III

Evidence Base for the QUILS: ES

The following is an excerpted and lightly edited version of Chapter 9 from the *User's Manual for the Quick Interactive Language Screener™: English–Spanish (QUILS™: ES): A Measure of Vocabulary, Syntax, and Language Acquisition Skills in Young Bilingual Children*. This paper provides information on the normative sample behind the QUILS: ES, the development and testing of the items on the QUILS: ES, the validity and reliability of the QUILS: ES, the types of scores the QUILS: ES provides, and the results of other psychometric analyses. Readers may quote from this material provided their use is accompanied by the following credit line: Excerpted with permission from Iglesias, A., Golinkoff, R.M., de Villiers, J., Hirsh-Pasek, K., & Wilson, M.S. (2021). *User's manual for the Quick Interactive Language Screener™: English–Spanish (QUILS™: ES): A measure of vocabulary, syntax, and language acquisition skills in young bilingual children*. Baltimore: Brookes Publishing. Copyright © 2021 by Paul H. Brookes Publishing Co., Inc. www.quilscreener.com

9 Details on the Research Behind the QUILS: ES

Ratna Nandakumar

This chapter provides technical details on the studies conducted in the development of the QUILS: ES, a bilingual Spanish–English screener. A monolingual English version of the QUILS has also been developed. Technical data for the QUILS are reported in the User's Manual for that version. (See www.quilscreener.com for more information.)

Normative Sample

The following section describes the normative sample for the QUILS: ES.

Inclusion Criteria

The normative sample for the QUILS: ES included children with no reported visual or hearing difficulties who were screened in their child care centers, preschools, kindergartens, and Head Start programs in Massachusetts, Pennsylvania, Delaware, Florida, and Nebraska. Children who spoke a language other than English or Spanish were not included in the sample. The Language Questionnaire (see Figures 6.1 and 6.2) or school information was used to determine the degree to which English or Spanish was used in the child's home and/or school environment. Because the normative sample was designed to be representative of dual language learning Spanish–English children in this age range in the United States, it likely includes some children who had language disorders.

Sample Composition

The final normative sample for the QUILS: ES was made up of 362 children (184 girls, 177 boys, 1 unreported). This included 69 three-year-olds, 159 four-year-olds, and 134 five-year-olds. Children's ages ranged from 3;1 to 5;11 years (M = 4;8; standard deviation [SD] = 0;9). For the children in the norming sample, information on socioeconomic status (SES) was provided either in the form of mothers' self-reported educational attainment or by enrollment in a low-income child care center. The majority of the children tested were from low-SES families (79.5%), and 20.4% of the children were from mid-SES families (see Table 9.1). The percentage of mid-SES families approximates the percentage reported in the 2014 U.S. census data for Hispanic women. Hispanic women

Table 9.1. Composition of the norming sample for the QUILS: ES

	Final norming sample
Total *N*	362
Age	
3-year-olds: *n* (%)	69 (26.3)
4-year-olds: *n* (%)	159 (43.9)
5-year-olds: *n* (%)	134 (37.0)
Mean age (years): *M* (*SD*)	4;8 (0;9)
Gender	
Male: *n* (%)	177 (48.9)
Female: *n* (%)	184 (50.8)
No reported: *n* (%)	1 (.2)
SES	
Low: *n* (%)	288 (79.5)
Mid: *n* (%)	74 (20.4)

Key: SD, standard deviation; *SES*, socioeconomic status.

with children under 18 having an education level of an associate's degree and above was 26.1% in 2015 (National Center for Education Statistics, n.d.). However, that figure includes women who achieved a degree later in life. Rate of completion of bachelor's degrees or higher among Hispanic women in the years from 2006 to 2016 is between 12.9% and 16.6% (U.S. Census Bureau, 2017).

Demographic data for race were available for 66.6% of the final bilingual sample. Of those who reported this information, 55.8% were white, 6.6% were black/African American, 1.4% were multiracial, 0% were Asian, and 1.9% were other races. In addition, 82% of parents reported whether or not their child was of Hispanic origin and, of those, 91.2% of children were of Hispanic origin.

