LSHSS

Research Article

Development and Validation of the Spanish–English Language Proficiency Scale (SELPS)

Ekaterina Smyk,[a] M. Adelaida Restrepo,[a] Joanna S. Gorin,[a] and Shelley Gray[a]

Purpose: This study examined the development and validation of a criterion-referenced Spanish–English Language Proficiency Scale (SELPS) that was designed to assess the oral language skills of sequential bilingual children ages 4–8. This article reports results for the English proficiency portion of the scale.
Method: The SELPS assesses syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity based on 2 story retell tasks. In Study 1, 40 children were given 2 story retell tasks to evaluate the reliability of parallel forms. In Study 2, 76 children participated in the validation of the scale against language sample measures and teacher ratings of language proficiency.
Results: Study 1 indicated no significant differences between the SELPS scores on the 2 stories. Study 2 indicated that the SELPS scores correlated significantly with their counterpart language sample measures. Correlations between the SELPS and teacher ratings were moderate.
Conclusions: The 2 story retells elicited comparable SELPS scores, providing a valuable tool for test–retest conditions in the assessment of language proficiency. Correlations between the SELPS scores and external variables indicated that these measures assessed the same language skills. Results provided empirical evidence regarding the validity of inferences about language proficiency based on the SELPS score.

Key Words: English language proficiency, assessment, rating scale, sequential bilinguals

Language proficiency (LP) is the ability to speak and comprehend a language on a continuum from nonproficient to native-like proficiency. LP assessment remains a controversial area in language acquisition and testing due to ongoing debates about the theoretical framework and definition of the construct (e.g., Martin-Beltrán, 2010; McNamara, 1996; Solórzano, 2008). Over the years, numerous competing frameworks of LP have been proposed (e.g., Canale & Swain, 1980; Cummins, 1976; McNamara, 1996; Skehan, 1998), yet there is still little agreement on how to conceptualize LP and its underlying dimensions. For example, communicative competence models (Bachman, 1990; Bachman & Palmer, 1982; Canale & Swain, 1980) characterize the construct of LP as communicative competence, whereas processing-based models focus on the language tasks a learner can perform and available cognitive resources and processing constraints (e.g., Pienemann, 1998; Robinson, Tin, & Urwin, 1995; Skehan, 1998). In addition, according to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & the National Council on Measurement in Education [NCME], 1999), few reliable measures exist that provide valid score inferences for children's LP (Bedore & Peña, 2008; MacSwan, Rolstad, & Glass, 2002; Pray, 2005; Solórzano, 2008).

For sequential bilinguals, LP assessment is particularly challenging with respect to a second language (L2). These children learn their native language (L1) from birth and are exposed to an L2 after the age of 3 years (Kohnert, 2004), while still developing their L1. Sequential bilinguals represent a fast-growing population in the public schools, accounting for ~10.4% of the total school student enrollment in the United States in 2009–2010 (Aud et al., 2012). When working with sequential bilingual children, it is crucial to distinguish between language difficulties that are caused by a low level of linguistic attainment due to insufficient exposure to a particular language and language difficulties that are due to language impairment.

When examining children's language ability, the language in which a child is tested is critical. The language used in testing can have a significant effect on the results and interpretation of test scores (Bedore & Peña, 2008; Bedore et al., 2012; Goldstein, 2006). When sequential bilinguals with typical language development are assessed in a language in which they have not attained native-like proficiency, their performance on traditional language measures, such as language sample analysis (LSA) or standardized tests, may resemble the

[a]Arizona State University, Tempe

Correspondence to Ekaterina Smyk: ekaterina.smyk@gmail.com

Editor: Marilyn Nippold
Associate Editor: Gary Troia

Received August 9, 2012
Revision received December 4, 2012
Accepted February 8, 2013
DOI: 10.1044/0161-1461(2013/12-0074)

performance of monolingual children with language impairment (e.g., Håkansson & Nettlebladt, 1996; Paradis, 2010; Paradis & Crago, 2000). Before administering measures of language impairment, the child's LP must be assessed to determine the appropriate language for testing.

Assessment of LP in Sequential Bilingual Children

Prior to the No Child Left Behind Act of 2001 (2002), state education agencies (SEAs) employed a variety of methods to identify LP in sequential bilinguals, including home language-use surveys, criterion-referenced tests, achievement tests, and LP tests (Kindler, 2002). According to the *Survey of the States' Limited English Proficient Students and Available Educational Programs and Services 2000–2001 Summary Report* (Kindler, 2002), out of 56 SEAs that participated in the survey, 51 reported using LP tests for student identification and placement purposes. Despite the popularity of these LP tests among educators and practitioners, the psychometric properties and conceptual framework of such measures have been frequently criticized (e.g., Larkin, Restrepo, & Morgan, 2007; MacSwan & Rolstad, 2006; Pray, 2005). For example, Pray (2005) administered the oral English subtests of the Language Assessment Scales (De Avila & Duncan, 2005), the IDEA Proficiency Test (Ballard, Tighe, & Dalton, 2005), and the Woodcock-Muñoz Language Survey (Woodcock, Muñoz-Sandoval, Ruef, & Alvarado, 2005) to 40 monolingual English-speaking non-Hispanic White and Hispanic students in the fourth and fifth grades from varying socioeconomic backgrounds. Although results indicated no significant differences between the non-Hispanic White and Hispanic students, and found no effect of socioeconomic status (SES) on the students' performance on the LP tests, none of the students was classified as proficient or advanced proficient on the Woodcock-Muñoz Language Survey, and only 85% of the students were identified as proficient in their L1 by the IDEA Language Proficiency Test. The Language Assessment Scales was the only measure that identified 100% of the monolingual English-speaking children as proficient in their L1.

Under the No Child Left Behind Act of 2001 (2002), states are required to develop standards of English LP that are focused on academic language skills and to periodically assess LP attainment based on these standards. However, focusing on academic language is problematic because oral LP and academic achievement are confounded in these assessments (MacSwan & Rolstad, 2006). Despite numerous attempts to define the construct of academic language, there is little systematic research on what language skills the construct includes and whether, in fact, this construct is distinctive from *social* language (Bailey & Huang, 2011). For instance, low performance on a measure confounded by academic skills may not indicate low LP, but instead may indicate limited academic skills due to suboptimal education (e.g., MacSwan, 2000; MacSwan & Rolstad, 2006). On the other hand, bilingual children who are learning English as their L2, but who have prior educational experience in their L1, may perform better on measures of academic language in comparison to bilingual peers who do not have similar experience in their L1.

Teacher ratings. Teacher ratings are another type of assessment that has been used to evaluate English LP in bilingual children (Bedore, Peña, Joyner, & Macken, 2011; Gutiérrez-Clellen & Kreiter, 2003; Restrepo, 1998; Restrepo & Gutiérrez-Clellen, 2001). Gutiérrez-Clellen and Kreiter (2003) argued that teachers are more accurate than parents in their assessment of English LP in bilingual children because parents' English LP levels may interfere with their ratings. In addition, Gutiérrez-Clellen and Kreiter reported that teacher ratings of English LP were moderately and significantly correlated with the proportion of grammatical utterances in language samples ($r = .44$). Similarly, Bedore et al. (2011) reported that teacher ratings on a 5-point Likert-type scale were weakly but significantly correlated with language scores on English semantics and morphosyntax subtests in Spanish–English bilingual children ages 4;0–5;11 (years;months) (Kendall's τ coefficient of .18 and .23, respectively). It is important to note, however, that these correlations may have reached significance because of the large sample size.

