Contents

About the Authors

Nickola Wolf Nelson, Ph.D., CCC-SLP, BCS-CL, Professor Emerita, Department of Language, Speech, and Hearing Sciences, Western Michigan University, Kalamazoo, Michigan

Dr. Nelson was awarded the status of Professor Emerita in Speech, Language, and Hearing Sciences in 2016 after 35 years as faculty at Western Michigan University (WMU). During some of her years at WMU, she served as Associate Dean for Research and Director of the Ph.D. program in Interdisciplinary Health Sciences in the College of Health and Human Services. Dr. Nelson continues to conduct research and publish regarding language/literacy development and disorders. She is the first author of the *Test of Integrated Language and Literacy Skills™* (*TILLS™*; Paul H. Brookes Publishing Co., 2016), is Editor of *Topics in Language Disorders*, and is a Fellow of the American Speech-Language-Hearing Association and the International Academy for Research in Learning Disabilities. Dr. Nelson received the American Speech-Language-Hearing Foundation's Frank R. Kleffner Lifetime Clinical Career Award and Honors of the American Speech-Language-Hearing Association.

Barbara M. Howes, Ph.D., LMSW, Private Consultant, Cassopolis, Michigan

For the past 20 years, Dr. Howes has served in the social work field to assist overburdened families. She is Program Coordinator for a number of problem-solving courts and an adjunct faculty member in the Western Michigan University School of Social Work. Her research interests lie in the study of interdisciplinary practice.

Michele A. Anderson, Ph.D., CCC-SLP, Research Affiliate, Western Michigan University, Kalamazoo, Michigan

Dr. Anderson served as coordinator for the national, multiyear norming and validation study of the *Test of Integrated Language and Literacy Skills™* (*TILLS™*; Paul H. Brookes Publishing Co., 2016). She earned her Ph.D. from Western Michigan University. Her research interests include language and literacy assessment as well as work in adult neurorehabilitation.

About the Contributors

E. Brooks Applegate, Ph.D., Professor, Western Michigan University, Kalamazoo, Michigan

Dr. Applegate is the program director for the graduate programs in Evaluation, Research and Measurement at Western Michigan University. Dr. Applegate has extensive experience in research design, measurement, and applied statistics. He teaches graduate courses in psychometrics, structural equation modeling, and research methodology. Dr. Applegate actively participates in funded research and evaluation projects nationally and internationally. He has authored and coauthored more than 100 peer-reviewed journal articles and more than 85 peer-reviewed presentations.

Elena Plante, Ph.D., CCC-SLP, Professor, Department of Speech, Language, and Hearing Sciences, The University of Arizona, Tucson, Arizona

Inspired by her clinical experiences as a speech-language pathologist, Dr. Plante's research has focused on the assessment and treatment of specific language impairment and language-based learning disabilities. She also has contributed knowledge concerning the neurobiology of such disorders as specific language impairment, dyslexia, and auditory processing disorder. She is coauthor of the *Pediatric Test of Brain Injury* (PTBI; with G. Hotz, N. Helm-Estabrooks, & N. W. Nelson; Paul H. Brookes Publishing Co., 2010) and the *Test of Integrated Language and Literacy Skills™* (*TILLS™*; 2016; Paul H. Brookes Publishing Co.). Dr. Plante also coauthored a widely used textbook on communication disorders and more than 100 peer-reviewed journal articles, three of which have won editors' awards.

CHAPTER 4

Reliability and Validity of the Student Language Scale

In evaluating assessment instruments, evidence is needed to determine whether a tool is reliable in its consistency and valid for its stated purposes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In this section, we describe evidence that the Student Language Scale (SLS) measures the constructs and content it purports to measure (validity) and does so consistently (reliability).

SCIENTIFIC METHODS

First, we summarize methods used to evaluate the scientific evidence for reliability and validity of the SLS. According to traditional test theory, establishing validity of an assessment instrument includes procedures for identifying the constructs the tool will measure and the content for doing so (American Educational Research Association et al., 2014).

Theoretical Models and Expert Consultation

In early planning for the SLS, we considered how to gather information that could be gained from ethnographic interviews of teachers, parents, and students, which could serve as a precursor to curriculum-based language assessment and intervention (Nelson, 2010). We also considered how school social workers use ethnographic interviewing to gain insights into multiple perspectives when interviewing parents and teachers, as contributed by coauthor Barbara Howes, Ph.D., LMSW. In addition, we outlined the key constructs to be rated with the SLS by referring to the language levels-by-modalities model for the co-normed *Test of Integrated Language and Literacy Skills*™ (TILLS™; Nelson, Plante, Helm-Estabrooks, & Hotz, 2016a).

The next step was to generate a set of preliminary content items to represent the targeted constructs. To refine early versions of the SLS, we followed this step by consulting a panel of interdisciplinary scientific experts and parents regarding content of the scale (see the Appendix for acknowledgements). This group included experts who could comment on the cultural-linguistic appropriateness of candidate SLS items for a diverse population of students and families.

The quantitative analysis methods were planned in consultation with TILLS coauthor Elena Plante, Ph.D., CCC-SLP, and standardization project design and analysis expert E. Brooks Applegate, Ph.D. Many of the analyses described in this chapter were conducted by Dr. Applegate.

