buros Mental Measurement Yearbook concluded that the test has good reliability and validity. Hutchinson (2010), however, noted limitations in the evaluation of test–retest and interexaminer reliability as well as in construct validity. With regard to the latter, Hutchinson noted that the authors did not statistically evaluate the age difference in scores and did not evaluate differences in scores between children with LI and children with TL. A study by Greenslade et al. (2009) investigated concurrent validity and diagnostic accuracy for the SPELT-P2. Participants included children aged 4;0–5;6 (*M* = 4;9). The correlation between SPELT-P2 and another expressive grammar test—Test for Examining Expressive Morphology (Shipley, Stone, & Sue, 1983)—was high (.866). Diagnostic accuracy at a cutoff score of 87 was good for both an exploratory study and a confirmatory study, with both sensitivity and specificity above the 90% level recommended by Plante and Vance (1995). This was important to determine because the normative sample included approximately 2.5% children with LI, which could potentially have reduced the test's ability to differentiate children with and without LI (Peña et al., 2006).

MLU$_m$

MLU$_m$ has long been suggested as an index of grammatical development (Miller & Chapman, 1981). Guo and Eisenberg (2015) recommended that conversational samples be at least 90 utterances to ensure acceptable reliability for MLU$_m$, and this recommendation was followed in the current study. MLU$_m$ was calculated on the first 100 complete and intelligible utterances using the SALT program. *z* scores (i.e., number of standard deviations from the mean) were calculated based on normative data (i.e., means, standard deviations) from Rice et al. (2010).

Several studies have reported a positive correlation between age and MLU both for children with TL and for children with LI (Conant, 1987; deVilliers & deVilliers, 1973; Klee, Schaffer, May, Membrino, & Mougey, 1989a; Miller & Chapman, 1981; Scarborough, Wyckoff, & Davidson, 1986). Analyzing data from Klee, Schaffer, May, Membrino, and Mougey (1989b), Eisenberg et al. (2001) reported specificity at 96% and sensitivity at 54%, at a −1.5-*SD* cutoff. They concluded that "while we may not be able to conclude that MLUs above a certain level mean that language is normal, we may be able to use a low MLU as evidence of language impairment" (p. 338).

FVMC

FVMC computes the overall percentage of correct use in obligatory contexts of four morphemes that mark verb tense and agreement—third-person singular present *–s*, regular past tense *–ed*, and copula and auxiliary *BE* (i.e., *am, are, is, was, were*)—with a single measure. The number of correct uses of the four verb tense morphemes is divided by the total number of obligatory contexts, and the resultant quotient is multiplied by 100% to obtain a percentage. Computation of FVMC was based on the entire sample and followed the procedure in Eisenberg and Guo (2016). Usage and obligatory contexts for each of the four morphemes were calculated using SALT and then summed. FVMC was then calculated by dividing the summed uses of the morphemes by the summed obligatory contexts and then multiplying by 100% to obtain a percentage. *z* scores for FVMC were calculated based on normative data from Guo and Eisenberg (2014). A study by Guo and Eisenberg (2014) reported acceptable specificity (89%) and sensitivity (83%) for this measure for 3-year-old children.

IPSyn

The IPSyn (Scarborough, 1990) is a type-based measure of grammar. Up to 2 points are assigned for each of 60 syntactic structures divided into four subscales. The appendix to the 1990 publication serves as the coding manual, providing definitions and examples for each item with further clarification about item scoring provided by Altenberg, Roberts, and Scarborough (2018). IPSyn scores were calculated by hand on the same 100 utterances as MLU$_m$. The IPSyn score was calculated by summing all of the points received on the four subscales. *z* scores for IPSyn were calculated based on normative data from Scarborough (1990). Although not examining diagnostic accuracy, a study by Rescorla et al. (2000) does suggest that IPSyn is more sensitive to LI than MLU$_m$. These authors compared performance on MLU$_m$ and IPSyn for children who were identified as late talkers at age 2 years and continued to show language deficits at ages 3 and 4 years. At age 3 years, 66% of the children scored below a −1.25-*SD* cutoff on the IPSyn compared to 59% for MLU$_m$; at age 4 years, 62% scored below the cutoff on the IPSyn compared to 29% for MLU$_m$.