LSA. LSA is considered the gold standard in the assessment of oral language skills in speech-language pathology (Heilmann et al., 2008; Heilmann, Miller, Nockerts, & Dunaway, 2010; Miller et al., 2006) and in LP assessment (MacSwan & Rolstad, 2006). LSA requires eliciting a language sample, transcribing the sample, and coding and analyzing language features such as syntactic length and complexity, grammatical accuracy, and lexical diversity. The purpose of this procedure in speech-language pathology is to identify whether or not a child has a speech and language disorder, but the purpose for LP assessment is to characterize the child's L2 productions. LSA offers several advantages over norm-referenced tests because it simultaneously assesses multiple areas of an individual's language and provides a direct representation of an integrated language performance in a meaningful communication context (Heilmann et al., 2010). Further, LSA is a reliable form of assessment. Heilmann et al. (2008) showed the stability of language sample scores over a period of 2 months, with correlation coefficients ranging from .65 for the mean length of utterance (MLU) to .79 for the total number of different words (NDW). Coding reliability was 98%–100%.

Rating scales. One drawback to using LSA as a measure of LP is the time it takes to transcribe and code the language samples. Nevertheless, LSAs are important for validating LP assessments. One possible alternative to LSA is online rating of language, which may be more efficient than transcription if raters can be trained to high accuracy levels and if the validity of score interpretations can be demonstrated. Rating scales are frequently used in LP assessment (Brindley, 1998; North, 2000) to address this problem. Results may be represented by scores plotted on two intersecting axes, with one axis representing a continuum of increasing language ability from nonproficient to native-like proficiency that describes the stages or levels of acquisition and the second axis representing categories or domains that are related to a theoretical framework of LP (Brindley, 1998; North, 2000). Rating scales such as this offer a number of advantages for defining LP globally in terms of real-life performance within a specific context of language use. They establish a meaningful frame of reference and provide a useful method for comparing individuals within a population (North, 2000).

Study Purpose

The purpose of this study was to develop a criterion-referenced Spanish–English Language Proficiency Scale (SELPS) that yields valid and reliable score interpretations based on the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and on L2 acquisition theory. According to the testing standards, there are several important steps in test development, including description of a theoretical model of an ability or trait, generation of an operational definition, administration of measures to collect observations, and evaluation of the evidence in terms of the initial hypothesis (Gorin, 2006). The SELPS aims to measure the level of oral LP in 4- to 8-year-old sequential bilingual children learning English as an L2 in order to identify whether a child has sufficient skills in the L2 to be tested in English. The present study reports the development and preliminary validation of the SELPS for assessment of the English LP component.

To this end, we first reviewed the extant literature on LP, from which we derived our definition of the construct for scale development. Then, we operationalized the construct of LP and developed a rating scale and two parallel language tasks that elicited behavioral evidence of the construct of L2 proficiency. This was accomplished in two studies. Study 1 was dedicated to the evaluation of parallel forms of the language elicitation task using the same scale. Study 1 addressed the following research question: Do parallel forms of the language elicitation task elicit comparable SELPS ratings? Study 2 was dedicated to the evaluation of the construct validity of the score inferences based on the SELPS by examining the relationship of the subscale and scaled scores with external measures. Study 2 addressed the following research questions: (a) Are SELPS subscale and scaled scores related to scores on counterpart language sample measures? and (b) Are SELPS scores related to teacher ratings of L2 proficiency?

Theoretical Framework of LP

Syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity refer to the domains of learners' performance that are used to describe the continuum of LP (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008; Norris & Ortega, 2009; Skehan, 2009). Although there is some variability in how these domains have been operationalized, there is consensus regarding their usefulness in describing L2 acquisition levels (Housen & Kuiken, 2009; Iwashita, 2010; Pallotti, 2009; Skehan, 2009).

Syntactic complexity. Syntactic complexity is the length, elaborateness, and variety of verbal utterances produced by an L2 learner. The MLU based on the number of words per sentence is associated with oral L2 proficiency in young adults. The more proficient the language learner is, the higher the sentence complexity in his or her oral and written language is (Iwashita, 2006; Iwashita et al., 2008). In addition, the MLU in English as an L2 has been found to increase with age in sequential bilingual children ages 4–7 years (Miller et al., 2006; Muñoz, Gillam, Peña, & Gulley-Faehnle, 2003).

Grammatical accuracy. Grammatical accuracy is indexed by the degree of deviation from a target norm, which is measured using metrics such as frequencies, ratios, and formulas focused on the identification of grammatical errors (Housen & Kuiken, 2009; Pallotti, 2009). Muñoz et al. (2003) reported that predominantly English-speaking bilingual children ages 3;10–4;8 produced on average 59% grammatically correct utterances in a storytelling task, whereas predominantly English-speaking bilingual children ages 5;0–5;6 produced on average 80% grammatically correct utterances in the same task. Similarly, Fiestas and Peña (2004) reported that Spanish–English bilingual children ages 4;0–6;11 who were fluent speakers of English had ~80% grammatical utterances in a storytelling elicited by a wordless picture book. Chondrogianni and Marinis (2011) found that after 6 years of English exposure, school-age bilingual children demonstrated increased grammatical accuracy in their production of tense and nontense morphemes, except for a few children, who continued to demonstrate difficulties with English past tense.

Verbal fluency. Verbal fluency is a learner's control over his or her language knowledge manifested by temporal variables such as speech rate, pauses, false starts, reformulations, and repetitions (e.g., Housen & Kuiken, 2009; Riazantseva, 2001). LP has been found to affect verbal fluency in terms of duration, pause frequency (Riazantseva, 2001), and speech rate, as measured by the total number of syllables per minute (Kormos & Denes, 2004). In a study by Fiestas, Bedore, Peña, and Nagy (2005), sequential bilingual children ages 4–7 years demonstrated higher rates of repetition at the sound, word, and phrase level in storytelling in English and Spanish compared to dominant Spanish- or English-speaking children who had been in contact with Spanish or English <20% of the time, respectively, indicating that language proficiency impacts verbal fluency in children.

Lexical diversity. Lexical diversity is the variety and specificity of words that are produced by a speaker. Lexical diversity is measured using a variety of methods, including the NDW and type-token ratios (e.g., Golberg, Paradis, & Crago, 2008; Muñoz et al., 2003). Lexical diversity has been shown to predict L2 proficiency (Iwashita et al., 2008; Zareva, Schwanenflugel, & Nikolova, 2005). Studies suggest that the NDW is a sensitive developmental measure of lexical skills in young Spanish–English bilingual children (Golberg et al., 2008; Muñoz et al., 2003).

Stages of LP Development

Tabors (2008) described four general stages of English language acquisition in sequential bilingual children: (a) home language use, (b) nonverbal period, (c) telegraphic and formulaic use, and (d) productive language use. During the home language stage, children continue to speak their L1 as if people around them can understand and speak the same language (Saville-Troike, 1987; Tabors, 2008). When children realize that use of their home language is not effective, they enter a period of nonverbal communication. During the nonverbal period, children abandon any attempts to communicate in their L1 and may use nonverbal communication instead (Tabors, 2008). The presence of the nonverbal period has been documented across young L2 learners from a variety of linguistic backgrounds (e.g., in American children learning French: Ervin-Tripp, 1974; in Japanese-speaking children learning English: Hakuta, 1978; Itoh & Hatch, 1978).

The telegraphic language stage refers to a similar stage of language development in monolingual children (Brown & Fraser, 1963) in which utterances consist of functional content words with the omission of grammatical inflections and function words (e.g., **Mommy give applejuice;* Tabors, 2008). The use of formulaic language is referred to as the production of unanalyzed phrases, which other speakers have been observed to use (e.g., *I don't know;* Peters, 1983; Tabors, 2008). Finally, the stage of productive language use begins when children learn a number of vocabulary words and phrases that allow them to construct their own utterances and go beyond the production of telegraphic and formulaic sentences. During this stage of acquisition, children may use formulaic phrases as building blocks of their early language productions (e.g., * *I got a big*) and continue making grammatical errors (Tabors, 2008). Although these stages represent increasing LP, children eventually master the system, and thus, we argue that after the productive stage, there is a proficient or advanced stage in which children have command of the grammatical, fluency, and lexical components of language and exhibit minimal errors.