Developing the Items on the QUILS: ES

The creation of the items included on the QUILS: ES was based on extensive review of the research on children's language development, 3 through 6 years of age, and study of the most effective techniques to measure children's language abilities. (For more information on item development, see Chapter 3.) In addition, the development team was attentive to racial, ethnic, and cultural differences. For example, the team was mindful from the start that speakers of AAE would be tested, and children might be speaking any of several varieties of Spanish. Thus, all items included in the QUILS: ES had to contain words or linguistic structures that would not be biased against speakers of AAE or different Spanish dialects.

Another factor the development team kept in mind during item creation was ensuring that each item could be visually depicted in a way that young children could understand. For instance, verbs referring to mental state, such as *think/pensar* or *know/saber,* could not be visually represented. The verbs chosen entailed visible actions, such as *unlock/abrir.* Furthermore, the characters portrayed in the QUILS: ES vary in age, race, gender, and disability status.

Field Testing

Field testing included the recruitment process and preparation of the sites for the First Item Tryout and the Second Item Tryout.

Recruitment Process and Preparation of Sites

The development team's three labs (University of Delaware, Temple University in Pennsylvania, and Smith College in Massachusetts) worked with preschools and child care centers in those areas to recruit sites for pilot testing. Data in the norming sample also came from Nebraska and Florida. Researchers or schools with particular interest in preschool Spanish–English DLLs recruited participants in those areas from preschools and child care centers and were trained by the development team's experienced personnel using the Field Testing Guide, which served as a starting point for this User's Manual.

Administrators were shown screenshots of the software and given instructions on how to use the program to administer the screener. They practiced using the software and giving the screening instructions prior to working with children. Administrators were directed to e-mail development team staff at the main pilot testing sites with questions or problems with screening. After they completed screening a group of children, administrators sent the raw data to the development team staff at the University of Delaware for analyses.

Chapter 3 describes in detail how the QUILS: ES was developed over the course of 5 years, covering the four main phases in the QUILS: ES development process: 1) Item Development and pilot testing, 2) First Item Tryout, 3) Second Item Tryout, and 4) Creation of the Final Version of the QUILS: ES. Second Item Tryout—the source of the final version of the QUILS: ES—began in January 2014 and was completed in March 2015. This chapter briefly summarizes these phases and reports the analyses conducted in construction of the screener, its subsequent testing at laboratory and satellite sites, and its finalization. (See Chapter 3 for additional information about the item development process.)

Item Development and Pilot Testing

Before bilingual First Item Tryout testing, the development team tested all items in Spanish with a sample of monolingual Spanish children. These children were spending the summer at a Head Start in Springfield, Massachusetts, while their parents, low-income migrant workers from Mexico and Guatemala, worked on local farms. Testing the sample of monolingual Spanish children helped determine whether the test items would work for the target age range and show developmental trends from 3 to 6 years. A sample of 27 children, ages 3 years (n = 8), 4 years (n = 10), and 5 years (n = 9), was tested. The Spanish items proved able to capture developmental change in language skills for the monolingual Spanish speakers in this age range.

First Item Tryout

The First Item Tryout began with 96 English items and their equivalents in Spanish. The development team needed to determine which sets of items would be equivalent in English and Spanish. However, children could not give all of the items in both languages, because it is likely that they would remember an answer given in one language when the same item was presented in the second. Instead, the 96 items in each

language were partitioned into Set A (48 items) and Set B (48 items). Then a given child received 96 items consisting of either Spanish A and English B, or Spanish B and English A. In this way, each child received equivalent items but never the same ones in English and Spanish. Half of the children received the first test in Spanish and then English, and half received English first and then Spanish.

Only children who scored between 1.5 (Mostly English) and 4.5 (Mostly Spanish) on the Language Questionnaire took the test. A total of 76 children ages 3 to 6 years in child care centers throughout the Northeast took part in First Item Tryout. The children were in three age groups: 3-year-olds (n = 20), 4-year-olds (n = 37), and 5-year-olds (n = 19) (see Table 9.2). Children were randomly assigned to receive Set A or Set B.