Analyses

Multivariate ANOVAs were adopted to examine whether scores on SPELT-P2, MLU_m, IPSyn, and FVMC differed between children who passed and failed the Fluharty-2. We again used partial eta squared (η_p^2) to quantify the effect size or magnitude of the differences and interpreted effect size as small ($.01 \leq \eta_p^2 < .09$), medium ($.09 \leq \eta_p^2 < .25$), or large ($\eta_p^2 \geq .25$). Pearson product–moment correlations were used to evaluate the extent to which performance on the Fluharty-2 correlated with the SPELT-P2, MLU_m, and FVMC. We interpreted correlation coefficients as small ($.1 \leq r < .3$), medium ($.3 \leq r < .5$), or large ($r \geq .5$) following Cohen (1988).

Chi-square contingency tables were used to determine whether children who had passed or failed the Fluharty-2 GLQ scored below or above the cutoff on SPELT-P2, MLU_m, IPSyn, and FVMC. We used the cutoff standard score of 87 on the Fluharty-2 based on the test manual. For the criterion measures, we used a cutoff of 1.25 *SD*s below the mean (−1.25 *SD*). This cutoff was chosen as it corresponds roughly to the 10th percentile in a normal distribution; performance below that level is considered by many to be below average or disordered (e.g., Paul et al., 2018).

We then compared the degree of agreement for the pass/fail decisions between the Fluharty-2 and SPELT-P2, MLU_m, FVMC, and IPSyn using Cohen's kappa (κ). Following Landis and Koch (1977), we interpreted the degree of agreement as fair ($.21 \leq \kappa < .40$), moderate ($.41 \leq \kappa < .60$), substantial ($.61 \leq \kappa < .80$), or almost perfect ($\kappa \geq .80$).

To further explore the relationship between the RLQ and pass/fail designations, we ran the statistical analyses in two configurations; first, as per the Fluharty-2 manual, with any child who failed the GLQ, ELQ, or RLQ included in the FAIL group and then with children who failed only the RLQ removed from the FAIL group. There were no differences in the pattern of results. This demonstrated that low performance on the RLQ did not impact agreement between Fluharty-2 GLQ scores and the expressive diagnostic measures. Thus, we made the decision to implement pass/fail decisions as per the Fluharty-2 manual and to use the GLQ in our examination of pass/fail agreement.

Results

Question 1: Score Comparison Between PASS and FAIL Groups

The first research question addressed was whether children who failed the Fluharty-2 scored significantly lower on the SPELT-P2, MLU_m, FVMC, and IPSyn than children who passed the Fluharty-2. A multivariate ANOVA was conducted, with SPELT-P2 standard scores, MLU_m *z* scores, FVMC *z* scores, and IPSyn *z* scores as dependent variables and PASS/FAIL on the Fluharty-2 as the independent variable. *z* scores were used for MLU_m, FVMC, and IPSyn results since standard scores were not available for these measures and using raw scores could misrepresent age-expected differences within groups. Table 3 shows group data for each diagnostic measure. The analysis revealed significant group differences for all four measures. The FAIL group received lower standard scores on the SPELT-P2, $F(1, 60) = 38.297$, $p < .001$, $\eta_p^2 = .390$; lower *z* scores for MLU_m, $F(1, 60) = 4.844$, $p = .032$, $\eta_p^2 = .075$; lower *z* scores for FVMC, $F(1, 60) = 6.468$, $p = .014$, $\eta_p^2 = .097$; and lower *z* scores for IPSyn, $F(1, 60) = 13.418$, $p = .001$, $\eta_p^2 = .183$. Effect size varied across the measures, with a large effect size observed for the SPELT-P2 score difference, a medium effect size for FVMC and IPSyn score differences, and a small effect size for the MLU_m score difference between the PASS and FAIL groups.