Development of the SELPS

The present study used contemporary measurement theory and an assessment design that incorporated cognitive theory into the test development process (e.g., AERA, APA, & NCME, 1999; Embertson & Gorin, 2001; Gorin, 2006; National Research Council, 2001). Scale development proceeded in three stages: (a) description of a theoretical model and generation of an operational definition, (b) administration of measures to collect observations, and (c) evaluation of the evidence in terms of the initial hypothesis (Gorin, 2006).

Construct definition. We defined LP as a continuum of oral language skills ranging from nonproficiency at one end to native-like proficiency at the other, which is based on a functional linguistic approach to L2 acquisition. LP is viewed as linguistic knowledge and the ability to implement the linguistic knowledge to a specified level of performance in a situation defined by cognitive and linguistic demands (Bialystok, 2001). Unlike most available standardized LP measures, in the present study, the construct of LP was considered independently from literacy and other school-related skills because their inclusion would confound interpretation of LP assessment (MacSwan & Rolstad, 2006).

The SELPS was constructed to be administered in conjunction with a story retell task that elicits a language sample. The story retell task was adopted based on literature suggesting that in sequential bilinguals, such tasks are likely to elicit longer utterances and more complex grammatical structures than spontaneous conversation or story generation (Gazella & Stockman, 2003; Gutiérrez-Clellen, 2002; Restrepo & Gutiérrez-Clellen, 2001). Further, the story retell discourse context is authentic for young children and incorporates language comprehension and production (Skarakis-Doyle & Dempsey, 2008). Language comprehension is considered a prerequisite for the ability to produce a story (Pickert & Chase, 1978; Skarakis-Doyle & Dempsey, 2008). In addition, story retelling elicits the different components of language, which allows for efficient assessment of LP.

SELPS construct map. The scores from the SELPS were used to make criterion-referenced decisions regarding L2 proficiency based on the four domains of oral language production—syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity—which were assessed by a corresponding subscale (see Figure 1). Each SELPS subscale differentiated among up to five levels of oral LP, with each higher level subsuming lower levels. The levels of LP were modified from Tabors' (2008) general stages of English language acquisition and were similar to some classifications of oral LP in L2 learners (Krashen & Terrell, 1983).

The construct map of the SELPS, with behavioral descriptors for each subscale assessing each of the four LP domains, the SELPS levels, and the corresponding Tabors' (2008) stages is presented in Figure 1. Tabors' first stage, home language use, is not assessed by the SELPS because during this stage, children continue speaking their L1. Tabors' stage of productive language use was divided into three SELPS levels: short sentences and phrases with multiple errors, full sentences with a few grammatical errors, and native-like production. Qualitative and quantitative differences in LP among levels were expected across the proposed domains of LP. Nondevelopmental errors in the grammatical accuracy subscale were defined as errors that are not found among L1 speakers, such as "are happy" instead of "they are happy" and "boy is get turtle" instead of "the boy is getting a turtle" or "the boy gets a turtle."

Scoring the SELPS subscales. For each of the proposed domains, the rating criteria were generated to correspond with the levels of the construct map (see Figure 1). The subscale ratings ranged from 1 to 4 for syntactic complexity and lexical diversity and from 1 to 5 for grammatical accuracy and verbal fluency; *1* indicated the lowest level of performance, and *5* indicated the highest level of performance. The subscale score ranges for syntactic complexity and lexical diversity were based on pilot work for the scale development, which suggested that the reduced ranges were optimal. Adding an additional level 5 to the syntactic complexity and lexical diversity subscales created difficulty with reliable separation of levels 4 and 5, suggesting that children with high levels of L2 proficiency demonstrate less distinct syntactic and lexical skills than children with low levels of L2 proficiency.

Scoring for each domain was aligned with expectations for each level of LP. The inferences about overall LP level were made by a single scaled score that was based on the four subscale scores. The highest scores on all subscales resulted in an overall score of 5. If there were at least three subscale scores assigned to one level, that score resulted in

Figure 1. The Spanish–English Language Proficiency Scale construct map. SC = syntactic complexity; GA = grammatical accuracy; VF = verbal fluency; LD = lexical diversity.

Language proficiency stages		Score	Language proficiency domains			
Tabors' stage	SELPS level		SC	GA	VF	LD
Home language use	-	-	-	-	-	-
Nonverbal period	Silent/ observer	1	Uses a few words or jargon; doesn't interact or uses gestures to communicate	Not applicable	Not applicable	< 10 different words
Telegraphic and formulaic use	A few words or formulaic phrases	2	Uses single words or short phrases	Telegraphic ungrammatical speech	Most of the time sounds disfluent. Most of the time uses fillers, frequent, and long pauses within sentences and/or repetitions. May frequently switch to non-target language	Basic level of vocabulary (frequent or common words and phrases)
Productive language use	Short sentences and phrases with multiple errors	3	Uses short or incomplete sentences	Multiple nondevelopmental errors	Half of the time sounds disfluent. Half of the time uses fillers, frequent, and long pauses within sentences and/or repetitions	Few different nouns, verbs, and adjectives
	Full sentences with a few grammatical errors	4	Uses variety of age-appropriate sentences	A few nondevelopmental errors	Sometimes sounds disfluent. Sometimes uses fillers, frequent, and long pauses within sentences and/or repetitions	Variety of different verbs, nouns, and adjectives
	Native-like production	5		Only age-appropriate errors or no errors	Seldom sounds disfluent. Seldom uses fillers, frequent, and long pauses within sentences and/or repetitions	

assignment of an overall LP score at that same level (e.g., three subscale scores of 3 resulted in an overall LP score of 3). If two subscales scores were assigned to one level and the other two to the adjacent level, those were averaged to provide an overall score (e.g., two scores of 3 and two scores of 4 resulted in an overall LP score of 3.5). In a few cases in which there were fewer than two subscale scores assigned to one level, or the two scores were assigned to a non-adjacent level, then the average score was interpreted as follows: (a) scores of 2.00–2.24 were assigned a level of 2, (b) scores of 2.25–2.99 were assigned a level of 2.5, (c) scores of 3.00–3.24 were assigned a level of 3, (d) scores of 3.25–3.99 were assigned a level of 3.5, (e) scores of 4.00–4.24 were assigned a level of 4, and (f) scores of 4.25–4.44 were assigned a level of 4.5.

Study 1: Subscale and Scaled Score Reliability

The purpose of this study was to examine the reliability of the proposed scale based on the theoretical SELPS structure and subscale construct map. This issue was examined via parallel forms reliability.

Method

Participants

Forty sequential bilingual children learning English as an L2 and attending English-only schools in Arizona were recruited as part of a larger study on the development of a language screener in predominantly Spanish-speaking children. Nineteen girls and 21 boys ages 54–101 months (M = 74.38; SD = 12.03) participated in the study. Different age groups were included to ensure that the participants had different levels of L2 proficiency. All participants were enrolled in English language development classrooms in the Phoenix metropolitan area and were from low-SES homes based on their participation in a free and reduced-price lunch program at school. Participants were selected based on the following criteria: (a) spoke Spanish at home at least 50% of the time based on parent report, (b) were exposed to English after 3 years of age based on parent report, (c) scored ≥75 on the nonverbal scale of the Kaufman Assessment Battery for Children, Second Edition (K-ABC–II; Kaufman & Kaufman, 2004), and (d) passed a bilateral hearing screening at 20 dB HL in the frequency region from 1000 through 4000 Hz (American National Standards Institute, 1969).

