



Test of Integrated Language and Literacy Skills™ (TILLS™)

TECHNICAL MANUAL

by

Nickola Wolf Nelson, Ph.D., CCC-SLP, BCS-CL
Western Michigan University
Kalamazoo

Elena Plante, Ph.D., CCC-SLP
The University of Arizona
Tucson

Nancy Helm-Estabrooks, Sc.D., CCC-SLP, BC-ANCDS
Western Carolina University
Cullowhee, North Carolina

and

Gillian Hotz, Ph.D., CCC-SLP
University of Miami
Miami, Florida

with contributions by

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

and

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



Baltimore • London • Sydney



Paul H. Brookes Publishing Co.

Post Office Box 10624

Baltimore, Maryland 21285-0624

www.brookespublishing.com

Copyright © 2016 by Paul H. Brookes Publishing Co., Inc.

All rights reserved.

"Paul H. Brookes Publishing Co." is a registered trademark of

Paul H. Brookes Publishing Co., Inc.

Test of Integrated Language and Literacy Skills, TILLS, and  **are trademarks of Paul H. Brookes Publishing Co., Inc.**

Typeset by Scribe Inc., Philadelphia, Pennsylvania.

Manufactured in the United States of America

by Potomac Printing Solutions, Inc., Landsdowne, Virginia.

The standardization research for the TILLS was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A100354 to Western Michigan University. The opinions expressed are those of the authors and do not represent views of the Institute or of the U.S. Department of Education.

This manual accompanies other components of the *Test of Integrated Language and Literacy Skills™* (TILLS™). For more information, contact Brookes Publishing Co., 1-800-638-3775; www.brookespublishing.com/tills.

Library of Congress Cataloging-in-Publication Data

The Library of Congress has cataloged the print edition of the *TILLS Technical Manual* as follows:

Nelson, Nickola

Test of integrated language and literacy skills (TILLS) technical manual / Nickola Nelson [and five others].

pages cm

Summary: "Information about the development and standardization of the Test of Integrated Language and Literacy Skills (TILLS)"—Provided by publisher.

Includes bibliographical references and index.

ISBN 978-1-59857-909-3 (paperback)

1. Communicative disorders—Diagnosis. 2. Language disorders—Diagnosis. I. Title.

RC423.N355 2016

616.85'5075—dc23

2015020952

British Library Cataloguing in Publication data are available from the British Library.

2019 2018 2017 2016 2015

10 9 8 7 6 5 4 3 2 1

Contents

About the Authors	vii
1 Normative Sample, Test Development, and Standardization of the TILLS	1
Criteria for Inclusion in the Normative Sample	1
Composition of the Normative Sample	2
Recruitment and Training of Test Administrators During Standardization	3
Test Development	4
Field Test 1	4
Field Test 2	5
Field Test 3	6
Frequently Asked Questions	6
2 Validity of the TILLS	9
Validity for Purpose 1: Identifying Language and Literacy Disorders	9
Identification of Language and Literacy Disorders at Different Ages	10
Results for Subgroups of Students with Language Disorders	12
Sensitivity to Students at Risk for but Not Identified as Having Language and Literacy Disorders	13
Validity for Purpose 2: Profiling Strengths and Weaknesses	15
Students with Language and Literacy Disorders versus Typically Developing Peers	15
Patterns of Strengths and Weaknesses for Different Diagnostic Groups	15
Validity for Purpose 3: Tracking Change Over Time	18
Determining True Gains or Losses	19
Other Aspects of Construct Validity	20
Evidence that TILLS Scores Reflect Age-Related Change	21
Evidence that TILLS Measures Skills at the Sound/Word and Sentence/Discourse Levels	21
Concurrent Validity	25
Frequently Asked Questions	25
3 Reliability of the TILLS	27
Intraclass Reliability	27
Test–Retest Reliability	27
Interrater Reliability	27
Measuring Students’ Abilities Reliably	28
Frequently Asked Questions	31
4 Evaluation and Reduction of Test Bias	33
Statistical Evaluation of Item Bias	33
Assessment of Item-Level Bias	33
Assessment of Subtest-Level Bias	35
Implications of Score Differences for Interpreting TILLS Results	36
Frequently Asked Questions	38
References	39
Appendix Contributors to TILLS Development and Standardization	41

