Understanding TILLS® Standardization and Psychometric Properties

In this handout, the developers of the *Test of Integrated Language and Literacy Skills* (TILLS®; Nelson, Plante, Helm-Estabrooks, & Hotz, 2016), led by Dr. Elena Plante, provide scientific evidence in response to concerns about the test's validity for identifying language and literacy disorders in school-age children and adolescents. Evidence supporting use of the TILLS for this purpose comes from peer-reviewed journal articles and alignment with the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014), which is considered the "gold standard" in guidance on testing by the American Psychological Association (www.apa.org/monitor/2014/12/testing-standards).

We will continue to add to this resource as further questions or concerns are raised.

The TILLS Normative Distribution

Concern 1: *Because the TILLS excludes children with disabilities from its normative sample, the TILLS does not reflect a bell-shaped distribution.*

This concern reflects a conflation of the separate concepts of a normative sample and normal (bell-shaped) distribution. Consider the facts:

- In a normative sample, most skills that show a developmental trajectory do not distribute as a bell-shaped curve at most of the ages over which the skill develops. For example:
  - Decoding skills will be very positively skewed at ages 5–6 years (more low scores and few high scores) because most children will not be able to decode most words they encounter on a test.
  - In contrast, by age 10, the distribution will be highly negatively skewed (more high scores, and few low scores) because most children can decode most words they encounter. This pattern also occurs for other skills that develop over the school-age years (e.g., articulation, morphology).
- Whether there are children with disabilities in the normative sample may affect the degree of skew, but it will not create a fully bell-shaped distribution at all ages for a skill that is not inherently normally distributed.
- Children with disabilities were not included in the TILLS normative sample because the authors wanted to develop a test that could discriminate children with language and literacy disorders from children who do not have these disorders. (Please see additional details in response to Concern 2.)

Concern 2: *The lack of children with disabilities in the normative sample invalidates the TILLS for diagnosing conditions that are not part of that sample.*

This is a common assumption that is based on an older and now outdated view about how diagnoses are made.

- The old assumption is that those with disabilities will score at the low end of the normal distribution and that distribution should include students with that disability for a "fair" comparison.

- There are two problems with this assumption:
  - There is a lack of evidence for the claim that the majority of children with language disorders score at the low end of a normal distribution. Spaulding, Plante, and Farinella (2006) demonstrated that this claim is untrue (doi:10.1044/0161-1461(2006/007)).
  - It is also untrue that students with learning disabilities, including dyslexia, typically score at the low end of test distribution. This can be confirmed by examining the mean difference between samples of typical and impaired children reported in the manuals of the many tests intended to assess these conditions.

The TILLS includes only typical children in its norms because its first purpose is to identify children with language and literacy disorders.

- This is conceptually akin to asking whether the language and literacy performance of a student being tested is consistent with that of typical children or departs from what is typical.
  - Achieving this purpose is enhanced by not confounding the standard for typical performance with the performance of children who are not typical in any number of ways. Indeed, we have demonstrated mathematically the negative impact on diagnostic accuracy when normative samples include children with disabilities involving language (Pena, Spaulding, & Plante, 2006; doi:10.1044/1058-0360(2006/023).
  - This is consistent with the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) Standard 5.8 which states that the normative sample "should contain groups with whom test users will ordinarily wish to compare their own examinees" (p. 104). In this case, the purpose of TILLS is to compare a student with a suspected language/literacy disorder with others who do ***not*** have the disorder. Hence, only those with typical language should be in the comparison group.
  - We note here that the current Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) do not require that normative samples contain a mix of individuals with normal and impaired skills, or even that the normative sample approximate a normal distribution.
- The current best practice for identifying the presence or absence of a disorder requires a method that accurately identifies those with the disorder as having the disorder and those without the disorder as being free of the disorder.
  - Several metrics (e.g., sensitivity/specificity, positive and negative likelihood ratios, ROC curve data) are available that express the level of diagnostic accuracy, given certain cut scores.
  - This is consistent with the current Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) statement that "A critical step in the development and use of some test is to establish one or more cut scores dividing the score range to partition the distribution of scores into categories" (p. 100) and Standard 5.23, which states that interpretation relative to cut scores "should be informed by sound empirical data" (p. 108).
  - The TILLS provides age-dependent cut scores for classification of scores as consistent or inconsistent with the presence of a language/literacy disorder. The accepted standard for sensitivity and specificity is 80% or higher (Vance & Plante, 1994, doi:10.1044/0161-1461.2501.15).