Measures

Qualification measures. Parent report and children's enrollment in English language development classrooms were used to qualify children for participation in the study. Parent report included questions pertaining to the child's language background, degree of exposure to two languages, and exposure patterns to English. Parent report has been used to identify the age of L2 exposure and the amounts of L2 input outside school in sequential bilingual children (Gutiérrez-Clellen & Kreiter, 2003; Restrepo, 1998).

The K-ABC–II is a standardized measure of cognitive ability for children ages 3 to 18 years. It was used to exclude children with a nonverbal inventory (NVI) standard score <75. According to the test manual, the NVI can be used with bilingual children who are not fluent in English. The test includes instructions and a scoring guide for administering the scale in Spanish and nonverbally using pantomime and motoric responses. The norming sample included >500 Hispanic children, minimizing the effects of ethnic background on the score interpretations. In addition, the manual reported that ethnicity accounted for <5% of the variance of the NVI after controlling for SES, mother's level of education, and gender. The composite score test–retest reliabilities ranged from .87 to .95.

Experimental measures. Two parallel forms of the SELPS were developed by manipulating the story retell component of the assessment. Each form used a different story, but the same rating scale was applied to both language samples. The parallel SELPS story retell tasks used the modified wordless storybooks, *Frog on His Own* (Mayer, 1973) and *A Boy, a Dog, a Frog, and a Friend* (Mayer & Mayer, 1971). Scripts for each story were adapted from the publicly available scripts from the Systematic Analysis of Language Transcripts (SALT) software website (http://www.saltsoftware.com/resources-elicitationaids/frogStories/index.cfm). To ensure equivalence in story length and complexity, the number of T-units (Hunt, 1970), mean length of T-units, and number of pictures presented to participants were controlled in each story and across stories. A T-unit was defined as an independent clause with its dependent (subordinate) clauses. There were 43 T-units and 25 pictures in each story that was presented to the children. The mean length of T-units for *Frog on His Own* was 12.00, and the mean length of T-units for *A Boy, a Dog, a Frog, and a Friend* was 12.65.

We developed standardized procedures for task administration and prompting to encourage oral productions. Raters were allowed to ask general questions if a child struggled with the retell (e.g., *What happened here? Can you tell me more about it?*) and to briefly affirm and encourage the children, especially in the beginning of the retell. If children started to switch to their L1 or spoke about unrelated topics, the raters redirected the children multiple times if necessary.

Procedure

Rater training. Raters were bilingual English–Spanish-speaking graduate students in the department of speech and hearing science. Because use of the SELPS requires a rater to understand the linguistic bases of oral language (e.g., to differentiate between telegraphic speech and multiple grammatical errors), each rater was required to attend a series of training sessions. During the training sessions, a number of the prescored and recorded language samples of children at different LP levels were scored by all raters together and were discussed until score agreement was reached. Raters had to reach 100% accuracy in rating all of the samples before they could score independently.

Story retell administration. All children completed the two story retell tasks on the same day, but the sequence of administration was counterbalanced across children to account for story effects. The examiner read the story script to the child while the child looked at the pictures in the book. Then the child was asked to look at the pictures and to tell the story back to the examiner. To minimize memory load, the book was made available to the children. The children's story retells were audio-recorded using a digital Ipod nano (Apple Inc.) or Olympus WS-600 digital voice recorder. The audio recordings of the story retells were rated, using the SELPS, by two trained raters. Each story was rated on a different day without knowledge of other retell scores.

Analysis

Interrater reliability. Before evaluating the parallel forms reliability, we examined the interrater agreement by having a third rater independently score 15 randomly selected story retells (18%). One hundred percent of the SELPS scores were within a .5 score difference between the two raters. A linear weighted kappa was calculated to evaluate the agreement based on the distance between the two SELPS scores. The greater the distance between the two scores, the less weight it was assigned. We used Cohen's (1968) formula for calculating weights:

$$\text{weight} = 1 - \frac{|\text{distance}|}{\text{maximum possible distance}}$$

Kappa (κ) values range from –1 to 1, with negative values indicating poorer than chance agreement, 0 indicating exactly chance agreement, and positive values indicating better than chance agreement (Fleiss & Cohen, 1973). According to Landis and Koch (1977), κ values <0 represent poor agreement, values of .01–.20 represent slight agreement, values of .21–.40 represent fair agreement, values of .41–.60 represent moderate agreement, values of .61–.80 represent substantial agreement, and values of .81–1.00 represent almost perfect agreement. The weighted κ value yielded a value of .81, which indicates almost perfect agreement between the two SELPS scores.

The agreement between the subscale scores within a range of +/– one score point was 100% and yielded a κ value of 1 for the syntactic complexity scores (perfect agreement), a value of .47 for the grammatical accuracy scores (moderate agreement), a value of .68 for the verbal fluency scores (substantial agreement), and a value of .86 for the lexical diversity scores (almost perfect agreement).

Parallel forms reliability. To estimate parallel forms reliability for the two story retell tasks, we conducted three analyses: (a) a Wilcoxon test evaluating differences between the medians of scores on the two story retell tasks; (b) Spearman's rho correlations investigating whether there was a linear relationship between the corresponding subscale scores on the two story retell tasks and a linear relationship between the overall scaled scores; and (c) the percentage agreement between the two SELPS scores and the linear weighted κ evaluating the proportion of weighted agreement corrected for chance.

The Wilcoxon test represents an alternative to paired-samples *t* tests and is more appropriate for ordinal-level data.

It involves ranking the difference between the scaled scores on the two story retells for each child. The Spearman's rho correlations were selected as more appropriate for the ordinal-level data than a Pearson product–moment correlation coefficient.

Results

Descriptive statistics (median, interquartile range, mean, standard deviation, and skew) for the subscale and scaled scores on each story retell task are presented in Table 1. Syntactic complexity and lexical diversity subscale score distributions for the two story retells were negatively skewed as a result of the large number of children who received the highest possible scores on the two subscales: 33 children (82.5%) and 34 children (85.0%) received a score of 4 on the syntactic complexity subscale for *Frog on His Own* and *A Boy, a Dog, a Frog, and a Friend,* respectively, and 30 children (75.0%) and 27 children (67.5%) received a score of 4 on the lexical diversity subscale for *Frog on His Own* and *A Boy, a Dog, a Frog, and a Friend,* respectively. The scaled score distributions on both story retells were negatively skewed as a result of the large number of children who received high overall scaled scores: 23 children (57.5%) and 19 children (47.5%) received an overall scaled score of 4.5. or 5.0 for *Frog on His Own* and *A Boy, a Dog, a Frog, and a Friend,* respectively. The scaled scores ranged from 2 to 5 on both story retell tasks, indicating that the participants' LP was past the nonverbal period of L2 acquisition. The scale describes a continuum of L2 proficiency from a nonverbal period to native-like proficiency; thus, the skewed score distributions were due to characteristics of the study participants.

For estimates of parallel forms reliability, results of the Wilcoxon test indicated a nonsignificant difference between all of the subscale scores for the two story retell tasks ($z = -1.00$, $p = .32$, $r = -.11$ for syntactic complexity; $z = -.27$, $p = .78$, $r = -.03$ for grammatical accuracy; $z = -1.70$, $p = .09$, $r = -.19$ for verbal fluency; and $z = .00$, $p = 1.0$, $r = 0$ for lexical diversity). The results also indicated a nonsignificant difference between the overall scaled scores on the two story retell tasks, $z = -1.37$, $p = .17$, $r = -.15$. In addition, the effect size estimates demonstrated that the story accounted for minimal variance in the subscale and scaled scores. The mean of the ranks in favor of the scaled scores on the *Frog on His Own* story was 11.86; the mean of the ranks in favor of the scaled scores on the *A Boy, a Dog, a Frog, and a Friend* story was 10.88.