The children received the bilingual screener in two sessions, and both Set A and Set B were given in counterbalanced order, namely either English first or Spanish first. The Spanish portion of the screener was always given by testers who were either native or highly proficient Spanish speakers. Instructions for the Spanish and English sections were given in the target language of the screener. In circumstances where the child responded to the tester in the non-target language, the tester administered the instructions in a bilingual manner, providing instructions in the target language first and then repeating them in the non-target language. Children were administered both sections of the screener (English and Spanish) in 1 day or on separate days within 14 days.

First Item Tryouts on the bilingual screener were used to guide the assignment of items to each language to come up with equivalent sets. The data allowed the development team to choose which items behaved better in each language. The team assessed whether an item was satisfactory or not by examining performance on items against general child ability level across all of the items. They sought out items that behaved coherently, in that more linguistically able children in that language (defined by overall score) were more likely to pass them than less linguistically able children. Each item was examined to see if the children who passed it had a total score that exceeded the total score of the children who chose one of the foils instead of the correct answer. Using that method, items were chosen with the best discrimination between ability levels in each language, and the 96 items were partitioned again into two final sets—48 in Spanish and 48 in English—for Second Item Tryout.

Table 9.2. Composition of Piloting and Item Tryout sample populations

	First Item Tryout	Second Item Tryout
Total *N*	76	568
Age		
3-year-olds: *n* (%)	20 (30.39)	148 (26.05)
4-year-olds: *n* (%)	37 (38.56)	232 (40.8)
5-year-olds: *n* (%)	19 (31.05)	185 (32.6)
Not reported: *n* (%)		3 (.5)
Mean age (years): *M* (*SD*)	4;7 (0;9)	4;96 (0;90)
Gender		
Male: *n* (%)	38 (50.0)	278 (48.1)
Female: *n* (%)	38 (50.0)	285 (49.3)
Not reported: *n* (%)		5 (.9)
SES		
Low: *n* (%)	76 (100)	482 (84.8)
Mid: *n* (%)	0 (0)	81 (14.3)
Not reported: *n* (%)		5 (.9)

Key: SD, standard deviation; *SES,* socioeconomic status.

Second Item Tryout

The new 96-item version of the screener had 48 items chosen to be presented in Spanish and 48 items chosen to be presented in English, with 16 in each area (Vocabulary, Syntax, and Process), and 4 in each type within the areas. Half the children received the Spanish screener first, and half of the children received the English screener first. The final sample for Second Item Tryout came from preschools, child care centers, and Head Start programs in Massachusetts, Pennsylvania, Delaware, Florida, and Nebraska.

In total, 568 children were tested in the Second Item Tryout, an equal number of boys and girls. The majority of the children came from low-SES families (84.8%), and the remaining children were from mid-SES families (14.3%), with 5 being unreported. Table 9.2 shows the breakdown by age, gender, and SES.

During the bilingual Second Item Tryout, some children also were randomly assigned to receive one validity or reliability measure: 49 were tested on the Preschool Language Scales–Fifth Edition (PLS-5), 44 on the Bilingual English Spanish Oral Screener (BESOS), and 51 on the QUILS: ES retest. An additional 20 received the English version of the Peabody Picture Vocabulary Test (PPVT).

After completion of the Second Item Tryout, only children who received all the items on both versions of the test were included for further analyses. These 446 children provided the data for the Rasch analyses, a procedure used to remove items that were either redundant or nondiscriminating (see the Rasch Analyses section). The best 45 items in each language make up the final QUILS: ES, with the following number of items in each area by language:

English:

Vocabulary: 16 items

Syntax: 15 items

Process: 14 items

Spanish:

Vocabulary: 16 items

Syntax: 14 items

Process: 15 items

There are four items in each type within the areas, except for *Wh*-Questions (three per language section) and Converting Active to Passive (two items in English, three items in Spanish).

The final QUILS: ES items in Spanish and English are presented in Tables 9.3 and 9.4; the tables show the areas, types, and items in the sequence in which the items are presented in each langauge section. The correct answers for all items appear in a gray box.