Question 2: Correlation Between Fluharty-2 and Diagnostic Scores

Next, we considered whether performance on the Fluharty-2 was significantly correlated with SPELT-P2, MLU_m, IPSyn, and FVMC (see Table 4). Pearson product–moment correlations were used to evaluate the extent to which the Fluharty-2 GLQ score correlated with the standard or *z* scores on each of the diagnostic measures. All measures were significantly intercorrelated, indicating that low performance on the screening measure was correlated with low performance on each of the criterion measures. Based on Cohen (1988), the correlation between GLQ and SPELT-P2 was interpreted as "large" ($r = .722$, $p < .001$). The correlations between GLQ and MLU_m, IPSyn, and FVMC scores all fell in the "medium" range.

Question 3: Pass–Fail Agreement Between Fluharty-2 and Diagnostic Measures

We then determined the extent to which the pass/fail decisions for the Fluharty-2 agreed with pass/fail decisions for SPELT-P2, MLU_m, FVMC, and IPSyn using −1.25 *SD* as the cutoff score. First, new variables were created to reflect whether a participant passed or failed each diagnostic measure. Pass/fail decisions between the Fluharty-2 and each of the diagnostic measures were compared via chi-square cross-tabulations (see Table 5), and degree of agreement was measured via Cohen's kappa, with the degree of agreement interpreted as fair ($.21 \leq \kappa < .40$), moderate ($.41 \leq \kappa < .60$), substantial ($.61 \leq \kappa < .80$), or almost perfect ($\kappa \geq .80$).

SPELT-P2

Only two participants in the FAIL group, and no participants in the PASS group, failed the SPELT-P2. Overall agreement between the SPELT-P2 and the Fluharty-2 was 53%, with 100% agreement for passing but only 7% agreement for failing. This resulted in a Pearson chi-square value (χ^2) of 2.067 ($p = .151$). Thus, the degree of agreement was not statistically significant and was interpreted as very low ($\kappa = .065$).

Table 3. Mean (*SD*) of raw scores for the diagnostic measures.

Fluharty-2 rating	SPELT-P2	MLU$_{\text{m}}$	IPSyn	FVMC
PASS (*N* = 31)	25.06 (4.02)	4.19 (0.67)	74.77 (7.26)	92.91% (7.70%)
FAIL (*N* = 31)	17.39 (5.22)	3.81 (0.68)	66.42 (8.38)	81.21% (19.40%)

Note. SPELT-P2 = Structured Photographic Expressive Language Test–Preschool, Second Edition; MLU$_{\text{m}}$ = mean length of utterance in morphemes; IPSyn = Index of Productive Syntax; FVMC = finite verb morphology composite.

MLU$_{\text{m}}$

Only two participants in the FAIL group and one in the PASS group fell below the cutoff for MLU$_{\text{m}}$. Overall agreement between MLU$_{\text{m}}$ and the Fluharty-2 was 52%, with 97% agreement for passing but only 7% agreement for failing. This resulted in a chi-square statistic that was not significant ($\chi^2 = 0.350$, $p = .554$). The degree of agreement was even lower for this measure ($\kappa = .032$).

FVMC

Twenty-one participants in the FAIL group and 15 in the PASS group fell below the cutoff for FVMC. Overall agreement was 60%, with 52% agreement for passing and 68% agreement for failing. The chi-square statistic for this comparison was not significant ($\chi^2 = 2.385$, $p = .123$), and agreement was low ($\kappa = .194$).

IPSyn

Twenty-one participants in the FAIL group and six in the PASS group fell below the cutoff for IPSyn. Overall agreement between the IPSyn and the Fluharty-2 was 74%, with 81% agreement for passing and 68% agreement for failing. This resulted in a statistically significant chi-square ($\chi^2 = 14.762$, $p < .001$), suggesting moderate agreement between the measures ($\kappa = .484$)

Discussion

In this study, we compared performance on a screening test, namely, the Fluharty-2, to performance on four diagnostic assessment measures; a standardized test, namely, SPELT-P2; and three LSA measures, namely, MLU$_{\text{m}}$, FVMC, and IPSyn. Children who failed the Fluharty-2 showed significantly lower performance on each of the diagnostic measures than children who passed the screening test. However, the effect size for MLU$_{\text{m}}$ was small. This means that there was actually little difference in MLU$_{\text{m}}$ between children who passed and children who failed the screening. Performance on the Fluharty-2 was significantly correlated with performance on each of the diagnostic measures at least on a medium level.

