III

Evidence Base for the QUILS

The following is an excerpted and lightly edited version of Chapter 9 from the *User's Manual for the Quick Interactive Language Screener™ (QUILS™): A Measure of Vocabulary, Syntax, and Language Acquisition Skills in Young Children*. This paper provides information on the normative sample behind the QUILS, the development and testing of the items on the QUILS, the validity and reliability of the QUILS, the types of scores the QUILS 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 Golinkoff, R.M., de Villiers, J., Hirsh-Pasek, K., Iglesias, A., & Wilson, M.S. (2017). *User's manual for the Quick Interactive Language Screener™ (QUILS™): A measure of vocabulary, syntax, and language acquisition skills in young children*. Baltimore: Brookes Publishing. Copyright © 2017 by Paul H. Brookes Publishing Co., Inc. www.quilscreener.com

9 Details on the Research Behind the QUILS

This chapter provides technical details on the studies conducted in the development of the QUILS. The following information is for the monolingual English version of the QUILS. A bilingual English–Spanish version of the QUILS, the QUILS: ES, has also been developed. Technical data for the QUILS: ES 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.

Inclusion Criteria

The normative sample for the QUILS included children 3 (3;0) through 5 (5;11) years old 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 were not dominant in a language other than English were not included in the sample. The Language Questionnaire (included as Figure 6.1 in the User's Manual) was given as needed to confirm a child was sufficiently familiar with English. Since the normative sample was designed to be representative of monolingual 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 was made up of 415 children (216 female, 199 male). This included 130 three-year-olds, 154 four-year-olds, and 131 five-year-olds. Children's ages ranged from 3;04 to 5;11 years (M = 4;5; SD = 0;9). For 414 children, 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. (Information was not reported for one child.) The majority of the children tested were from low SES families (61.2%), and 38.6% of the children were from mid-SES families. The percentage of mid-SES families is close to the percentage reported in the 2014 U.S. census data for females age 18–39 years having an education level of an associate's degree and above (40.6%) (see Table 9.1).

Demographic data for race were available for 43.6% of the final monolingual sample. Of those who reported this information, 57.8% were White, 31.6% were Black/African American, 8.8% were multiracial, fewer than 1% were Asian, and 1% were other races.

Table 9.1. Composition of the norming sample for the QUILS (English monolingual version)

	Final norming sample
Total *N*	415
Age	
3-year-olds: *n* (%)	130 (31.32)
4-year-olds: *n* (%)	154 (37.11)
5-year-olds: *n* (%)	131 (31.57)
Mean age (years): *M* (*SD*)	4;5 (0;9)
Gender	
Male: *n* (%)	199 (47.95)
Female: *n* (%)	216 (52.05)
SES	
Low: *n* (%)	254 (61.20)
Mid: *n* (%)	160 (38.55)
Not reported	1 (.24)

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

Additionally, 45.9% of parents reported whether their child was of Hispanic origin; of those who reported on it, 23.3% of children were of Hispanic origin.

Developing the Items on the QUILS

The creation of the items included on the QUILS was based on extensive review of the research on children's language development, 3 through 6 years of age (including previous work by the QUILS development team), and study of the most effective techniques to measure children's language abilities. (For more information on type development, see Chapter 3 of the User's Manual.) 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 African American English, as well as English-proficient Hispanic children would be tested. Thus, all items included in the QUILS had to contain words or linguistic structures that would not be biased against speakers of African American English or Spanish-influenced English.

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* or *know,* could not be visually represented. The verbs chosen entailed visible actions. Furthermore, the characters portrayed in the QUILS show a variety of ages, races, and genders, and they are representative of a range of ability levels.

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 (at the 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. Researchers in other areas (Omaha, Nebraska, and Miami, Florida) recruited participants in those areas from preschools and child care centers and were trained by the development team's experienced personnel.

Screener administrators in each laboratory and in each of the satellite locations were trained using a Field Testing Guide consisting of the screening administration instructions included in 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 was developed over the course of 5 years, covering the four main phases in the QUILS 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. Second Item Tryout—the source of the final version of the QUILS—began in January 2014 and was completed in July 2014.

First Item Tryout

Following conventional evidence-based practice in psychometrics (Schmeiser & Welch, 2006), the development team tried out twice the number of items to appear in the final version of the QUILS. The original 96-item screener was conducted as the First Item Tryout with 306 monolingual English-speaking preschoolers from diverse socioeconomic backgrounds in Massachusetts, Delaware, and Pennsylvania. The sample consisted of 93 three-year-olds, 118 four-year-olds, and 95 five-year-olds (see Table 9.2). Based on this first round of data collection, Rasch and DIF analyses were conducted to identify the best 60 items out of the 96 used in the First Item Tryout. The 60 items scaled with age such that, on all items, 5-year-olds showed highest performance and 3-year-olds showed lowest performance.