Data Gathering

Following try-outs with multiple preliminary versions of the SLS, we gathered quantitative data for analyzing the validity and reliability of the standardization version of the tool. This work occurred in conjunction with standardization research on the TILLS. The work was conducted from 2010 to 2015 and was coordinated by SLS coauthor Michele Anderson, Ph.D., CCC-SLP. Dr. Anderson also trained the test administrators for TILLS and interacted with parents and students regarding submission of forms and provision of incentives (i.e., parents, teachers, and students all received modest gift cards for helping us gather data) for both the SLS and TILLS.

As part of the broader TILLS standardization research (Nelson et al., 2016a), we tested more than 1,900 students from age 6 through 18 years with TILLS. In addition, we gathered information from parents and students using the SLS for the majority of this sample. Procedures for gathering informed parental permission/consent and child assent and for protecting identities were approved by two universities' Human Subjects Institutional Review Boards (Western Michigan University and the University of Arizona). In addition, some parents were asked to give permission for their children's teachers to complete SLS forms, and in such cases, teachers were asked to consent for their SLS data to be used for research purposes. Most of the students in this smaller sample for whom teacher SLS responses were gathered were part of a substudy in which we collected data longitudinally at approximately 6-month intervals over two or more time points across two school years. The exact numbers of participants in each analysis are detailed in Tables 4.1–4.8 later in this chapter.

Identifying Student Participants' Status

Identifying the sensitivity and specificity of a new assessment tool for screening purposes requires the independent establishment of each person's status with regard to the condition of concern: in this case, language/literacy disorder. Although this calls for a gold standard against which the new measure can be evaluated (Dollaghan, 2007), no gold standard exists that is widely accepted for identifying language and literacy disorders in school-age students. Rather, a set of procedures has been approved for establishing eligibility under the Individuals with Disabilities Education Improvement Act (IDEA 2004). The best option, therefore, was preexisting identification of a student as having such a disorder (i.e., language impairment, reading disorder, dyslexia, or specific learning disability in oral and/or written language) by a multidisciplinary school-based team or, in a few cases, by a private practitioner. Specific criteria for assigning students to different language status groups are outlined for each participant group in the following subsections.

Data used for assigning student participants to groups were based on parental completion of a student information form, which requested demographic information. This form, which was part of the approved parental consent packet, also asked parents to check *yes*, *no*, or *unsure* for a list of possible eligibility categories that could have been used for identifying their student as having a disability. In addition, the form asked parents to indicate whether anyone had expressed concerns about the student's reading or language ability (and if so, to explain). Finally, the form asked parents to indicate whether the student had an individualized education program (IEP) and, if so, whether the research team could have permission to see it. Test administrators, who had been trained by Dr. Anderson in test administration and human subjects protections, were asked to review the student information forms after parents completed them and to follow up on any responses that were unclear or inconsistent. When parents gave permission (and essentially all did), test administrators were asked to review available records, to provide scores on any related measures

of language/literacy skills, and to check off any goal areas that were targeted on the student's IEP.

Criteria for Normal Language Group Criteria for inclusion in the group of students with "normal language" (NL) were met if a student was progressing through school on time (i.e., had not repeated a grade), had never had language intervention, did not have a diagnosed disability (with a few minor exceptions), and was learning to read and write without difficulty. The minor exceptions were that if no other risk factors were present, 1) a student could have attention deficit disorder or attention-deficit/hyperactivity disorder (ADD/ADHD) and still be classified as having NL or 2) a student could have a speech impairment only (i.e., articulation problems affecting individual speech sounds or problems involving voice or fluency only, with no signs of language or literacy difficulty) and be included in the NL group. This decision about speech sound production disorders was made after preliminary analyses showed that students could be receiving services for misarticulating single sounds and still perform no differently on language/literacy tasks than students in the NL group on TILLS as long as no other exclusionary criteria were present. It is also important to note that consistent misarticulations are not counted as errors on TILLS subtests. Exclusion factors for the NL groups were if the student had been identified as having any other disability; if the student had been tested, treated, or referred for any language or literacy concerns; or if there were concerns about the student's hearing or vision (beyond visual acuity problems that were treatable with eyeglasses).

Criteria for Language Learning Disabilities Group Criteria for inclusion in the group of students with language learning disabilities (LLD) were met if the parent checked *yes* for any of the following from a list of diagnosed disabilities: language impairment, reading impairment, or learning disability. Students in the LLD group could not have been identified with any other disability, although they could have ADD/ADHD or speech impairment, as long as they had one of the eligible LLD categories comorbidly.

Criteria for Language and Literacy Risk Group Criteria for inclusion in the group of students with language and literacy risk (LLR) were met if a parent indicated that a student had been tested previously or received any services (e.g., "private therapy, special reading instructions, child study team services, or response-to-intervention services") for "any concerns about learning to use language or to read and write." This criterion was used to include any student in the LLR group who was receiving a second tier of multi-tier support services for language or literacy concerns as part of a response to intervention approach, but who did not meet criteria for the LLD group.