Given the medium to large effect sizes for the Fluharty-2 score differences and the medium to large correlation with the Fluharty-2, we might have expected good agreement for passing and failing between the Fluharty-2 and SPELT-P2, IPSyn, and FVMC. However, only the comparison of pass/fail decisions between the Fluharty-2 and IPSyn was significant with a moderate level of agreement. For the other diagnostic measures, agreement for pass/fail decisions was low. This finding highlights the importance of examining agreement between measures and of considering cutoff levels for interpreting performance. While the group differences and correlations showed trends in performance (e.g., that lower performance on the screening was associated with lower performance on the diagnostic measures), those statistics did not align with clinical decisions for the diagnostic measures relative to cutoff levels of performance.

Since follow-up assessments would typically involve more than one assessment, we also wanted to compare performance on the Fluharty-2 to a combined criteria that considered performance on both a standardized test and LSA measures. For this comparison, we defined "failing" as scoring below the −1.25 *SD* cutoff on the standardized

Table 4. Correlations between the General Language Quotient (GLQ) screening composite and standard or z-scores on diagnostic measures.

Score	1	2	3	4	5	*M*	*SD*
1. Fluharty-2 GLQ Composite	—					94.4	11.04
2. SPELT-P2 standard score	.722[a]	—				103.08	12.17
3. MLU$_{\text{m}}$ *z* score	.351[a]	.485[a]	—			0.12	0.97
4. IPSyn *z* score	.476[a]	.577[a]	.451[a]	—		−1.24	1.33
5. FVMC *z* score	.413[a]	.509[a]	.352[a]	.335[a]	—	−3.00	5.03

Note. Fluharty-2 = Fluharty Preschool Speech and Language Screening Test–Second Edition; SPELT-P2 = Structured Photographic Expressive Language Test–Preschool, Second Edition; MLU$_{\text{m}}$ = mean length of utterance in morphemes; IPSyn = Index of Productive Syntax; FVMC = finite verb morphology composite.

[a] Correlations are significant at the .01 level (two-tailed).

Table 5. Chi-square cross-tabulations comparing pass/fail decisions.

	SPELT-2		MLU$_m$		IPSyn		FVMC	
Fluharty-2	**Above cutoff**	**Below cutoff**	**Above cutoff**	**Below cutoff**	**Above cutoff**	**Below cutoff**	**Above cutoff**	**Below cutoff**
PASS	31	0	30	1	25	6	16	15
FAIL	29	2	29	2	10	21	10	21
Total	60	2	59	3	35	27	26	36
χ^2	2.067, p = .151		0.350, p = .554		14.762, p < .001		2.385, p = .123	
κ	.065		.032		.484		.194	

Note. Fluharty-2 GLQ = Fluharty Preschool Speech and Language Screening Test–Second Edition, General Language Quotient; SPELT-P2 = Structured Photographic Expressive Language Test–Preschool, Second Edition; MLU$_m$ = mean length of utterance in morphemes; IPSyn = Index of Productive Syntax; FVMC = finite verb morphology composite.

test and on at least one of the LSA measures and "passing" as scoring at or above the cutoff on the test and all three LSA measures. However, given that only two participants in the FAIL group failed the SPELT-P2, this comparison resulted in a very low agreement rate (≤ 7%) for failing regardless of which LSA measure was used. Since all participants in the PASS group passed the SPELT-P2, agreement for passing was a function of the specific LSA measure, ranging from 52% for FVMC, 81% for IPSyn, to 97% for MLU$_m$.

We had compared performance on the Fluharty-2 to what we believed would be a valid assessment protocol used by SLPs for diagnosing LI in young children. However, we could not rule out that any lack of agreement between the Fluharty-2 and that assessment protocol reflected the criterion measures themselves rather than the Fluharty-2. This was particularly of concern for the two criterion measures showing high pass rates, including for children in the FAIL group.