Table 3.3. Identification Core subtests that must be administered for the purpose of diagnosing language and literacy disorders at different age bands, with corresponding cut scores and sensitivity and specificity

Age band	Subtests that constitute Identification Core standard scores	Cut score	Sensitivity	Specificity
6;0–7;11 years	Vocabulary Awareness (VA) Phonemic Awareness (PA) Nonword Repetition (NWRep)	24	84%	84%
8;0–11;11 years	Vocabulary Awareness (VA) Nonword Spelling (NWSpell) Nonword Reading (NWRead) Written Expression–Discourse Score (WE-Disc)	34	88%	85%
12;0–18;11 years	Phonemic Awareness (PA) Nonword Spelling (NWSpell) Reading Comprehension (RC) Reading Fluency (RF) Written Expression–Word Score (WE-Word)	42	86%	90%

Table 3.3 above from the TILLS Examiner's Manual (p. 99) reports 84%–88% sensitivity (correctly identifying high percentages of students with a known disorder as having the disorder) and 84%–90% specificity (correctly identifying high percentages of those without a disorder as not having the disorder) for the core subtests and related cut scores provided for three different age groups.

- Table 3.4 below from the TILLS Examiner's Manual (p. 101) shows sensitivity by smaller age groups ranging from 83%–97% and specificity for those same age groups ranging from 81%–100%. These high specificity results, in particular, should allay fears that the TILLS might over-identify students as having disabilities.

Table 3.4. Sensitivity and specificity levels by age for all ages tested by the TILLS

Age groups	Sensitivity	Specificity
6-year-olds	84%	82%
7-year-olds	84%	86%
8-year-olds	97%	100%
9-year-olds	83%	81%
10-year-olds	81%	81%
11-year-olds	86%	82%
12-year-olds	83%	100%
13-year-olds	84%	86%
14- to 18-year-olds	87%	87%

TILLS does not include separate age norms between the ages of 14 and 18 years. This is because the skills tapped by the TILLS no longer show age-related change after age 13 years.

Sensitivity/specificity results provide the evidence behind the evidence-based use of TILLS as a diagnostic measure. When tests or diagnosticians disagree on a diagnosis, the discussion might fruitfully begin with the question "How accurately does the contrasting test (or method) identify this disorder?" The answer should be numeric and based on data regarding sensitivity/specificity. If not, it is not an evidence-based method. This requires that the normative sample be composed in a way that facilitates intended interpretations of test scores (AERA/APA/NCME, 2014, p. 97). The TILLS norms are constructed to facilitate determining student status as having or not having a language and/or literacy disorder.

Concern 3: *Lack of children with disabilities in the normative sample makes it impossible to compare the scores from the TILLS with other test scores.*

Generally speaking, it is risky to directly compare scores that are based on different normative samples, whether or not those samples include or exclude children with disabilities.

- It is tempting to compare performance across tests because the common scaling of scaled scores (e.g., mean of 100, SD of 15) makes it seem as if scores from different tests are equivalent and can be compared directly. However, differences in the relative aptitude of those in the normative sample of one test compared to another can yield substantially different standard scores.
    - For this reason, to truly have comparable scores requires co-norming, in which test scores that are relevant for identifying patterns of strengths and weaknesses are based on the *same* normative sample, using procedures referred to as "score linking" or "equating" (see the Standards for Educational and Psychological Testing, AERA/APA/NCME, 2014, pp. 97–98).
    - In the absence of co-norming and score equating (not available for most tests one might wish to compare), it is not clear whether otherwise hidden differences in the performance of different normative samples is driving differences in test scores. (A demonstration of this principle was published by Plante & Vance, 1994, illustrated in Figure 2; doi:10.1044/0161-1461.2501.15)

Scores Provided

Concern 4: *The TILLS converts raw scores into percentiles, rather than converting standard scores into percentiles, making it impossible to compare with other tests.*

There are three relatively common types of percentile scores provided by different behavioral tests: (1) percentiles based on the area under a bell-shaped curve, (2) normal curve equivalent (NCE) percentiles (i.e., based on the bell-shaped curve but adjusted to create equal appearing intervals), and (3) actual percentile ranks (based on the percentage of the normative sample whose scores fell below the score of the student in question).