Next, we examined the Spearman's rho correlations between the four subscale scores on the two retell tasks. Using the Sidak (1967) adjustment of the Bonferroni procedure to control for Type I error, a p value <.001 was required for significance. The following formula was used to calculate a p value: $\alpha_{\text{S-B}} = 1 - (1 - \alpha_{\text{FWE}})^{1/c}$, where $\alpha_{\text{S-B}}$ is the Sidak-Bonferroni alpha level used to determine significance, α_{FWE} is the desired familywise error of .05, and c is the number of correlations (Sidak, 1967). This approach has less impact on statistical power and is considered less conservative in comparison with the Bonferroni procedure (Keppel & Wickens, 2004).

The results indicated statistically significant correlations for the syntactic complexity scores, $\rho = .76$, $p < .001$; the grammatical accuracy scores, $\rho = .63$, $p < .001$; the verbal fluency scores, $\rho = .59$, $p < .001$; and the lexical diversity scores, $\rho = .82$, $p < .001$. The correlation between the overall scaled scores on the two stories was statistically significant and strong in magnitude, $\rho = .78$, $p < .001$. Finally, the percentage agreement within half of a point between the two overall scaled scores was 87.5%. The weighted κ value was .62, which indicates substantial agreement between the scores assigned to the two story retell tasks.

Discussion

The SELPS was constructed as a criterion-referenced rating scale to assess the level of English LP in sequential bilingual children learning English as an L2 in order to identify their English developmental stage. The theoretical framework of the SELPS considers syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity as the four areas of language performance describing the continuum of L2 proficiency (Iwashita, 2010; Iwashita et al.,

Table 1. Descriptive statistics for the subscale and scaled scores on the two story retell tasks: *Frog on His Own* (Mayer, 1973) and *A Boy, a Dog, a Frog, and a Friend* (Mayer & Mayer, 1971).

	Frog on His Own						***A Boy, a Dog, a Frog, and a Friend***					
					Skew						**Skew**	
Scale domain	***M***	***SD***	**Median**	**Interquartile range**	**Statistic**	**SE**	***M***	***SD***	**Median**	**Interquartile range**	**Statistic**	**SE**
SC	3.79	0.52	4.00	0.00	–2.59	.38	3.84	0.49	4.00	0.00	–3.20	.38
GA	3.82	0.76	4.00	1.00	–0.45	.38	3.82	0.68	4.00	1.00	–0.28	.38
VF	3.72	0.79	4.00	1.00	–0.11	.38	3.85	0.90	4.00	2.00	–0.13	.38
LD	3.64	0.71	4.00	0.00	–1.71	.38	3.64	0.58	4.00	1.00	–1.42	.38
OSS	4.01	0.87	4.50	1.00	–1.01	.37	4.09	0.85	4.00	1.50	–0.72	.37

Note. SE – standard error; SC = syntactic complexity; GA = grammatical accuracy; VF = verbal fluency; LD = lexical diversity; OSS = overall scaled score.

2008; Kormos & Denes, 2004; Norris & Ortega, 2009; Skehan, 2009). The five levels of oral LP in the SELPS were based on the modified Tabors' (2008) stages of L2 acquisition. To elicit the behavioral evidence of L2 proficiency, the two story retell tasks were based on wordless storybooks. The story retell task incorporated both language comprehension and language production (Skarakis-Doyle & Dempsey, 2008) and was more likely to elicit longer utterances and complex grammatical structures than spontaneous conversation or story generation in sequential bilinguals (Gutiérrez-Clellen, 2002; Restrepo & Gutiérrez-Clellen, 2001).

To investigate whether the two story retell tasks elicited comparable language performance in bilingual children based on the SELPS, we conducted three analyses. The first analysis indicated that there were no differences between the corresponding subscale and scaled scores on the two story retells. These results suggest that on average, children's performance on the two tasks did not differ. The second analysis revealed significant medium-to-large correlations between the corresponding subscale scores on each of the stories and large correlations between the corresponding scaled scores. Thus, children who received high scores on one story retell task were likely to receive high scores on the second story retell task, regardless of which story was administered first, and children who received low scores on one story retell task were likely to receive low scores on the second story retell task. The third analysis revealed substantial agreement between the two scaled scores.

Overall, results indicated that the SELPS provides comparable ratings of LP based on the two story retell tasks. Using the two stories provided a valuable tool for test–retest conditions in the assessment of L2 proficiency and allowed memory effects to be eliminated. However, it is important to note that all of the participants were expected to have some exposure to the task of storytelling; therefore, the elicited language productions were not influenced by novelty. The results may be different for children who did not have any previous exposure to the story retell task.

Study 2: Validity Evidence

Study 2 aimed to evaluate the appropriateness of the SELPS as a measure of English LP. Specifically, we examined the relationships of the SELPS scores with theoretically related external variables such as language sample measures and teacher ratings of English LP.

Method

Participants

Seventy-six sequential bilingual children learning English as an L2 attending English-only schools in Arizona participated in this study. Of the 76 participants, 21 (9 boys and 12 girls) also participated in Study 1. Because these children had completed two story retells, the SELPS scores from one administration were randomly selected for analysis in Study 2. There were 43 girls and 33 boys ages 59–108 months (M = 76.71; SD = 11.76). All participants were enrolled in English language development classrooms in schools in the Phoenix metropolitan area and were from low-SES homes based on participation in the free and reduced-price lunch program. Participants were selected using the same selection criteria described in Study 1. In addition, to ensure that participants had typical language development, all children were required to pass the Clinical Evaluation of Language Fundamentals—4, Spanish Edition (CELF–4; Wiig, Secord, & Semel, 2006) with a standard score ≥85. Participants with typical language development were randomly selected from a larger study of Spanish–English-speaking children. Children with language impairment were not selected for the validation because their L2 characteristics could be influenced by language impairment rather than level of L2 proficiency.

The CELF–4 is a standardized measure that is used to identify language impairment in primarily Spanish-speaking children ages 5 to 21. The standardization sample included bilingual Spanish–English students, and the test manual reported that Spanish was the L1 of all participants. The core subtests administered to our study participants included Concepts and Directions, Word Structures, Sentence Repetition, and Formulating Sentences. The test–retest reliability across subtests ranged from .80 to .95. The test mean is 100 and the standard deviation is 15.

External Validation Measures

Teacher ratings. A questionnaire was used to obtain teacher ratings of children's English LP based on a 5-point Likert-type scale, with a score of *1* indicating that a child cannot speak English at all and a score of *5* indicating that a child has native-like English LP. The rating scale was similar to the teacher scale of LP examined in Restrepo (1998) and in Gutiérrez-Clellen and Kreiter (2003).

Language sample measures. Story retelling samples used for obtaining the SELPS scores were orthographically transcribed using SALT (Miller & Iglesias, 2008). Running speech was segmented into T-units and then coded for subordinate clauses and ungrammatical sentences. On average, children produced 28 T-units in their language samples. SALT was used to generate values for the following variables from the transcribed language samples: (a) mean length of utterance in words (MLUw); (b) NDW; (c) ungrammaticality index (UG; a total number of ungrammatical sentences divided by a total number of T-units); (d) subordination index (SI; a total number of subordinate clauses divided by a total number of T-units); and (e) percentage of maze words (PMW), such as false starts, repetitions, and reformulations, to total number of words produced in the transcript. MLUw and SI were selected as measures of syntactic length and complexity (Hunt, 1970; Iwashita et al., 2008); NDW was selected as a measure of lexical diversity. Although NDW is affected by utterance length (Klee, 1992), it is considered a robust indicator of lexical skills when the language elicitation procedure is standardized. UG was used as a global measure of grammatical accuracy (Iwashita, 2010; Iwashita et al., 2008), and PMW was used as a measure of verbal fluency (Fiestas et al., 2005; Heilmann et al., 2010).