Order of Testing

For most of the sample, the order in which the screeners were given was recorded. Of the total 345 recorded, 175 children were given the Spanish section first and 170 children were given the English section first. Analysis of the data using analysis of variance (ANOVA) suggests that the effect of order was irrelevant once other factors such as age and SES were considered. The development team therefore recommended that the sections be given in the order likely to be most comfortable for the child.

Validity

Validity of an instrument is examined to ensure that a test is measuring what it purports to measure. For the QUILS: ES, the development team examined construct validity, or whether the screener actually measures language development. This also entailed a statistical test (Cronbach's alpha) of whether the items formed a coherent set. The QUILS: ES was also assessed for convergent validity, which answers the question "Does children's performance on the QUILS: ES correlate with their results on other established language assessments of Spanish and English in DLLs?"

Construct Validity

Construct validity is an index of whether a test measures the abilities that it is designed to measure. That is, the screener or assessment must be based on phenomena that expert researchers, teachers, and other educators regard as linguistically significant and educationally meaningful for children in the age range being examined. Without adequate theoretical and empirical backing to establish construct validity, no screener or test can be considered adequate. The foundation for the construct validity of the QUILS: ES is explained in Chapter 2, which describes the theoretical and empirical bases of item and item type selection for the QUILS: ES.

A test must also have internal integrity; that is, the items on the test must intercorrelate or form a coherent set, even though the items may vary in difficulty. To ensure this for the QUILS: ES, an analysis called Rasch modeling was used, described later in this chapter. In exploring internal integrity, the goal is to identify which items serve the intended purpose and which items are poor at doing so or are redundant because other items test the same thing. Item response theory, tested for the QUILS: ES using Rasch modeling, provides a way to evaluate the worth of the individual items to the test as a whole. These studies are detailed in the Rasch Analyses section.

Convergent Validity

To establish convergent validity, results from existing measures were compared with QUILS: ES. Standardized language measures were administered within 4 weeks of QUILS: ES testing to establish concurrent validity. For an additional test of concurrent validity only with the English half of the QUILS: ES, the PPVT in English was used with a small sample of 20 children.

The PLS-5 was chosen to check concurrent validity for the QUILS: ES as it also provides both a Spanish and an English score. A subgroup of 44 children tested on QUILS: ES completed the English PLS-5 (Zimmerman et al., 2011) and the PLS-5 Spanish (Zimmerman, Steiner, & Pond, 2012)—including the Expressive Communication and Auditory Comprehension portions of the test. This group completed the full PLS-5 administered in English and in Spanish in counterbalanced order. The PLS-5 was either administered in one session, for a duration of approximately 40 minutes, or in two separate sessions of approximately 20 minutes each, depending on the child's attention and tolerance for testing.

The PLS-5 has two components, Expressive Competence (EC) and Receptive Competence (AC), and provides a total score in each language. To prepare the data for the validity analyses, a total score was derived for the QUILS: ES by adding together the 45-item scores in each language. To compare with the standard scores of the PLS-5 and PPVT, these totals were then converted to standard (Z) scores by age group. Bivariate

Table 9.5. Bilingual validity results: PLS-5

	PLS-5 English	PLS-5 Spanish	PPVT-4 English
QUILS: ES English	.693** (44)		.735** (19)
QUILS: ES Spanish		.449*** (44)	

**$p < .001$; ***$p < .01$.

Numbers in parentheses denote the sample size.

Key: PLS-5, Preschool Language Scales–Fifth Edition (Zimmerman, Steiner, & Pond, 2011, 2012); PPVT-4, Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007).

correlations between the QUILS: ES in English and the PLS-5 total English reveal a moderately high correlation ($r(44) = .693$, $p < .001$). Bivariate correlations between the QUILS: ES total Spanish scores and the PLS total in Spanish reveal a smaller but still highly significant correlation ($r(44) = .449$, $p < .002$). These data are all summarized in Table 9.5.