Three Additional Groups of Students in Special Populations In addition to the three primary groups, three "special population" groups were formed of students who were recruited to allow evaluation of the TILLS for use with students with diverse special needs. These groups were made up of students who had been identified as having autism spectrum disorder (ASD), being deaf or hard of hearing (DHH), or having mild intellectual developmental disability (IDD).

CONSTRUCT AND CONTENT VALIDITY: FOCUS GROUPS AND FACTOR ANALYSIS

When investigating preliminary versions of the SLS, several names and formats were used. An early version was entitled Language and Literacy Questionnaire (LLQ). It was much longer than the 12-item SLS, which became the final published version. That is, the

LLQ incorporated a 52-item rating scale, which was consistent with the advice of our panel of scientific experts to incorporate an orthogonal set (i.e., complete array) of fine-grained questions to ask how good the student was at language tasks that paired varied abilities at each language level with all four communication modalities (i.e., listening, speaking, reading, and writing). Among the 52 items were multiple items asking about sound/word abilities in different modalities—such as those involving reading, spelling, and phonological awareness—as well as multiple items asking about varieties of discourse within the curriculum—such as those that were narrative or expository. Thus, the questionnaire asked about reading, writing, and oral ability when using varied forms of expository, narrative, and social discourse and different forms of sound/word structure knowledge.

Focus Groups

After piloting the 52-item version, we held qualitative focus groups with teachers, parents, and students. Members of these focus groups indicated, almost unanimously, that the 52-item scale was too long. In addition, quantitative analyses showed low correlations between ratings of LLQ items and preliminary TILLS performance measures. This evidence also suggested that informants were interpreting the finer grained questions on this version in unreliable ways and in ways that were inconsistent with students' performance.

Thus, we trimmed the SLS length to 12 items and rewrote items to represent key content of the TILLS model as clearly and as simply as possible. At this point, we chose the word *story* to represent connected discourse for informants more clearly, rather than trying to differentiate narrative and expository discourse items on the rating scale. We also decided to ask only one question about spelling, describing it as "spelling words correctly when writing" (to differentiate it from performance on memorized spelling tests) and one question about word recognition or reading decoding, describing it as "figuring out new words while reading." The 12-item version of the SLS was used in the TILLS standardization research. At that point, it was called the Student Rating Scale (SRS). To avoid confusion with other tools with the same acronym, the SRS was renamed the Student Language Scale (SLS) but the items on the scale did not change.

Factor Analysis

Data from the standardization study were submitted to separate exploratory factor analyses (EFAs) for teachers, parents, and students. With 1,837 participants, the parent sample was large enough to conduct separate EFAs for the three age bands that are differentiated with the TILLS (Nelson et al., 2016a). That is, differential function analysis for the TILLS previously had identified three sets of core subtests of the TILLS that were best for identifying language/literacy disorders for three age bands of students: 6–7 years; 8–11 years; and 12–18 years. Because EFA results for the large set of parents' data showed minimal differences across the three age bands, final factor analyses for the SLS were conducted on collapsed age groups for each of the three informant types: teachers, parents, and students.

Details for the primary maximum likelihood EFAs for the SLS data showed support for a two-factor solution. Following oblique rotation (Promax, Power = 3), a clear pattern of loading on two factors was evident in the factor reference structure. The reference structure in Table 4.1 shows the relationship of factors 1 and 2 to each of the 12 items on the rating scale after partialling out effects of the factor. This table shows that the first factor (comprising Items 1–8) reflects the primary language/literacy construct, as measured with the TILLS; the second factor (comprising Items 9–12) reflects related

Table 4.1. Factor reference structure based on exploratory factor analyses

	Teachers (*N* = 325)		Parents (*N* = 1,837)		Students (*N* = 662)	
	Factor 1: language/ literacy	Factor 2: cognitive/ social	Factor 1: language/ literacy	Factor 2: cognitive/ social	Factor 1: language/ literacy	Factor 2: cognitive/ social
1. Listening Vocabulary	**.68**	–.01	**.74**	–.05	**.57**	–.02
2. Speaking Vocabulary	**.63**	.05	**.68**	–.02	**.41**	.12
3. Reading Decoding	**.60**	–.03	**.68**	–.04	**.44**	–.02
4. Spelling	**.48**	.16	**.53**	.08	.36	.10
5. Listening Comprehension	**.55**	.11	**.55**	.15	.31	.19
6. Oral Expression	**.59**	.09	**.49**	.21	**.39**	.14
7. Reading Comprehension	**.61**	.07	**.63**	.08	**.51**	–.02
8. Written Expression	**.53**	.17	**.55**	.15	**.45**	.09
9. Following Directions	.19	**.49**	.14	**.53**	.07	**.49**
10. Organization	.05	**.63**	–.02	**.67**	.07	**.44**
11. Attention	.03	**.67**	.03	**.67**	–.05	**.66**
12. Social	.12	.34	.02	**.42**	.04	.27

Note: Numbers in bold (≥ .39) are clearly loaded on the factor; numbers < .39 are not clearly loaded on the factor. The three exploratory factor analyses were conducted separately for ratings by teachers (interfactor correlation of .68), parents (interfactor correlation of .61), and students (interfactor correlation of .60).

cognitive/social skills. This same factor structure held for all three informant groups. The factor-structure correlation shows the relationship of both factors and their ability to define the observed variance of the SLS items. The interfactor correlations were 0.68 for teachers, 0.61 for parents, and 0.60 for students. These were considered low enough to represent different factors but high enough to reflect general relatedness.