Procedure

The administration of the story retell task was based on the procedures discussed in Study 1. Forty-one children (53.9%) listened and retold the story *A Boy, a Dog, a Frog, and a Friend,* and 35 children (46.1%) listened and retold the story *Frog on His Own*. A research assistant orthographically transcribed the recorded story retelling samples and then coded them using SALT (Miller & Iglesias, 2008). To ensure reliability of transcriptions and coding, two independent transcribers checked 100% of the transcriptions. Any discrepancies between transcriptions or codes were resolved by consensus.

Analysis

Interrater reliability. Before examining validity evidence, we examined the reliability of the scores for this sample by having a third rater independently score 15 randomly selected story retells (19.6%). The agreement between the two SELPS scaled scores within a .5 score difference was 100%. The weighted κ value yielded a value of .88, which indicates almost perfect agreement between the two SELPS scaled scores. The agreement between the subscale scores within a range of ±1 score point was 100% and yielded a κ value of 1 for syntactic complexity scores (perfect agreement), a value of .72 for grammatical accuracy scores (substantial agreement), a value of .81 for verbal fluency scores (almost perfect agreement), and a value of .63 for lexical diversity scores (substantial agreement). Kappa values were interpreted according to Landis and Koch (1977).

Validity evidence. As evidence of the external relationships between the SELPS and other methods of LP assessment, we computed bivariate correlations among the measures. The relationship between the SELPS scaled scores and the teacher ratings of LP were examined using nonparametric correlations, as this represents the most appropriate analysis for the ordinal-level data.

Large significant correlations were predicted between the language sample measures and the subscale and scaled scores, which would support the validity of the inferences about English LP based on the SELPS; small correlations would suggest that language sample variables and the SELPS scores measure different constructs. It was further expected that large significant correlations between the teacher ratings and the SELPS scores would support the validity of the inferences about English LP based on the SELPS; small correlations would suggest that the teacher ratings and the SELPS measure different constructs.

Results

Descriptive statistics (mean, standard deviation, median, interquartile range, skew, standard error of skew, and minimum and maximum scores) for the SELPS subscale and scaled scores, the language sample variables, and the teacher ratings are presented in Table 2. The minimum observed subscale and scaled score on the SELPS was 2, which indicated that all of the participants were past the nonverbal period of L2 acquisition. The means and standard deviations of the language sample variables at each SELPS level are presented in Table 3. Due to the small sample size for some of the SELPS levels, statistical analysis was not used to evaluate whether there were significant mean differences for all language sample variables across different scaled scores. However, descriptive statistics indicated an increase in MLU_w, SI, and NDW and a decrease in UG and PMW with the increasing SELPS scores. There was a slight increase in the mean of PMW from the SELPS scaled score of 3.5 (M = .10) to the SELPS scaled score of 4.0 (M = .18), possibly due to the increase in the length of retells.

Pearson product–moment correlations were calculated to investigate whether there was a linear relationship between

Table 2. Descriptive statistics for the Spanish–English Language Proficiency Scale (SELPS) subscale and scaled scores, language sample variables, and teacher ratings.

Measure	Variable	*M*	*SD*	Median	Interquartile range	Skew	Std. error of skew	Min	Max
SELPS scores	SC	3.84	0.43	4.00	0	–2.86	.28	2.00	4.00
	GA	3.87	0.74	4.00	1	–0.20	.28	2.00	5.00
	VF	4.21	0.84	4.00	1	–0.56	.28	2.00	5.00
	LD	3.79	0.50	4.00	0	–2.38	.28	2.00	4.00
	OSS	4.34	0.78	4.50	1	–1.04	.28	2.00	5.00
Language sample variables	MLU_w	6.59	1.27	6.60	-	–0.66	.28	2.31	9.59
	SI	1.10	0.08	1.10	-	0.92	.28	1.00	1.37
	UG	0.42	0.24	0.40	-	0.25	.28	0.00	1.00
	PMW	0.15	0.09	0.14	-	0.75	.28	0.00	0.47
	NDW	61.75	23.57	62.50	-	0.43	.28	15.00	127.00
Teacher ratings		3.71	0.83	4.00	1	–0.56	.28	1.00	5.00

Note. MLU_w = mean length of utterance in words; SI = subordination index; UG = ungrammaticality index; PMW = percentage of maze words to total number of words; NDW = total number of different words.

Table 3. The mean and standard deviation of the language sample variables at each SELPS level.

	SELPS overall scores											
	2.0 (*n* = 2)		3.0 (*n* = 6)		3.5 (*n* = 8)		4.0 (*n* = 16)		4.5 (*n* = 8)		5.0 (*n* = 36)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
MLU_w	4.16	2.61	5.31	1.86	5.96	0.88	6.49	1.00	6.51	0.70	7.14	1.04
SI	1.04	0.05	1.06	0.08	1.05	0.07	1.08	0.07	1.05	0.04	1.13	0.08
UG	0.83	0.24	0.68	0.16	0.57	0.24	0.51	0.17	0.43	0.15	0.27	0.20
PMW	0.40	0.11	0.21	0.03	0.10	0.06	0.18	0.07	0.16	0.10	0.13	0.08
NDW	24.00	12.73	43.67	16.18	48.25	19.84	57.88	15.57	57.13	16.20	72.61	24.52

the subscale and scaled scores and the corresponding language sample variables. Because the language sample variables represented continuous data, for the purposes of the analyses, the subscale and scaled scores were also treated as continuous variables. Using the Sidak (1967) adjustment of the Bonferroni procedure to control for Type I error, a *p* value <.001 was required for significance.

Results indicated statistically significant correlations between MLU_w and the syntactic complexity scores, $r = .42$, $p < .001$, between UG and the grammatical accuracy scores, $r = –.66$, $p < .001$, and between NDW and the lexical diversity scores, $r = .43$, $p < .001$. However, the syntactic complexity subscale scores were not significantly correlated with SI, $r = .28$, $p = .014$, and the verbal fluency scores were not significantly correlated with PMW, $r = –.35$, $p = .002$. The SELPS scaled scores were significantly correlated with MLU_w ($r = .53$, $p < .001$), UG ($r = –.63$, $p < .001$), and NDW ($r = .50$, $p < .001$), but not with SI ($r = .34$, $p = .003$) or PMW ($r = –.36$, $p = .002$).

To investigate whether there was a linear relationship between the SELPS and the teacher ratings, we conducted Spearman's rho correlations between the SELPS subscale and scaled scores and the teacher ratings of English LP. Using the Sidak-Bonferroni approach to control for Type I error, a *p* value <.004 was required for significance. There were moderate and significant correlations between the SELPS scaled score and the teacher ratings of LP, $\rho = .46$, $p = .001$. In addition, the teacher ratings were moderately and significantly correlated with the grammatical accuracy subscale scores, $\rho = .47$, $p = .001$, the verbal fluency subscale scores, $\rho = .43$, $p = .001$, and the lexical diversity subscale scores, $\rho = .36$, $p = .001$, but not the syntactic complexity subscale scores, $\rho = .46$, $p = .029$.