A prior study by Greenslade et al. (2009) had reported good sensitivity and specificity for the SPELT-P2. However, their study included only 4- and 5-year-olds, and their results might not be applicable to the 3-year-old children in the current study. An examination of the content of the SPELT-P2 by Oetting and Hadley (2009) suggested why this might be the case. These authors noted (a) that items known to be difficult for children with LI (such as verb tense markers) are mixed in with other structures that are less diagnostically sensitive (such as progressive *–ing,* plural *–s*) and (b) that the less diagnostically sensitive items account for the majority of earlier items on the test. They thus concluded that performance on the SPELT-P2 might overestimate language performance in younger children. This might explain why so few of the 3-year-old children in the current study failed the SPELT-P2 and the low agreement rate for failing with the Fluharty-2.

In addition, Greenslade et al. (2009) used an empirically determined cutoff for SPELT-P2 of −0.86 *SD*, whereas the current study used a prescriptive cutoff of −1.25 *SD*. We did not use that cutoff for two reasons: (a) Applying an −0.86 *SD* cutoff did not appreciably change the results as only three children in the FAIL group scored below that level on the SPELT-P2, and (b) a −0.86 cutoff would not conform to the eligibility guidelines that are typically used in schools/clinical practice. That is, children who score only 0.86 *SD*s below the mean would not be eligible for speech-language services in most school settings.

We had anticipated that MLU$_m$ might be insufficiently sensitive for diagnosing LI. Eisenberg et al. (2001) calculated predictive values for MLU$_m$ based on an assumed prevalence rate for LI of 8% using data from Klee et al. (1989b). The positive predictive value was 96% (meaning that MLU$_m$ would correctly identify 96% of children with LI), while the negative predictive value was only 54% (meaning that MLU$_m$ would correctly identify only 54% of children with TL) for MLU$_m$. This suggests that, although a low MLU$_m$ (i.e., below the cutoff) can be used as evidence of LI, a high MLU$_m$ (i.e., at or above the cutoff) cannot be used to conclude that language is normal. This means that a large number of children with LI are likely to have an MLU$_m$ above the cutoff, which could have accounted for the low failure rate for MLU$_m$ and the low agreement for failing.

Limitations

It is important to note the lack of independent confirmation about the children's language status. This means that we did not have results of a prior diagnostic assessment by an SLP for the children in the current study. That is why we reported agreement rates rather than diagnostic accuracy, per se, for the Fluharty-2. However, the diagnostic assessment used for the current study included a standardized test, the SPELT-P2, and other norm-referenced measures based on LSA. This is consistent with state regulations that mandate the use of norm-referenced assessments to qualify children for special education services and reflects the measures likely to be used by clinicians to determine language status and eligibility for special education services. To the extent that these measures accurately classify children as having LI or TL, an assessment using these measures could be considered an indicator of language status.

An additional limitation is the lack of evidence about diagnostic accuracy for 3-year-olds for some of the criterion measures used in the current study, specifically, the SPELT-P2 and MLU$_m$. However, the current study also included two other measures, FVMC and IPSyn, for which there is evidence of sensitivity to detect language deficits. Thus, it seems reasonable to conclude that the low agreement for pass/fail decisions reflects limitations of the Fluharty-2 itself rather than being attributable to these latter diagnostic measures.

Finally, participants were limited to mainstream English-learning children. Thus, we cannot be certain that our conclusion about the Fluharty-2 is applicable to children speaking nonmainstream dialects. In addition, the racial composition of the PASS and FAIL groups was not equally balanced, with a lower percentage of White children in the FAIL group. However, since the groups were balanced for socioeconomic status and did not include nonmainstream speakers, this was unlikely to have affected the results.