SENSITIVITY AND SPECIFICITY EVIDENCE SUPPORTING VALIDITY FOR SCREENING

The validity of the SLS for the purpose of screening depends on how accurately it can predict which students are likely to have language/literacy disorders. Screening tools generally lead to pass/fail decisions about whether a person has a high risk of the disorder in question. Passing a screening procedure indicates low risk for the disorder; failing it indicates high risk. (See Chapter 3 for a description of how to use the SLS for screening.)

To evaluate the evidence supporting validity of the SLS for screening purposes, we sought a pass/fail cut score that would maximize both sensitivity (based on the criterion that 80% or more of students known by other measures to have LLD should fail the screening test) and specificity (based on the criterion that 80% or more of the people known NOT to have language/literacy disorders should pass the screening test). Making a cut score less stringent could maximize sensitivity for picking up all students who might have the disorder, but that could be at the expense of specificity; that is, it could overidentify students who do not have the disorder. Making a cut score more stringent,

on the other hand, could increase the measure's specificity, but at the expense of its sensitivity. In this case, the danger would be that setting the cut score too low would underidentify students who do have the disorder.

To identify the best cut score for achieving the optimal balance between sensitivity and specificity with the SLS, a table was created to quantify the percentages of participants identified with the SLS as being at risk of having LLD (i.e., failing the screening) or not having LLD (i.e., passing the screening) when ratings of different levels on different items were used. The goal was to find a cut score 1) that met the 80% criterion for both sensitivity (i.e., percentage of students known to have the disorder who fail the SLS) and specificity (i.e., percentage of students known not to have the disorder who pass the SLS) and 2) that yielded the highest percentages for both. By considering multiple options, we could see that scores of 5 or above generally were associated with typical development but that scores of 4 or below (when assigned by teachers or parents) were associated with other indicators of language/literacy disorder or risk.

Just one score of 4 or below was not enough to signal failing the SLS as a screener, but two scores less than 5 were.

We tested the cut score criterion of "2 or more less than 5 on the first 8" by creating tables for each of the informant groups so that we could evaluate sensitivity and specificity for each group if we were to alter the cut score. Table 4.2 provides the sensitivity and specificity results for teachers, Table 4.3 for parents, and Table 4.4 for students. Using these tables, it is possible to see how the percentages of pass/fail agreements with prior identification as NL, LLD, and LLR would shift if different cut score criteria were adopted. The shaded areas of the data table for teachers (Table 4.2) show the results when using the recommended cut score of two or more less than 5 on the first eight items as

Table 4.2. Sensitivity and specificity for the Student Language Scale (SLS) as completed by teachers

		Number of Items 1 through 8 rated below 5 by teachers								
Group	*N*	0	1	2	3	4	5	6	7	8
NL	203	158	24	5	5	4	1	2	2	2
		77.8%	11.8%	2.5%	2.5%	2.0%	0.5%	1.0%	1.0%	1.0%
LLD	66	4	1	4	3	5	5	12	9	23
		6.1%	1.5%	6.1%	4.5%	7.6%	7.6%	18.2%	13.6%	34.8%
LLR	48	7	0	8	1	3	4	8	3	14
		14.6%	0.0%	16.7%	2.1%	6.3%	8.3%	16.7%	6.3%	29.2%

Key: NL = normal language; LLD = language learning disabilities; LLR = language and literacy risks.

Teacher ratings of lower than 5 on two or more of the first eight items correctly identified 61 of 66 students (92%) with known LLD (sensitivity) and 182 of 203 students (90%) with NL (specificity). Sensitivity to LLR using this criterion was 41 of 48 students (85%), which is above the 80% threshold, suggesting that teachers tend to rate struggling students lower even if they are not yet identified as having a disability; this also fits criteria for screening.

Table 4.3. Sensitivity and specificity for the Student Language Scale (SLS) as completed by parents

		Number of Items 1 through 8 rated below 5 by parents								
Group	*N*	0	1	2	3	4	5	6	7	8
NL	1,290	917	148	83	53	38	19	10	10	12
		71.1%	11.5%	6.4%	4.1%	2.9%	1.5%	0.8%	0.8%	0.9%
LLD	239	18	18	25	28	28	28	31	22	41
		7.5%	7.5%	10.5%	11.7%	11.7%	11.7%	13.0%	9.2%	17.2%
LLR	192	53	17	29	24	17	7	19	12	14
		27.6%	8.9%	15.1%	12.5%	8.9%	3.6%	9.9%	6.3%	7.3%

Key: NL = normal language; LLD = language learning disabilities; LLR = language and literacy risks.