- Perhaps the most common in educational applications is the percentile that is derived from test standard scores. These percentiles reflect the area under the normal (bell-shaped) curve that is at or below a given standard score (e.g., 50% of area under the normal curve falls below the mean; 16% of the area under a normal curve falls below a standard score of 85).
    - Because these percentiles are based on the standard scores, and not on the underlying normative distribution, they do not communicate information that is inherently different from the standard scores on which they are based.
    - In addition, because standard scores are intended to normalize the normative distribution (i.e., make it appear more bell-shaped than it is in reality), these types of percentiles can be misleading when the underlying distribution is not normal.
        - For example, the percentile of 16 that is linked to the standard score of 85 might over-estimate the actual percentage of student scores if the distribution for the normative sample is negatively skewed, or underestimate performance if the distribution is positively skewed.
        - However, because these percentiles are linked directly to the standard scores, and not to the normative sample, they can be readily calculated using any of the online calculators (e.g., onlinestatbook.com/2/calculators/normal_dist.html) that return the area under a normal curve.
        - This includes the option of calculating such commonly expressed percentiles from TILLS standard scores.

- In contrast, the TILLS provides a *percentile rank* that is directly referenced to the underlying normative sample. It expresses the percentile as the actual percentage of test takers in the normative sample who scored lower than the student at hand.
    - This percentile rank is not linked to the standard score, but to the raw scores of children in the actual normative sample. Therefore, this type of percentile provides information that ***supplements rather than duplicates*** the information provided by the standard score.
    - Nevertheless, the availability of these percentile rank scores does not prevent anyone from calculating the other type of percentile score, which duplicates the standard score, from TILLS or any other test's standard scores if they want them.
- The advantage of using percentile ranks derived from the actual normative distribution is that they align with the purpose of identifying patterns of strengths and weaknesses, by referencing the actual performance of the children in the normative sample at the same age as the current student. This is consistent with Standard 5.0 of the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) that indicates scores should be derived in a way that supports the purpose of the test (p. 102).

Test Administrators and Administration

Concern 5: *Because the TILLS covers literacy skills, it should not be administered by speech-language pathologists.*

The TILLS was developed to be used by a variety of professionals who have training in the administration and scoring of individually administered standardized tests. A variety of these professionals (including psychologists and speech-language pathologists) contributed testing results both for typical students and for students with previously identified language and literacy disorders during standardization. In addition:

- Segregation of oral and written skills by profession is inconsistent with the numerous professions that contribute to the research base of both oral language and written language (literacy), as well as the position statements of multiple state departments of education, and the American Speech-Language-Hearing Association (ASHA; doi:10.1044/policy.PS2001-00104).
- The conceptual division between language skills in the oral or print domains is inconsistent with the current state-of-science for language and literacy. Indeed, the most widely cited model of reading (The Simple View; Gough & Tunmer, 1986), which is supported by a large body of research, rests on an explanation of reading comprehension as the product of two dimensions: (1) decoding (closely tied to knowledge of word sounds) and (2) language (listening) comprehension (encompassing a wide range of oral language skills).

  Multiple large-scale studies of language dimensionality show high correlations among language in the oral and print modalities (e.g., Foorman et al., 2015a, doi:10.1037/edu0000026; Foorman et al., 2015b, doi:10.1007/s11145-015-9544-5). We have shown this same two-dimensional model with the TILLS data (Nelson, Plante, Anderson, & Applegate, 2022; doi:10.1044/2022_JSLHR-21-00534), and others have reported similar findings with datasets examining language comprehension in oral language and reading (e.g., Catts, Adlof, & Weismer, 2006; doi/abs:10.1044/1092-4388%282006/023%29).
- The TILLS allows comparison of oral and written language skills using co-normed subtests to reveal patterns of strengths and weaknesses that can differentiate dyslexia from developmental language disorder, which often co-occur (Catts et al., 1999, doi:10.1207/s1532799xssr0304_2; Catts et al., 2003, doi:10.1111/jcpp.13140).
    - As discussed previously, co-norming is essential to allow valid comparison of this nature. For example, by directly comparing Listening Comprehension to Reading Comprehension using co-normed measures on the TILLS, evaluation teams can gain information that can help them

distinguish patterns consistent with relatively pure dyslexia (in which Listening Comprehension may be higher than Reading Comprehension) from developmental language disorders characterized by problems of language comprehension in both oral and written modalities.