Clinical Implications

A good screening should result in at least 90% agreement for passing (i.e., no more than 10% failure on a follow-up assessment) and at least 70% agreement for failing (i.e., no more than 30% passing on a follow-up assessment) relative to the follow-up diagnostic assessment (McCauley, 2001; Plante & Vance, 1995). Even for IPSyn, which was the only diagnostic measure for which pass/fail agreement with the Fluharty-2 was statistically significant, the agreement rates for passing and for failing did not meet an acceptable level. This means both that too many children would be missed for a follow-up evaluation and that too many children would be referred for a follow-up assessment, thus making the Fluharty-2 an inefficient tool for screenings. Given the limited evidence about diagnostic accuracy for some of the follow-up measures, clinicians may want to monitor children who fail a screening with the Fluharty-2 but pass a follow-up evaluation.

Concluding Thoughts

The determination of diagnostic accuracy for a screening instrument reflects the extent of agreement between performance on the screening instrument and the specific diagnostic measures used for the follow-up evaluation. In their review of the diagnostic accuracy of language screening assessments, Law et al. (1998) observed that screenings are more effective for ruling out LI than for identifying children at risk for LI. The better agreement rates for passing between the Fluharty-2 and the criterion measures were consistent with this finding. The difference in outcome among the diagnostic measures and the limited evidence about diagnostic accuracy call into question our ability to trust not only the screening test but also the measures that would be used for follow-up assessments at age 3 years. This highlights the need for more research to determine which diagnostic measures have good diagnostic accuracy as well as studies about how well screening measures agree with those diagnostic measures.

Acknowledgments

Data for this study were from an archival data originally collected with support to the first author from National Institute on Deafness and Other Communication Disorders Award R21DC009218. The authors are grateful to the children who participated as well as to their parents who allowed them to participate and to the research assistants who collected and transcribed the samples. Portions of this study were presented at the 2013 Symposium for Research on Child Language Disorders in Madison, Wisconsin, and at the 2013 convention of the American Speech-Language-Hearing Association in Chicago, Illinois.

References

Altenberg, E. P., Roberts, J. A., & Scarborough, H. S. (2018). Young children's structure production: A revision of the Index of Productive Syntax. *Language, Speech, and Hearing Services in Schools, 49,* 995–1008.

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192.

Black, L. I., Vahratian, A., & Hoffman, H. J. (2015, June). *Communication disorders and use of intervention services among children aged 3–17 years: United States, 2012* (NCHS Data Brief, no. 205). Hyattsville, MD: National Center for Health Statistics.

Blaxley, L., Clinker, M., & Warr-Leeper, G. (1983). Two language screening tests compared with Developmental Sentence Scoring. *Language, Speech, and Hearing Services in Schools, 14,* 38–46.

Blosser, J. (2012). *School programs in speech-language pathology*. San Diego, CA: Plural.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language (CASL)*. Circle Pines, MN: AGS.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Conant, S. (1987). The relationship between age and MLU in young children: A second look at Klee and Fitzgerald's data. *Journal of Child Language, 14,* 169–173.

Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K. (2005). *Structured Photographic Expressive Language Test–Preschool, Second Edition (SPELT-P2)*. DeKalb, IL: Janelle Publications.

deVilliers, J. G., & deVilliers, P. A. (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research, 2,* 331–341.

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). Use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10,* 323–342.

Eisenberg, S. L., & Guo, L. (2016). Using language sample analysis in clinical practice: Measures of grammatical accuracy for identifying language impairment in preschool and school-age children. *Seminars in Speech and Language, 37,* 106–116.

Fluharty, N. (1978). *Fluharty Preschool Speech and Language Screening Test*. New York, NY: Teaching Resources.

Fluharty, N. (2001). *Fluharty Preschool Speech and Language Screening Test–Second Edition (Fluharty-2)*. Austin, TX: Pro-Ed.

Goldstein, P. A. (1994). *A comparison of language screening procedures in the identification of children with language* *delays in prekindergarten classes*. Retrieved from ProQuest Dissertations and Theses. (Order No. 9501606).

Greenslade, K., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construction validity of the Structured Photographic Expressive Language Test–Preschool: Second Edition. *Language, Speech, and Hearing Services in Schools, 40,* 150–160.