Discussion

Study 2 examined construct validity evidence based on the relationship between the SELPS scores and two theoretically related variables: language sample measures and teacher ratings of English LP. LSA was selected as an unbiased measure in the assessment of oral language skills in sequential bilingual children (Gutiérrez-Clellen & Simon-Cereijido, 2009; Restrepo, 1998). Teacher ratings were selected as a screener of L2 proficiency that is similar in scoring format (i.e., a rating scale; Bedore et al., 2011; Gutiérrez-Clellen & Kreiter, 2003; Restrepo & Gutiérrez-Clellen, 2001). No existing standardized measures of L2 proficiency were selected because they confound oral language skills and academic achievement (MacSwan & Rolstad, 2006; Pray, 2005).

There were significant moderate-to-large correlations between the SELPS subscale and scaled scores and the corresponding language sample measures, indicating that the SELPS and the language sample measures assessed similar language abilities. However, it is possible that there was some degree of nonnormality in the distributions of the subscale and scaled scores that subsequently reduced variability in the observed data and affected the magnitude of the correlation analysis. The SELPS offers a general description of L2 proficiency levels and is intended to be used as a screener; thus, higher correlations with the language sample measures were not expected.

The SELPS levels captured the continuum of oral LP skills as demonstrated by the increase in the MLU_w, SI, and NDW, and by the decrease in the UG and PMW, with increasing overall SELPS scores. The decrease in the UG and PMW was expected and indicated that children with greater L2 proficiency made fewer grammatical errors and used fewer maze words than those with lower L2 proficiency.

When we consider the results in relation to other studies, there are some similarities in the performance of children at a higher level of LP based on the language sample measures. Specifically, children at the highest SELPS level had, on average, 73% grammatical utterances and 13% maze words in their story retells, which is comparable with the mean percentage of grammatical utterances ($M = 79.64\%$) and the mean percentage of maze words ($M = 20\%$) in story tells elicited by a wordless picture book in proficient Spanish–English bilingual children ages 4;0–6;11 (Fiestas et al., 2005; Fiestas & Peña, 2004). However, children at the highest SELPS level demonstrated, on average, a higher MLU_w ($M = 7.14$) than reported in the previous research on proficient Spanish–English bilingual children ($M = 5.44$; Fiestas & Peña, 2004). It is possible that differences in the story elicitation technique influenced these results. Fiestas and Peña (2004) used a story generation technique based on a wordless picture book rather than story retelling; thus, the higher MLU_w in the current study was expected.

Among the language sample measures, the SI and PMW were not significantly correlated with the scores from the corresponding SELPS subscales. The nonsignificant correlations between the syntactic complexity subscale and the SI may have been influenced by a lack of subordinate clauses in the children's story retells in which the total number of clauses was almost equal to the number of T-units. The stories used in the story retell tasks included a linear temporal progression of events and had comparable values on the SI (1.23 and 1.22) that were slightly higher than the mean of the SI in children's retells, indicating that the story scripts could have influenced the amount of subordination produced by the children. The syntactic complexity subscale, however, significantly correlated with the MLU_w. It is possible that the PMW was not as sensitive to differences in verbal fluency across LP levels as expected. Other measures of verbal fluency, such as percentage of maze words per minute, should be examined in future studies.

Teacher ratings of English LP were considered as a second external validation measure that assesses language skills. The SELPS scaled scores and teacher ratings of English LP on a 5-point Likert-type scale were moderately and significantly correlated. The current finding corroborates those of Gutiérrez-Clellen and Kreiter (2003), who reported moderate correlations between teacher rating and the proportion of grammatical T-units in Spanish–English bilingual children, and those of Bedore et al. (2011), who found significant associations between teacher ratings and language scores on English semantics and morphosyntax scores. This finding indicates that higher scores on the SELPS are more likely to be associated with higher scores on the teacher scale and, conversely, lower scores on the SELPS are more likely to be associated with lower scores on the teacher scale.

Overall, these results provide empirical evidence regarding the validity of inferences about L2 proficiency based on the SELPS scores. Although the same language samples were used for the validation phase and to assign the SELPS scores, the study validated that the rating matched the observed language behavior. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) recommend that evidence to support the validity of a measure should be considered in terms of several aspects or sources of evidence. Based on the complexity of validity evidence, it is crucial to recognize that no single study can examine all possible aspects to support the intended purpose of the measure and thus, a series of studies should be conducted to build a sufficient validity argument to support the desired interpretations.

LSA is considered the gold standard in language assessment (Heilmann et al., 2008, 2010; Miller et al., 2006). Therefore, high and significant correlations between the subscale and overall scaled scores with the language sample measures indicated that the operationalization of the construct of LP into different subcomponents is valid. Further, validation of the ratings with objective and quantifiable measures provided evidence that the scale is serving its intended purpose while providing a more efficient and ecologically valid assessment method. Moreover, the increase in values by increase in proficiency stage indicates that the scale is sensitive to changes in LP.

Study Limitations

Several limitations of the current study warrant discussion. The SELPS required raters to have sufficient training and background knowledge about the different aspects of L2 proficiency. Interrater differences may introduce some challenges in terms of establishing reliability and consistency of scoring. Thus, it is recommended that raters receive a substantial amount of training and demonstrate an adequate level of interrater reliability before administering the task and scoring it. Further, it is possible that raters may need refresher trainings to prevent rater drift in reliability.

The SELPS subscale scores range from 1 to 5; however, further investigations are necessary to examine whether this range is sufficient to capture the differences in L2 proficiency levels. None of the participants recruited for this study received a score <2, which restricted the variability in the observed scores and may have subsequently influenced the results of the statistical analyses. Future studies should include participants with a wider range of L2 proficiency; however, including children who are in the nonverbal period of L2 acquisition may not add much to the validity analysis. In addition, the participants had some exposure to the story retell task as a part of their academic experience; thus, the elicited language productions were not influenced by the novelty of the task. Younger children and children with no prior schooling experiences may not perform well on this task. Different methods may be necessary for these children. Also, the current scale requires validation with older children.

Finally, the use of the SELPS to rate a child's native language is not appropriate and has not been validated for this purpose. Therefore, the use of the scale for both languages in a bilingual child needs further study. In addition, raters should be exposed to native speakers of each language to ensure that they can differentiate between L1 and L2 proficiency.

Conclusions

The SELPS was designed as a criterion-referenced rating scale of L2 proficiency that assesses syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity based on two story retell tasks. The scale can be used as a screener of English LP in sequential bilingual children ages 4–8 to determine whether a child has sufficient skills in their L2 to be further tested in English. The scale is not intended to diagnose language impairment in sequential bilingual children because it is based on the framework of typical L2 development. Further, language disorders may impact performance on this measure; thus, a child's language use must be documented as well as any concerns about language development as a critical component of the LP assessment process.

The story retell task was selected due to its relative universality and the general structural organization across cultures. The task incorporates both language comprehension and language production (Pickert & Chase, 1978;

Skarakis-Doyle & Dempsey, 2008) and is more likely to elicit longer utterances and complex grammatical structures than any other elicitation technique (Gutiérrez-Clellen, 2002; Restrepo & Gutiérrez-Clellen, 2001). The results of this study indicated that the two stories elicited comparable levels of L2 proficiency based on the SELPS scores. Thus, the SELPS represents a valuable tool for test–retest conditions in the assessment of L2 proficiency by eliminating a memory effect due to repeated use of the same storybooks.

This study provided empirical evidence of validity of score inferences based on the SELPS when it is evaluated against spontaneous language samples measures, the gold standard in oral language assessment (Heilmann et al., 2008, 2010; Miller et al., 2006), and teacher ratings of LP. Results indicated that the SELPS and language sample measures, including $\mathrm{MLU_w}$, NDW, and UG, are measuring similar language skills. Further, results suggested that the SELPS and teacher ratings tap related aspects of L2 proficiency.