As part of the concurrent validity testing, 44 other children completed the QUILS: ES and the BESOS. This test, designed for ages 4 to 7 years, contains Morphosyntax and Semantics subtests in both English and Spanish (Lugo-Neris et al., 2015). The intercorrelations between the Spanish and English BESOS with the Spanish and English QUILS: ES are shown in Table 9.6. Because the BESOS has only been normed for ages 4 years and up, the analysis includes the 29 children (out of 44 total) who were older than 4 years.

The correlations reported for the relation between QUILS: ES and BESOS in Table 9.6 are only modest, although all but the semantics are still statistically significant. The content of each screener is quite distinct. The semantics items on the BESOS (e.g., analogies, categorization) are entirely different from those on the QUILS: ES, which taps knowledge of nouns, verbs, prepositions, and conjunctions. The BESOS–Morphosyntax (BESOS-MS) focuses on morphological markers that are specific to SLI in Spanish or in English (e.g., regular plural in English, adjective agreement in Spanish, and prepositional phrases in both languages). Although QUILS: ES includes prepositional phrases, it does not test morphology except for the past auxiliary and copula, because we tried to avoid areas of specific difficulty for speakers of some English dialects. Finally, the BESOS was specifically designed to screen children who might have a language impairment, whereas QUILS: ES was designed as a screener for all bilingual children in a classroom. For all these reasons, the two screeners are likely to be complementary in the picture they paint of language abilities.

Table 9.6. Concurrent validity with the BESOS

	Bilingual validity results: BESOS			
	BESOS-MS English	BESOS-S English	BESOS-MS Spanish	BESOS-S Spanish
QUILS: ES English	0.39* (29)	0.39* (29)		
QUILS: ES Spanish			0.37* (29)	0.22 (29)

*$p < .05$.

Numbers in parentheses denote the sample size.

Key: BESOS, Bilingual English Spanish Oral Screener (Peña et al., 2008); BESOS-MS, Bilingual English Spanish Oral Screener–Morphosyntax (Peña et al, 2008); BESOS-S, Bilingual English Spanish Oral Screener Spanish (Peña et al., 2008).

For additional concurrent validity only with English QUILS: ES, the PPVT-4 (Dunn & Dunn, 2007) in English was used. The total English QUILS: ES score correlated well with the English PPVT ($r(19) = .735$, $p < .001$) in a separate group of 20 children who received both (see Table 9.5).

Using Existing Tests With QUILS: ES Best Scores, Combining English and Spanish

The previous section compared children's performance on the QUILS: ES, correlating English and Spanish scores separately against a second, existing test for bilingual children. This section reports correlations of the QUILS: ES Best Scores with best scores from existing tests of bilingual children's language (de Villiers, 2015; Peña et al., 2018b). Best Scores capture the fact that a bilingual child's knowledge can be distributed between the child's two languages. The DLL may know some things well in one language and other things well in the second language (Peña et al., 2002). *The Best Score uses the maximum score on the individual types of language items from each language to get an overall view of the child's functioning.* Thus, Best Scores were computed from the types of language items tested in each language: *Wh*-Questions, Noun Learning, and so forth. The comparison of the types was carried out by looking at the proportions correct in each language, because the numbers of items per type varied (see Table 4.1 and related discussion in Chapter 4). Credit was given for the maximum proportion correct a child received for each type across the two languages. For example, if a student did better on the proportion of *wh*-questions correct in Spanish than English, that *wh*-score would be counted. If a child did better on the proportion of past auxiliary correct in English than Spanish, that score was counted, and so on. Finally, these type Best Scores were totaled to give total scores for each area. The total scores provide area Best Scores (e.g., Vocabulary Best Score, Syntax Best Score, Process Best Score) and Overall Best Scores. These were converted to *Z*-scores by age group to compare with the PLS-5.

A best score was also computed for the PLS-5 by taking the maximum score of Expressive Competence (EC) and Auditory Comprehension (AC) across the two languages and averaging them. No finer grained subcomponent scores other than AC and EC are available for the PLS-5. Table 9.7 shows the intercorrelations. The Overall Best Score also has concurrent validity against the overall best score of the PLS-5.