Parent ratings of lower than 5 on two or more of the first eight items correctly identified 203 of 239 students (85%) with known LLD (sensitivity) and 1,065 of 1,290 students (83%) with NL (specificity). This leads to the conclusion that parents' responses on the SLS *are* valid for screening for LLD. Sensitivity to LLR using this criterion was 122 of 192 students (64%). This suggests that parent ratings might not identify all borderline students who may have LLR.

the criterion for failing the screening. Note that this cut score, when using teacher data, shows both high sensitivity to LLD (92%) and high specificity for correct classification of NL (90%). Although not quite as high, the data table for parents (Table 4.3) also shows acceptable sensitivity (85%) for correct identification of students with known disorders and acceptable specificity (83%) for correct identification of students without disorders. Thus, evidence also supports using the SLS results for using parent SLS data to contribute to screening decisions about the need for further testing. Sensitivity (73%) and specificity (61%) results for students (see Table 4.4) were too low to support validity of using student self-ratings on the SLS for screening. These results are summarized in Table 4.5 for all three informant groups.

Table 4.4. Sensitivity and specificity for the Student Language Scale (SLS) as completed by students

		Number of Items 1 through 8 rated below 5 by students								
Group	*N*	0	1	2	3	4	5	6	7	8
NL	419	176	81	66	47	25	12	9	3	0
		42.0%	19.3%	15.8%	11.2%	6.0%	2.9%	2.1%	0.7%	0.0%
LLD	90	14	10	8	13	19	11	14	1	0
		15.6%	11.1%	8.9%	14.4%	21.1%	12.2%	15.6%	1.1%	0.0%
LLR	72	12	10	15	10	9	8	4	4	0
		16.7%	13.9%	20.8%	13.9%	12.5%	11.1%	5.6%	5.6%	0.0%

Key: NL = normal language; LLD = language learning disabilities; LLR = language and literacy risks.

Applying the cut-score criterion of lower than 5 on two or more of the first eight items for students' self-ratings correctly identified only 66 of 90 students with known LLD (73% sensitivity) and 257 of 419 students with known NL (61% specificity). Sensitivity to LLR using this criterion was 50 of 72 students (69%). Because these values all fall below the 80% criterion, we concluded that students' responses on the SLS are *not* valid for screening purposes. They could, however, provide meaningful insights into students' views on their own strengths and weaknesses.

Table 4.5. Sensitivity/specificity results for Student Language Scale (SLS) ratings by teachers, parents, and students

Informant	Sensitivity	Specificity
Teacher	61/66 = .92[a]	182/203 = .90[a]
Parent	203/239 = .85[b]	1,065/1,290 = .83[b]
Student	66/90 = .73	257/419 = .61

[a]Meets criterion for high sensitivity and specificity.
[b]Meets criterion for good sensitivity and specificity.

Teams also may want to consider the data for students in our LLR group when establishing cut scores for screening. Tables 4.2 (for teachers) and 4.3 (for parents) provide data for the LLR group showing that teachers are a bit more sensitive to risks among students with LLR who are struggling but not identified with LLD. The sensitivity to risk among the LLR group for teacher ratings using the usual cut score was 85.4%, but the sensitivity to risk among the LLR group for parent ratings of students in the LLR group using this cut score was only 63.6%. This suggests that using this cut score based

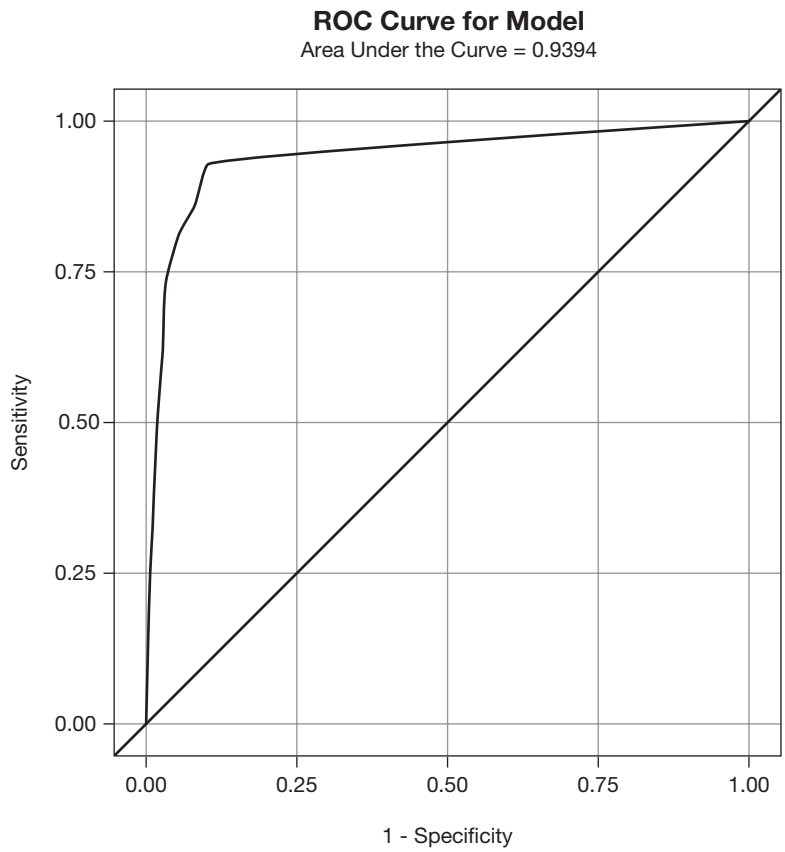

Figure 4.1. Receiver Operating Characteristic curve (ROC curve) analysis for teacher responses on the SLS showing their accuracy for differentiating two groups of students (i.e., language learning disabilities [LLD] or normal language groups) using a criterion of two or more ratings of less than 5 on the first eight items. *Note:* The shape of this curve and the area under the curve (.94) *do* support the validity of using teacher SLS ratings for making pass/fail decisions when screening for LLD.

on parent ratings only could miss some students at risk for language/literacy difficulty when screening. It underlies the higher validity of using teacher ratings as the best source for making decisions related to screening.