Guo, L. Y., & Eisenberg, S. (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology, 23*(2), 203–212.

Guo, L. Y., & Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in three-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools, 46,* 141–153.

Hurford, D. P. (2003). Review of the Fluharty Preschool Speech and Language Screening Test–Second Edition. In B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook* (pp. 395–397). Lincoln, NE: Buros Institute.

Hutchinson, T. L. (2010). Review of the Structured Photographic Expressive Language Test–Preschool 2. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 595–598). Lincoln, NE: Buros Institute.

Kamphaus, R. W., & Reynolds, C. R. (2003). *Reynold's Intellectual Screening Test (RIST)*. Lutz, FL: Psychological Assessment Resources.

Kelly, D. J., & Rice, M. L. (1986). A strategy for language assessment of young children: A combination of two approaches. *Language, Speech, and Hearing Services in Schools, 17,* 83–94.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989a). A comparison of the age–MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders, 54,* 226–233.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989b). *Predictive value of MLU in normal and language impaired preschool children*. Unpublished manuscript.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (1998). Screening for primary speech and language delay: A systematic review of the literature. *Health Assessment Technology, 2,* 1–184.

Lee, L. L. (1974). *Developmental sentence analysis*. Evanston, IL: Northwestern University Press.

Loeb, D. F., Kinsler, K., & Bookbinder, L. (2000). *Current language sampling practices in preschools*. Poster presented at the convention of the American Speech and Hearing Association, Washington, DC.

McCauley, R. J. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Erlbaum.

McCauley, R. J. (2003). Review of the Fluharty Preschool Speech and Language Screening Test–Second Edition. In *The fifteenth mental measurements yearbook* (pp. 395–397). Lincoln, NE: Buros Institute.

Merrill, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28,* 50–58.

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24,* 154–161.

Miller, J. F., & Iglesias, A. (2003–2007). Systematic Analysis of Language Transcripts (SALT, Version 9) [Computer software]. Madison: University of Wisconsin–Madison, Waisman Center, Language Analysis Laboratory.

Newcomer, P., & Hammill, D. (1997). *Test of Language Development–Primary: Third Edition (TOLD-P:3)*. Austin, TX: Pro-Ed.

Oetting, J. B., & Hadley, P. A. (2009). Morphosyntax in child language disorders. In R. Schwartz (Ed.), *Handbook of child language disorders* (pp. 341–364). New York, NY: Psychology Press.

Paul, R., Norbury, C. F., & Gosse, C. (2018). *Language disorders from infancy through adolescence* (5th ed.). St. Louis, MI: Elsevier.

Pawlowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research, 57,* 2261–2273.

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15,* 247–254.

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

Rescorla, L., Dahlsgaard, K., & Roberts, J. (2000). Late-talking toddlers: MLU and IPSyn outcomes at 3:0 and 4:0. *Journal of Child Language, 27,* 643–664.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairment. *Journal of Speech, Language, and Hearing Research, 53,* 333–349.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11,* 1–22.

Scarborough, H. S., Wyckoff, J., & Davidson, R. (1986). A reconsideration of the relation between age and mean utterance length. *Journal of Speech and Hearing Research, 29,* 394–399.

Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language & Communication Disorders, 45,* 691–702.

Semel, E., Wiig, E., & Secord, W. (2004). *Clinical Evaluation of Language Fundamentals–Preschool: Second Edition (CELF-P: 2)*. Austin, TX: Pearson.

Shipley, K., Stone, T., & Sue, M. (1983). *Test for Examining Expressive Morphology (TEEM)*. Austin, TX: Pro-Ed.

Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Disorders, 27,* 456–465.

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

Sturner, R. A., Heller, J. H., Funk, S. G., & Layton, T. L. (1993). The Fluharty preschool speech and language screening test: A population-based validation study using sample-independent decision rules. *Journal of Speech and Hearing Sciences, 36,* 738–745.

Towne, R. L. (2010). Review of the Structured Photographic Expressive Language Test–Preschool 2. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 598–600). Lincoln, NE: Buros Institute.