Table 9.7 also depicts the Best Scores on QUILS: ES correlated with the BESOS best scores, but only for the children age 4 years and older. For this small group ($n = 29$) of 4- and 5-year-olds, we created a best score on the BESOS by adding the maximum scores on BESOS–MS (Morphosyntax) and BESOS–S (Semantics). That BESOS best score also correlates with the Best Score on the QUILS: ES, although not very highly ($r(29) = .368$, $p < .05$), again potentially reflecting the very different contents of the two tests. Nevertheless, these results suggest that a very useful profile could emerge from testing children on both screeners.

Table 9.7. Best scores validity

	PLS-5 best score	BESOS best score
QUILS: ES	.486*	.368**
Best Score total	(44)	(29)

*$p < .01$; **$p < .05$.

Numbers in parentheses denote the sample size.

Key: BESOS, Bilingual English Spanish Oral Screener (Peña et al., 2008); PLS-5, Preschool Language Scales–Fifth Edition (Zimmerman, Steiner, & Pond, 2011, 2012).

Reliability

The reliability of a test asks whether the scores are stable for the same individuals when tested at different times. Another aspect of reliability is whether children's scores on the items cluster in meaningful ways. That is, children should pass items that reflect their ability and not pass a random selection of easy and hard items.

Test–Retest Reliability

A second session of QUILS: ES testing was administered 4–6 weeks after initial QUILS: ES testing to establish test–retest reliability. Children received both English and Spanish sections of the QUILS: ES after 4–6 weeks of their initial QUILS: ES session. English and Spanish sections of the QUILS: ES were administered in the same order in which the initial QUILS: ES was administered. As with the initial QUILS: ES session(s), for the retest, the two language sections of the QUILS: ES were given within 2 weeks of each other.

Table 9.8 represents the test–retest reliabilities for the Spanish and English sections of the QUILS: ES. Forty-four children were given the retest but complete data were only available from 39 tests in English and 42 in Spanish. The overall values shown in Table 9.8 are higher for English than for Spanish, but both are significant. The conventional view is that the value should be higher than .7. For English, the test–retest value (.88) falls well into an acceptable range, but the Spanish value (.59) is slightly lower than is conventionally acceptable.

Internal Consistency Reliability

Demonstrating that a test has internal consistency of its items is another metric of reliability. Cronbach's (1951) coefficient alpha is used to calculate internal consistency reliability. Cronbach's coefficient alpha provides a lower bound value of test reliability and is considered to be a conservative estimate of a test's reliability (Allen & Yen, 1979; Carmines & Zeller, 1979; Reynolds, Livingston, & Willson, 2009). Table 9.9 reports the Cronbach's coefficient alphas for both languages of QUILS: ES and their areas. These good-to-high coefficient values demonstrate that items are coherent in measuring the unidimensional construct underlying each area of the screener and also each language of QUILS: ES, making it a useful language comprehension screener for young DLLs of Spanish and English.

Interrater Reliability

Interrater reliability is an analysis that quantifies the amount of agreement between two or more raters of the same phenomenon, in this case student performance on a language screening. Measurement error is introduced into scores when different people

Table 9.8. Test–retest reliabilities for bilingual sample

	Overall score
Spanish test–retest	.59* (42)
English test–retest	.88* (39)

*$p < .001$.
Numbers in parentheses denote the sample size.

Table 9.9. Internal consistency of the QUILS: ES by language

	Bilingual reliability results			
	English (*n* = 417)		Spanish (*n* = 446)	
	No. of items	Cronbach's alpha	No. of items	Cronbach's alpha
Overall score	45	0.89	45	0.85
Syntax	15	0.75	14	0.69
Vocabulary	16	0.65	16	0.70
Process	14	0.82	15	0.70

administer or score a test on the same individual's performance differently. However, because the QUILS: ES administration and scoring are automated and by definition standardized, concerns regarding interrater reliability are minimized. Nevertheless, different testers could still induce effects indirectly, in terms of affecting the child's comfort level taking the test. The development team tested this proposition by comparing standard scores at the different sites at which testing occurred. ANOVAs were conducted comparing overall *Z*-scores on each language with between subject variables of gender, SES, and test location. There were no significant effects of test location or interaction between test location and other variables. Thus, any differences between individuals' scores on the QUILS: ES are not likely to be attributable to testing at different sites with different testers. The development team concluded that the QUILS: ES is sufficiently standardized that different testers have little effect.