Acknowledgments

The data in this research were collected with support from Grant R324A080024 by the U.S. Department of Education, Institute of Education Science. Portions of the results were presented at the Symposium on Research in Child Language Disorders in 2009 and the annual convention of the American Speech-Language-Hearing Association in 2009 and 2011.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American National Standards Institute. (1969). *Specifications for audiometers*. New York, NY: Author.

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., . . . Zhang, J. (2012). *The condition of education 2012* (NCES 2012-045). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449–465.

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing, 28,* 343–365.

Ballard, W. S., Tighe, P. L., & Dalton, E. F. (2005). *Idea Proficiency Test*. Brea, CA: Ballard & Tighe.

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 17*(1), 1–29.

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 15*(5), 489–511.

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., . . . Gilliam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition, 15,* 616–629.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge, UK: Cambridge University Press.

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge, UK: Cambridge University Press.

Brown, R., & Fraser, C. (1963). The acquisition of syntax. In C. N. Cofer & B. Musgrave (Eds.), *Verbal behavior and verbal learning: Problems and processes* (pp. 158–197). New York, NY: McGraw-Hill.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Chondrogianni, V. V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism, 1*(3), 318–345.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism, 9,* 1–43.

De Avila, E., & Duncan, S. (2005). *Language Assessment Scales*. New York, NY: McGraw-Hill.

Embertson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Ervin-Tripp, S. M. (1974). Is second language learning like the first? *TESOL Quarterly, 8*(2), 111–127.

Fiestas, C. F., Bedore, L. M., Peña, E. D., & Nagy, V. J. (2005). Use of mazes in the narrative language samples of bilingual and monolingual 4- to 7-year old children. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 730–740). Somerville, MA: Cascadilla Press.

Fiestas, C. F., & Peña, E. D. (2004). Narrative discourse in bilingual children. *Language, Speech, and Hearing Services in Schools, 35,* 155–168.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619.

Gazella, J., & Stockman, I. J. (2003). Children's story retelling under different modality and task conditions: Implications for standardizing language sampling procedures. *American Journal of Speech-Language Pathology, 12,* 61–72.

Golberg, H., Paradis, J., & Crago, M. (2008). Lexical acquisition over time in minority L1 children learning English as a second language. *Applied Psycholinguistics, 29*(1), 41–65.

Goldstein, B. A. (2006). Clinical implications of research on language development and disorders in bilingual children. *Topics in Language Disorders, 26*(4), 305–321.

Gorin, J. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21–35.

Gutiérrez-Clellen, V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education, 13,* 175–197.

Gutiérrez-Clellen, V. F., & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24,* 267–288.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language, 30,* 234–245.

Håkansson, G., & Nettlebladt, U. (1996). Similarities between SLI and L2 children: Evidence from the acquisition of Swedish word order. In C. E. Johnson & J. H. V. Gilbert (Eds.), *Children's language* (pp. 135–151). Mahwah, NJ: Erlbaum.

Hakuta, K. (1978). A report on the development of grammatical morphemes in a Japanese girl learning English as a second language. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 132–147). Rowley, MA: Newbury House.

Heilmann, J., Miller, J., Iglesias, A., Fabiano-Smith, I., Nockerts, A., & Andriacchi, K. (2008). Narrative transcription accuracy and agreement in two languages. *Topics in Language Disorders, 28,* 178–187.

Heilmann, J., Miller, J., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology, 19,* 154–166.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development, 35*(1), Serial No. 134. Chicago, IL: University of Chicago Press.

Itoh, H., & Hatch, E. (1978). Second language acquisition: A case study. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 76–88). Rowley, MA: Newbury House.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 2*(4), 151–170.

Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In M. T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected Proceedings of the 2008 Second Language Research Forum* (pp. 32–47). Somerville, MA: Cascadilla Proceedings Project.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, Second Edition*. Circle Pines, MN: AGS.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services, 2000–2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*(2), 28–41.

Kohnert, K. (2004). Processing skills in early sequential bilinguals. In B. A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers* (pp. 53–76). Baltimore, MD: Brookes.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145–164.

Krashen, S. D., & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom*. Hayward, CA: Alemany Press.

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

Larkin, E., Restrepo, M. A., & Morgan, G. (2007, November). *How the bilingual syntax measure measures up. A comparison of Spanish/English proficiency scores to the Stanford English Language Proficiency Test and parent report*. Presentation at the annual convention of the American Speech-Language-Hearing Association, Boston, MA.

MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contributions to a deficit view of linguistic minorities. *Hispanic Journal of Behavioral Sciences, 22*(1), 3–45.

MacSwan, J., & Rolstad, K. (2006). How language tests mislead us about children's abilities: Implications for special education placements. *Teachers College Record, 108*(11), 2304–2328.

MacSwan, J., Rolstad, K., & Glass, G. V. (2002). Do some school-age children have no language? Some problems of construct validity in the pre-LAS Español. *Bilingual Research Journal, 26*(2), 213–238.

Martin-Beltrán, M. (2010). Positioning proficiency: How students and teachers (de)construct language proficiency at school. *Linguistics and Education, 21*(4), 257–281.

Mayer, M. (1973). *Frog on his own*. New York, NY: Dial Books for Young Readers.

Mayer, M., & Mayer, M. (1971). *A boy, a dog, a frog, and a friend*. New York, NY: Dial Books for Young Readers.

McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice, 21,* 30–43.

Miller, J. F., & Iglesias, A. (2008). Systematic Analysis of Language Transcripts, Bilingual SE (Version 2008) [Computer software]. Madison, WI: SALT Software, LLC.

Muñoz, M. L., Gillam, R. B., Peña, E. D., & Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools, 34,* 332–342.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

North, B. (2000). *The development of a common framework scale of language proficiency: Vol. 8*. Bern, Switzerland: Peter Lang.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics, 31,* 3–28.

Paradis, J., & Crago, M. (2000). Tense and temporality: Similarities and differences between language-impaired and second-language children. *Journal of Speech, Language, and Hearing Research, 43*(4), 834–848.

Peters, A. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Pickert, S. M., & Chase, M. L. (1978). Story retelling: An informal technique for evaluating children's language. *The Reading Teacher, 31*(5), 528–531.

Pienemann, M. (1998). *Language processing and second language development. Processability theory*. New York, NY: Benjamins.

Pray, L. (2005). How well do commonly used language instruments measure English oral-language proficiency? *Bilingual Research Journal, 29*(2), 387–409.

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

Restrepo, M. A., & Gutiérrez-Clellen, V. F. (2001). Article use in Spanish-speaking children with specific language impairment. *Journal of Child Language, 28*(2), 433–452.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23,* 497–526.

Robinson, P., Tin, S. C., & Urwin, J. J. (1995). Investigating second language task complexity. *RELC Journal, 26*(2), 62–79.

Saville-Troike, M. (1987). Dilingual discourse: The negotiation of meaning without a common code. *Linguistics, 25*(1), 81–106.

Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

Skarakis-Doyle, E., & Dempsey, L. (2008). Assessing story comprehension in preschool children. *Topics in Language Disorders, 28*(2), 131–148.

Skehan, P. (1998). *A cognitive approach to language testing*. Oxford, UK: Oxford University Press.

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78,* 260–329.

Tabors, P. O. (2008). *One child, two languages. A guide for early childhood educators of children learning English as a second language*. Baltimore, MD: Brookes.

Wiig, E. H., Secord, W. H., & Semel, E. (2006). *Clinical Evaluation of Language Fundamentals-4 Spanish edition*. San Antonio, TX: The Psychological Corporation.

Woodcock, R. W., Muñoz-Sandoval, A. F., Ruef, M. L., & Alvarado, C. G. (2005). *Woodcock-Muñoz Language Survey*. Itasca, IL: Riverside.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27,* 567–595.