Concern 6: *Virtual administration of the TILLS is not valid*.

Practitioners are wise to be skeptical about virtual administration, as different formats of testing are not guaranteed to return the same result. For this reason, prior to releasing a telepractice version of the TILLS, the authors conducted a validation study to assure that the two methods produced comparable outcomes (Nelson & Plante, 2022; doi:10.1044/2022_LSHSS-21-00056).

- The validation study found that the scores yielded the same diagnostic decisions 96% of the time.
  - Scores were highly correlated for all but one subtest (Nonword Repetition at younger ages). The authors concluded that extra caution is warranted when administering and interpreting this particular subtest, which requires good audibility on both the administrator's and test taker's sides.
  - Recommendations concerning optimized technology and setup are provided along with the Tele-TILLS administration materials.
- The conduct of the study of virtual administration of the TILLS® (Tele-TILLS) and its recommendations are consistent with the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) Standard 5.7 regarding changes in administration for subgroups of test takers (p. 103) and the need to caution test administrators about comparability of scores from the modified versions of the test.

Elena Plante, Ph.D., CCC-SLP
Nickola Wolf Nelson, Ph.D., CCC-SLP, BCS-CL
Astrid Pohl Zuckerman, M.S.
Michele A. Anderson, Ph.D., CCC-SLP

October 5, 2024

Acknowledgment
The TILLS was validated with support from Grant R324A100354 from the U.S. Department of Education, Institute of Educations Sciences. However, the opinions expressed herein are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Biographical Information

Elena Plante, Ph.D., CCC-SLP, is Professor in the Department of Speech, Language, and Hearing Sciences at the University of Arizona. For more than 35 years, she has published articles concerning assessment methods and how norm-referenced tests can and cannot be used in valid ways to address clinical concerns. Her 1994 article established the standard for sensitivity and specificity in the field of speech-language pathology. Four of her other journal articles have won Editor's awards. Dr. Plante's numerous professional recognitions include being named a Fellow of the American Speech-Language-Hearing Association (ASHA), winning the ASHA highest honors for career achievements (Honors of the Association), and an honorary doctorate (Psychology) from the University of Bergen, Norway.

Nickola Wolf Nelson, Ph.D., CCC-SLP, BCS-CL, is Professor Emerita in the Department of Language, Speech, and Hearing Sciences at Western Michigan University. Dr. Nelson is the first author of the Test of Integrated Language and Literacy Skills (TILLS; 2016) and the textbook, *Language and Literacy Disorders: Infancy Through Adolescence* (2010; Pearson/Allyn & Bacon). Her publications include journal articles on language and literacy assessment and intervention, two of which have won Editor's awards. She was Editor-in-Chief of *Topics in Language Disorders* (2005–2018) and is a Fellow of the American Speech-Language-Hearing Association (ASHA) and recipient of the Kleffner Clinical Career Award and Honors of ASHA.

Astrid Pohl Zuckerman, M.S., is a Ph.D. candidate in developmental psychology at The Ohio State University. Her research focuses on risk and promotive factors of early language and literacy skills for later reading outcomes, both at the child and family level. Prior to graduate school, Pohl Zuckerman was Acquisitions Editor for Brookes Publishing, including development and publication of the first edition of the Test of Language and Literacy Skills (TILLS). She is currently part of the research team working on the second edition of TILLS and the Student Language Scale (SLS).

Michele A. Anderson, Ph.D., CCC-SLP, received her master's and doctoral degrees from Western Michigan University. She was Project Coordinator for the national validation studies of the Test of Integrated Language and Literacy Skills (TILLS). Dr. Anderson's research includes the dimensionality of language and literacy, the role of verbal working memory in language assessment, and procedures for training phonological awareness. She has taught child language development courses at Western Michigan University and given numerous national presentations on topics related to school-age child language assessment. She is an author of the Student Language Scale (SLS) and part of the research team working on the second edition of TILLS and the SLS.