Second Item Tryout

After the First Item Tryout was complete and analyzed to select the best and least redundant items, a 60-item version of the screener was administered to the final sample for norming from preschools, child care centers, and Head Start programs in Massachusetts, Pennsylvania, Delaware, Florida, and Nebraska. A majority of the children tested were from low-SES families (76.8%), and the remaining children were from mid-SES families (23.2%). There were 213 three-year-olds, 315 four-year-olds, and 146 five-year-olds (see Table 9.2). There were a total of 674 children tested in the Second Item Tryout.

After completion of the Second Item Tryout, problematic items were removed following analyses similar to those from the First Item Tryout. The final QUILS consists of the best 48 items culled from the two rounds of item tryouts. Table 9.3, not included in this excerpt but in the User's Manual for the Quick Interactive Language Screener™ (QUILS™), presents the final QUILS for monolingual English; it shows the areas, types, and items in the sequence in which the items are presented. The correct answers for all items are highlighted.

Table 9.2. Composition of First Item Tryout and Second Item Tryout sample populations

	First Item Tryout	Second Item Tryout
Total *N*	306	674
Age		
3-year-olds: *n* (%)	93 (30.39)	213 (31.60)
4-year-olds: *n* (%)	118 (38.56)	315 (46.74)
5-year-olds: *n* (%)	95 (31.05)	146 (21.66)
Mean age (years): *M* (*SD*)	4;55 (0;90)	4.47 (0;80)
Gender		
Male: *n* (%)	149 (48.69)	322 (47.77)
Female: *n* (%)	157 (51.31)	352 (52.23)
SES		
Low: *n* (%)	172 (56.21)	518 (76.82)
Mid: *n* (%)	134 (43.79)	156 (23.18)

Key: SES, socioeconomic status.

Validity

Validity of an instrument is examined to ensure the tool is valid for the specific purposes for which it will be used. For the QUILS, the development team examined construct validity, or whether the screener actually measures language development. This also entailed a statistical test (Cronbach alpha) of whether the items formed a coherent set. The QUILS was also assessed for convergent validity, which answers the question "Does children's performance on the QUILS correlate with their results on other established language assessments?"

Construct Validity

Construct validity demonstrates that a test measures the abilities that it is designed to measure. One of the most important requirements for an assessment is to have construct validity. That is, the screener or assessment test 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 is explained in Chapter 2 of the User's Manual, which describes the theoretical and empirical bases of item and item type selection for the QUILS.

A test must also have internal integrity; that is, the items on the test must form a coherent set that intercorrelates even though the items may vary in difficulty. To ensure this for the QUILS, an analysis called Rasch modeling was used, described later in this chapter. In seeking 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 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.

Table 9.4. Convergent validity coefficients

		Vocabulary standard score	Syntax standard score	Process standard score	Overall standard score
PPVT-4 standard score	Pearson correlation	.672**	.544**	.577**	.670**
	n	116	116	116	116
PLS-5 Auditory Comprehension standard score	Pearson correlation	.593**	.540**	.616**	.645**
	n	112	112	112	112

**Correlation is significant at the 0.01 level (2-tailed).

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

Convergent Validity

To assess convergent validity, 40 children from the Second Item Tryout were randomly assigned to also be tested on the Auditory Comprehension Subtest of the Preschool Language Scale, 5th Edition (PLS-5; Zimmerman et al., 2011), and 44 children from the Second Item Tryout were randomly assigned to be tested on Form A of the Peabody Picture Vocabulary Test, 4th Edition (PPVT-4; Dunn & Dunn, 2007). In addition, 72 children attending certain Head Start programs were administered both the PLS-5 and PPVT-4 as part of concurrent research projects at these schools. Both the PPVT-4 and the PLS-5 assess aspects of language development, have been normed on a representative population, and have demonstrated validity and reliability. Tests measure and emphasize different aspects of language; nonetheless, we would expect reasonably high correlations among the different tests. This was achieved for the QUILS, comparing it to these two well-known assessments (see Table 9.4). Table 9.4 presents the convergent validity coefficients for the group tested as part of the normative sample. The overall QUILS standard score correlates highly with PLS-5 and PPVT-4 standard scores. The area scores (i.e., Vocabulary, Syntax, and Process) also correlate highly with these assessments. Given that the PPVT-4 is a measure of vocabulary, the development team predicted that the area of the QUILS that would correlate most strongly with the PPVT-4 would be the Vocabulary area, and analyses confirmed this prediction. Thus, these results, together with the results of the construct validity tests, provide confirmation that the QUILS is measuring important aspects of language development for young children from the ages of 3;0 through 5;11. Children's performance on the QUILS predicts their performance on other omnibus tests of language development (e.g., PLS-5) as well as on tests that measure a single area (e.g., PPVT-4).