Another way to evaluate validity of the SLS for the purpose of screening is to use an analysis method for plotting hits and misses. This method, which is known as Receiver Operating Characteristic curve (ROC curve) analysis, involves plotting sensitivity by 1-specificity for different cut scores and measuring the area under the curve. More discriminative instruments with the most precise cut scores produce curves with a sharper turning point and maximum area under the curve, whereas less discriminative instruments produce flatter curves and smaller values for the area under the curve. The closer the ROC is to the diagonal reference line (which represents decisions no better than chance), the less discriminative is the information provided by the tool. Figures 4.1–4.3 show the ROC results for teachers, parents, and students respectively. Consistent with results of other analysis methods, the values for the areas under the curve were highest for teacher ratings (.94), next highest for parent ratings (.89), and smallest for student ratings (.74). The data for teacher and parent ratings supported the validity of the SLS for the purpose of making pass/fail screening decisions, whereas the data for students did not.

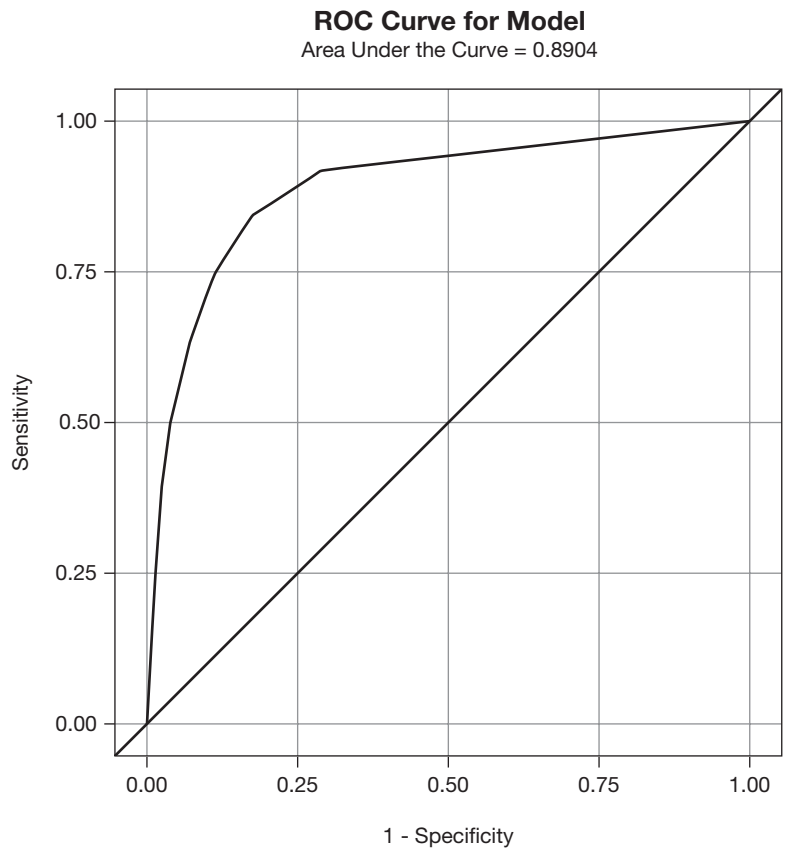

Figure 4.2. Receiver Operating Characteristic curve (ROC curve) analysis for parent responses on the SLS showing their accuracy for differentiating two groups of students (i.e., language learning disabilities [LLD] or normal language groups) using a criterion of two or more ratings of less than 5 on the first eight items. *Note:* The shape of this curve and the area under the curve (.89) support the validity of using parent SLS ratings for making pass/fail decisions when screening for LLD.