How Were QUILS: ES Scores Derived?

This section describes how the standard scores and percentile ranks for the QUILS: ES were derived by the development team and the psychometricians working with them. Rationale for deriving cut scores is also explained. For information on the scoring of the QUILS: ES (e.g., point assignment for correct answers), see Chapter 7.

Best Scores

As discussed in Chapter 7, best practice suggests using a best score to determine where the child's skills rank relative to his or her peers, not just compared to the norms in each language. The QUILS: ES provides Best Scores in Vocabulary, Syntax, Process, and Overall. The Best Score is used to determine the risk for language delay. The QUILS: ES automatically computes Best Scores by taking each type in the screener and taking the better of the two (proportioned) scores from Spanish and English, then summing these selected type proportions together to get a Best Score per area. An Overall Best Score is derived by summing the Best Scores for each area and converting them into a standard score and percentile rank. Each Best Score per area can similarly be converted into standard scores and percentile ranks.

Generation of Standard Scores

The QUILS: ES Best Score standard scores are generated based on age norms and the QUILS: ES raw scores. The standard score reflects each child's performance as compared to the norms generated from the final norming sample of children at each age (3, 4, and

5 years). The norming sample was a subsample of 362 children from the Second Item Tryout sample, consisting of those children who completed both Spanish and English sections of the screener completely with no missing data. In terms of a match to census data, the ratio of SES levels in the sample is a reasonable match to the overall population of Spanish–English bilingual families in the parents' age range.

The standard scores derived from the Best Score proportions for each composite area of the QUILS: ES were normalized to the bell-shaped distribution in each area (Vocabulary, Syntax, Process) scores, and these area scores were produced for each of the three age groups in the standardization sample. Next, each area score variable was transformed so that its shape matched the bell-shaped curve with a mean of 100 and a standard deviation of 15. A scaled score was then created by summing the three standardized area scores, and the norming process was repeated on this scaled score to derive an Overall Best Score standard score. As with the Best Scores per area, the scaled score was transformed so that its shape matched the bell-shaped curve with a mean of 100 and a standard deviation of 15. Finally, Best Score norm tables were constructed by comparing each standard score to its corresponding proportion. The normative tables of the Best Scores of the QUILS: ES, with standard scores and percentile ranks for the Vocabulary area, the Syntax area, the Process area, and Overall, are presented in Appendix 9.A1–9.A12. This process of transforming raw scores to normalized standard scores represents the most common application of a "nonlinear area conversion" (Thorndike, 1982, p. 115). *Note:* The conversion process differs greatly from the QUILS standardization process due to the dual language nature of the QUILS: ES and the necessity for Best Scores. The Best Score standardization process should not be applied to the individual language sections within the QUILS: ES.

Generation of Percentile Ranks

Percentile ranks reflect where children's standard scores fall compared to other children of the same age. The QUILS: ES provides both standard scores and percentile ranks for the Best Scores. A principal advantage of the normalized transformations used with the QUILS: ES is that percentiles corresponding to identical standard scores are equal because they follow well-known properties of the bell-shaped curve. Thus, all normalized standard scores of 115 will hover around a percentile rank of 84, and all normalized standard scores of 130 will hover around a percentile rank of 98. Standard scores are associated with percentile ranks based on the child's raw score and age (see Appendix 9.A1–9.A12).