Reliability

The reliability of a test asks whether the scores are stable for one individual 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

Seventy-five of the students participating in the Second Item Tryout were randomly assigned to take the QUILS a second time. Score stability was examined by using the data

Table 9.5. Test–retest reliability coefficients (n = 75)

	Measure			
Measure	Vocabulary	Syntax	Process	Overall
Vocabulary	.71			
Syntax		.73		
Process			.69	
Overall				**.83**

gathered from these 75 students. The time interval between the first and second testing ranged from 3 to 5 weeks for nearly all participants. Table 9.5 presents test–retest reliability coefficients, as well as averaged coefficients calculated with Fisher's *z* transformation. The average coefficient for the overall QUILS is somewhat higher than coefficients for the three areas of Vocabulary, Syntax, and Process because of the large number of items overall (48 items) versus in the areas (16 items). The overall coefficient is .83, and the coefficients ranged from .69 for Process, to .71 for Vocabulary, to .73 for Syntax. In sum, test–retest coefficients indicate that standard scores from the QUILS possess reasonable stability across short time periods. Test–retest reliability is an important aspect of the QUILS for two reasons: 1) 3- to 5-year-olds generally show variability in their behavior, and 2) they are in a growth phase for language development. Thus, the QUILS is capable of reliably capturing children's performance.

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. Coefficient alpha provides a lower bound value of test reliability and is considered to be a conservative estimate of a test's reliability (Allyn & Yen, 1979; Carmines & Zeller, 1979; Reynolds, Livingston, & Willson, 2009). For the Vocabulary and Syntax areas, the coefficient alpha is .79 for each area. The coefficient alpha is .87 for the Process area and .93 for the overall QUILS. These good to high coefficient values demonstrate that items are coherent in measuring the unidimensional construct underlying each of the areas of the screener and also the overall QUILS as a language comprehension screener for young children.

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 administer or score a test on the same individual's performance differently. However, because the QUILS administration and scoring are automated and by definition standardized, concerns regarding interrater reliability are minimized. The development team tested this proposition by comparing standard scores at the different sites at which testing occurred. Results indicated that standard scores on the QUILS are no different between the sites at which testing was conducted. Thus, any differences between individuals' scores on the QUILS cannot be attributable to testing at different sites with different testers. The QUILS therefore has a standardized delivery.

Scores

This section describes how the standard scores and percentile ranks for the QUILS 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 (e.g., point assignment for correct answers, generation of raw scores), see Chapter 7 in the User's Manual.

Generation of Standard Scores

The QUILS standard scores are generated based on age norms and the QUILS raw scores. The standard score reflects each child's performance as compared to the norms generated from the final norming sample of children for each age (3, 4, and 5 years). The norming sample was a subsample of 415 children from the Second Item Tryout sample, stratified by SES status and gender to match the U.S. census (see Table 9.1) and with a more equal representation by age band.

The standard scores for the QUILS were normalized to the bell-shaped distribution in the 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 a standardized overall score. As with the area scores, 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, norms tables were developed by comparing each standard score to its corresponding raw score. The normative tables of the QUILS, with standard scores and percentile ranks for the Vocabulary area, the Syntax area, the Process area, and overall, are presented in Tables 9A.1–9A.4 in the User's Manual. This process of transforming raw scores to normalized standard scores represents the most common application of a "nonlinear area conversion" (Thorndike, 1982, p. 115).

Generation of Percentile Ranks

Percentile ranks reflect where children's standard scores fall compared to other children of the same age. The QUILS provides both standard scores and percentile ranks. A principal advantage of the normalized transformations used with the QUILS 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.

Determination of Cut Scores

Cut scores were determined by considering the role of the QUILS in screening children at risk for language impairment. A language screener should try to identify all children who might be at risk, as 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 to be a conservative estimate of risk, given that the population of children with language impairments is estimated to lie between 7% and 12% (Tomblin et al., 1997; Leonard, 2014). The cut scores are based on this 25th percentile.

Rasch Analyses

Separate Rasch analyses were conducted on each area of the QUILS as well as on the overall screener. Fit statistics for each of the areas and overall were close to the expected value of 1. Fit statistics were also investigated at the item level for all areas and overall. More emphasis was placed on the Infit Mean-Square (MNSQ) because it is a weighted measure and is sensitive to the study subjects near the item level on the underlying ability continuum. (A mean-square value of .5–1.5 is regarded as productive for measurement. See https://www.rasch.org/rmt/rmt162f.htm.) Infit MNSQ values for all items in each of the three areas and overall were within the expected range of –0.7 to 1.3.

Further evidence of reliability was established by high (1.6–10) person and item separation values, suggesting that each of the areas and the screener overall can successfully differentiate the different proficiencies of students and that items are well spread along the measures of difficulty.

The mean of "person ability measure" indicates if item difficulty is within the range of participants' abilities. The mean of item difficulties is set to 0. This is done to fix the scale within a calibration. For example, if the mean of person ability measure is 1, then the screener is easier for this sample of students. If the mean of person ability measure is –1, the screener is more difficult for this sample of students. For all three areas and the screener overall, the mean of person ability measure is a close match to the mean of item difficulties of 0, demonstrating an excellent match of items to the sample population.

Evaluation and Reduction of Screener Bias

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: 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, 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. The DIF analysis ensures that the items are not in themselves biased against one gender or another, regardless of ability level. For each area, each item was tested for DIF across gender groups. When an item shows significant DIF, it means the item displays different difficulty levels for boys and girls. Although some individual items show DIF in favor of one or another gender, it can be argued that since on balance, the DIFs cancel out, neither of the groups is disadvantaged by including these items (Nandakumar, 1993).