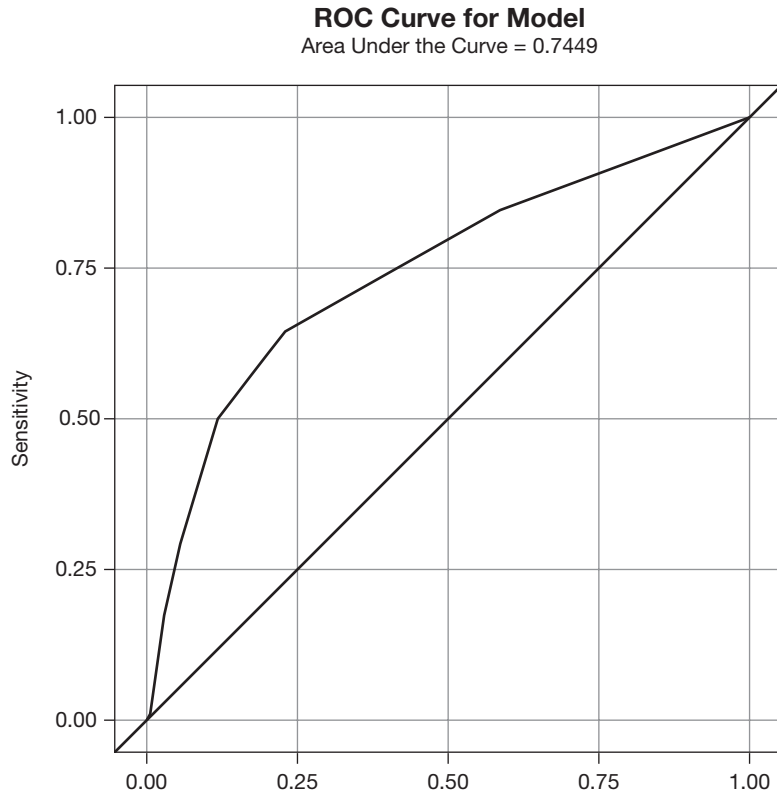

Figure 4.3. Receiver Operating Characteristic curve (ROC curve) analysis for student responses on the SLS showing that they lack accuracy for differentiating two groups of students (i.e., language learning disabilities [LLD] or normal language groups) using a criterion of two or more ratings of less than 5 on the first eight items. *Note:* The shape of this curve and the area under the curve (.74) do not support the validity of using student SLS ratings for making pass/fail decisions when screening for LLD.

EVIDENCE SUPPORTING VALIDITY FOR GATHERING MULTI-INFORMANT INPUT

Informant rating scales are valuable primarily for their ability to represent different perspectives on a phenomenon. Such responses should not be viewed as "correct" or "incorrect." In fact, one could say that informant ratings of strengths and weaknesses for a particular student have face validity because they are direct reflections of the informant's observations and evaluations, whether or not this person's ratings correlate to another person's ratings and to performance measures of the same phenomenon. That is, each person's ratings should be considered valid simply as indicators of that person's perspectives.

Still, it is helpful to know how closely informants' observations agree with each other and with a student's assessed performance in the areas in question. If items are valid for reflecting students' abilities, people should be able to provide ratings that agree, to some extent, with a direct measure of the student's performance of the ability in question. To evaluate this aspect of concurrent validity for the SLS, we examined binary correlations for data for subsets of informants' SLS ratings with theoretically aligned composite scores on the TILLS. Table 4.6 provides evidence of concurrent validity from these analyses. Although all of these correlation coefficients are statistically

Table 4.6. Correlation coefficient evidence for concurrent validity of Student Language Scale (SLS) ratings by teachers, parents, and students with related student performance on the *Test of Integrated Language and Literacy Skills*™ (*TILLS*™; Nelson, Plante, Helm-Estabrooks, & Hotz, 2016a)

	N_T	Pearson *r* for teachers	N_P	Pearson *r* for parents	N_S	Pearson *r* for students
SLS Items 3, 4 (sound/word items) with sound/word composite on TILLS	330	.671[a]	1,810	.595[a]	677	.299[a]
SLS Items 1, 2, 5–8 (sentence/discourse items) with sentence/discourse composite on TILLS	322	.720[a]	1,762	.570[a]	668	.302[a]
SLS Items 1–8 (language/literacy factor) with total TILLS	321	.752[a]	1,749	.613[a]	663	.329[a]
SLS Items 9–12 (cognitive/social factor) with total TILLS	323	.536[a]	1,762	.336[a]	677	.078[b]
SLS Items 1–12 (total SLS) with total TILLS	318	.725[a]	1,736	.572[a]	652	.279[a]

[a]$p < .001$.

[b]$p < .05$.

Key: N_T, number of teachers in each analysis; N_P, number of parents in each analysis; N_S, number of students in each analysis.

significant, only the teacher and parent ratings are correlated highly enough with the students' TILLS performance scores to meet standards for usefulness as representative measures. This finding supports use of teacher and parent SLS ratings for identifying areas of strength and weakness for planning. Whereas student ratings may provide insights into how a student perceives his or her abilities, correlations between students' ratings and actual performance on related TILLS subtests are weak, suggesting the need to interpret them with caution. However, student ratings may be taken at face value for how students feel about their abilities.

Of particular interest is the correlation of teacher ratings on Items 3 and 4 with the sound/word composite score on the TILLS. That is because Items 3 and 4 are the two SLS items that ask about sound/word-level skills (reading decoding and spelling), which are important diagnostic indicators of dyslexia. In this case, the correlation of SLS Items 3 and 4 and the sound/word composite on TILLS showed a significant correlation of .671. These results provide some support for use of the teacher–respondent SLS to screen for dyslexia. Evidence also supports use of the SLS as a screener for language/literacy disorders more generally. These results are signaled by the strong correlation between teacher ratings of sentence/discourse abilities on the SLS with actual student performance on the TILLS (.720 for the sentence/discourse composite on the TILLS and .752 for Items 1–8 with the total score on the TILLS).