Individual Language Scores

Although the team cannot recommend using the individual language scores in assessing risk for language delay, teachers and administrators may want to know where a bilingual child's individual English and Spanish scores compare relative to scores of bilingual peers. For this purpose, English and Spanish raw scores were also individually converted to standard scores and percentile ranks based on the process of standardization without taking the Best Score proportion, and these are also provided in some of the Student reports. If a child's language performance is significantly different in one language, teachers and administrators may want to look at the individual scores for that language and see if there is a specific pattern or other indication of the type of concepts with which the child is struggling. Use caution when interpreting the scores precisely because the performance of bilingual children lies on a continuum, and so a relative weakness in one language may be balanced by a strength in the other. That is why Best Scores are used for follow-up recommendations.

Table 9.10. Rasch analyses for the QUILS: ES sample

Test	No. of items/ No. of people	Person measure			Item measure			Item Infit MNSQ
		Mean	*SD*	Range	Mean	*SD*	Range	Range
Bilingual English	45/446	–0.12	1.05	–2.3–4.1	0	0.81	–2.4–1.6	0.8–1.2
Bilingual Spanish	45/446	–0.31	0.89	–2.4–2.6	0	0.87	–1.5–1.6	0.8–1.3

Key: SD, standard deviation; MNSQ, mean-square.

Rasch Analyses

Rasch analyses were undertaken separately for English and Spanish on data collected as part of the Second Item Tryout. Rasch results were highly promising on the 48-item versions of each section. In the Rasch analyses with 48 items and 446 children, 3 items were removed from each section on the basis of their misfit values, resulting in a final number of 45 items in each language.

Table 9.10 shows the results of Rasch analyses for the QUILS: ES sample, for both Spanish and English sections. The Infit Mean-Square (MNSQ) values for items are within the expected range (0.8 and 1.3), denoting good fit of these items to the scale. The person mean and item mean are close to each other, denoting a good match between items and people. Item maps for the English and Spanish sections show that these tests discriminate well between children of varying abilities. For both English and Spanish sections, children are discriminated well from under 2 standard deviations below the mean to over 2 standard deviations above the mean. Moreover, items have satisfactory spread throughout the scale.

Evaluation and Reduction of Screener Bias Using Differential Item Functioning Analysis

Differential item functioning (DIF) analysis is typically performed to ensure that items are of similar difficulty across groups. The distinctive features of DIF analysis are the following: 1) it is done at the item level and 2) examinees are matched on their underlying ability in the two groups before comparing their performance on the item. DIF analysis is typically performed with respect to two groups at a time. For QUILS: ES, DIF analysis was performed with respect to gender. Language acquisition research has not found marked difference in typical children's language development by gender, although there is a bias in children with language delays because boys outnumber girls (Salameh, Nettelbladt, & Gullberg, 2002; Tomblin et al., 1997).

English Section

DIF analysis of the English screener displayed four items that performed significantly differently between boys and girls. Two items were easier for boys, and two items were easier for girls. As an equal number of items favored each group, it can be argued that DIF cancels out at the screener level resulting in neither of the gender groups being disadvantaged.

Spanish Section

Five items functioned significantly differently between boys and girls. Four of these items were easier for girls, and one item was easier for boys. It is virtually impossible to have a test completely free of bias, so we allowed these items to remain because this is a screener and not a high-stakes test.

Determination of Cut Scores

Cut scores were determined by considering the role of the QUILS: ES in screening children at risk for language impairment. A language screener should try to identify all children who might be at risk, because the cost associated with missing vulnerable children is greater than the cost of unnecessarily screening children who will pass a more comprehensive test. For that reason, the development team judged scoring below the 25th percentile on the Overall Best Score, below the 25th percentile in the Process Best Score or in both the Vocabulary and Syntax Best Scores to be a conservative estimate of risk, given that the population of children with language impairments is estimated to lie between 7% and 12% (Leonard, 2014; Tomblin et al., 1997). The cut scores are based on this 25th percentile.

Future Directions

At the time of publication of this User's Manual, the QUILS: ES developers are collecting systematic data on children who have documented language difficulties to provide estimates of the sensitivity and specificity of the QUILS: ES for a clinical population. In addition, the development team plans to expand the concept of the QUILS: ES to develop screeners for DLL children ages 2;0 to 3;0.