In summary, multiple forms of evidence support the use of teacher ratings on the SLS (and parent ratings too, although not quite as strongly) for screening for language and literacy disorders, including dyslexia. Evidence also supports the validity of the SLS for providing perspectives on students' language/literacy strengths and weaknesses that can contribute to comprehensive evaluation for students with special needs and can support school–home communication for all students.

EVIDENCE SUPPORTING RELIABILITY

In addition to evidence of validity, evidence for the reliability of an assessment tool should indicate that it is acceptably stable, which means that it is internally consistent and also consistent across repeated uses and informants. Assessment devices with

higher reliability have less variability and more consistency; hence, they are easier to interpret with confidence.

One important measure of reliability for an assessment tool is its internal consistency. Internal consistency is an index of the degree to which multiple items on the tool are measuring the same thing. Internal consistency statistics that are closer to 1.0 indicate evidence of stronger reliability. This characteristic traditionally has been evaluated and reported as coefficient alpha. Coefficient alpha is problematic, however, in that it is influenced by the number of items in the analysis. Assessments with more items tend to have higher values for coefficient alpha, just by virtue of their length; therefore, to measure internal consistency of the SLS, we chose an alternative statistic, coefficient omega, which is less affected by length. Coefficient omega for the 12-item scale was .96 for teachers, .94 for parents, and .84 for students. For purposes of comparison, coefficient alpha values were nearly identical, at .96 for teachers, .93 for parents, and .84 for students. These results indicate strong internal reliability.

Another form of reliability, called intrarater (or test–retest) reliability, is an index of the consistency of scores found for the same person when comparing ratings by that person for the same student at two points in time. This is similar to test–retest reliability for a student repeating the same test. Intrarater reliability represents stability in the degree to which individual raters agree with themselves when rating the same student on two separate occasions. The duration between the two ratings should be close enough in time so that it is unlikely the student's status would have changed measurably, but far enough apart so that raters are unlikely to remember their prior ratings. Intrarater reliability for the SLS was measured by calculating interclass correlations on test–retest data that were gathered from the same informants at periods of less than 6 months. Table 4.7 summarizes these test–retest, intrarater reliability results for teachers, parents, and students. Good test–retest reliability is apparent for teachers and parents on all parts of the SLS. As in other analyses, teachers showed the highest coefficient (.92), parents were second (.83), and students were third (.61). An interesting outcome was that students were more consistent in rating themselves on the last four items at two time points than on the first eight. The first eight items rate language and literacy skills; the last four items rate related cognitive and social skills.

A third measure of reliability, which is called interrater reliability, evaluates agreement among different raters. The concept of interrater reliability is challenging when evaluating a multiple-informant scale because teachers' and parents' experiences with the same students may vary widely. This may lead to different opportunities to judge the abilities in question and also may encompass the possibility that abilities actually could be manifested differently in some contexts than in others, such as school compared to home. This is in contrast to the high interrater agreement one would expect when two examiners independently score the same set of responses on a traditional test. In fact, Achenbach, Krukowski, Dumenci, and Ivanova (2005) pointed out that results are more valuable in an additive way when they reflect different viewpoints and are not highly overlapping.

Table 4.7. Test–retest reliability estimates for Student Language Scale (SLS) ratings repeated over periods of less than 180 days

		Days between SLS					
	N	M	SD	Items 1–12	Items 1–8	Items 9–12	Interclass correlations (1,1)
Teacher	87	113	14.02	.92	.92	.84	.92
Parent	55	157	24.27	.84	.84	.80	.83
Student	49	122	36.03	.50	.42	.61	.61

Key: M = mean; SD = standard deviation.

Table 4.8. Interrater agreement on Student Language Scale (SLS) ratings

Informant agreement	*N*	ICC (1,1)
Teacher–Parent	107	.75
Teacher–Student	108	.61
Parent–Student	108	.66
Parent–Teacher–Student	108	.71

Note: Interrater agreement was conducted using Shrout and Fleiss (1979) intraclass correlations (ICCs) for the 12 items on the SLS rating scale.

Still, in the case of the SLS, if a student is experiencing difficulties in multiple situations, in and out of the classroom, multiple informants should be aware of those difficulties. Thus, one would expect some commonality across raters who are considering the same student in different settings. To analyze the degree to which teacher, parent, and student responses were correlated when each was rating the same items for the same student, we calculated binary correlations for the 12 scaled items on the SLS. As shown in Table 4.8, the correlations (calculated as Shrout & Fleiss [1979] intraclass correlations) are moderately strong and statistically significant, particularly between parent and teacher ratings. Correlations are weaker but still statistically significant between student and parent ratings and between student and teacher ratings.

SUMMARY OF EVIDENCE

In summary, the scientific evidence supporting the SLS indicates that ratings provided by teachers and parents using this tool are valid for the purposes of screening for language/literacy disorder and for describing a student's language strengths and weaknesses. Multi-informant input also can facilitate school–home communication regarding points of disagreement as well as points of agreement. Scientific evidence supports SLS reliability in terms of internal consistency, consistency of repeated ratings by the same informants, and relationship to performance data, particularly for ratings provided by teachers and parents. It is important to use evidence-based tools, such as the SLS, to contribute to decisions about students that may be critical to their access to education.