About the Authors

Nickola Wolf Nelson, Ph.D., CCC-SLP, BCS-CL, is a professor in the Department of Speech Pathology and Audiology and the director of the Ph.D. program in Interdisciplinary Health Sciences at Western Michigan University. She also is the editor of *Topics in Language Disorders* and author of *Language and Literacy Disorders: Infancy through Adolescence* (Pearson, 2010). Dr. Nelson's research has focused on providing curriculum-based language and literacy assessment and intervention and working with families, teachers, and other professionals in collaborative service-delivery models, including a writing lab approach, in school settings. Two of her journal articles have won editors' awards. She is coauthor of the Pediatric Test of Brain Injury™ (PTBI™; with coauthors G. Hotz, N. Helm-Estabrooks, & E. Plante; Paul H. Brookes Publishing Co., 2010), which is conceptually related to the Test of Integrated Language and Literacy Skills™ (TILLS™).

Elena Plante, Ph.D., CCC-SLP, is a professor in the Department of Speech, Language, and Hearing Sciences at The University of Arizona. Her 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. For more than 20 years, she has published journal articles concerning assessment methods, focusing on how norm-referenced tests can be used in valid ways to address clinical concerns. She is coauthor of the Pediatric Test of Brain Injury™ (PTBI™; with coauthors G. Hotz, N. Helm-Estabrooks, & N.W. Nelson; Paul H. Brookes Publishing Co., 2010), which is conceptually related to the Test of Integrated Language and Literacy Skills™ (TILLS™); a widely used textbook on communication disorders; and more than 100 peer-reviewed journal articles, three of which have won editors' awards.

Nancy Helm-Estabrooks, Sc.D., CCC-SLP, BC-ANCDS, is the Brewer Smith Distinguished Professor Emerita in the Department of Communication Sciences and Disorders at Western Carolina University. Her past positions include Professor of Neurology (Speech Pathology) at Boston University School of Medicine and Research Professor at both The University of Arizona and the University of North Carolina. Dr. Helm-Estabrooks's research has focused on the assessment and treatment of neurologic communication disorders and the interface between language and other domains of cognition. Among her numerous publications are eight books, including the *Manual of Aphasia and Aphasia Therapy, Third Edition* (with coauthors M.L. Albert & M. Nicholas; PRO-ED, 2014); eight treatment programs, including the Problem Solving Therapy Program and Melodic Intonation Therapy; and six standardized tests, including the Cognitive Linguistic Quick Test (Pearson, 2001), the Aphasia Diagnostic Profiles (PRO-ED, 1992), the Brief Test of Head Injury (with coauthor G. Hotz; PRO-ED, 1991), and the Pediatric Test of Brain Injury™ (PTBI™; with coauthors G. Hotz, N.W. Nelson, & E. Plante; Paul H. Brookes Publishing Co., 2010), which is conceptually related to the Test of Integrated Language and Literacy Skills™ (TILLS™).

Gillian Hotz, Ph.D., CCC-SLP, is a research professor and the director of the KiDZ Neuroscience Center in the Department of Neurosurgery and of the Miami Project to Cure Paralysis at the University of Miami Miller School of Medicine. Her research has focused on the clinical management and injury prevention of children with brain and spinal cord injuries. Her work in brain injury prevention includes evidence-based programs such as WalkSafe, BikeSafe, SkateSafe, and the Countywide Concussion Care program. Dr. Hotz has also coauthored two neurocognitive tests: the Brief Test of Head Injury (with coauthor N. Helm-Estabrooks; PRO-ED, 1991) for adults and the Pediatric Test of Brain Injury™ (PTBI™; with coauthors N. Helm-Estabrooks, N.W. Nelson, & E. Plante; Paul H. Brookes Publishing Co., 2010) for children and adolescents, which is conceptually related to the Test of Integrated Language and Literacy Skills™ (TILLS™).

ABOUT THE CONTRIBUTORS

E. Brooks Applegate, Ph.D., is a professor in the Department of Educational Leadership, Research, and Technology and a program coordinator 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 and teaches graduate courses in psychometrics, structural equation modeling, and research methodology. He actively participates in funded research and evaluation projects nationally and internationally and has authored and coauthored over 100 peer-reviewed journal articles and over 85 peer-reviewed presentations. Dr. Applegate's experience includes designing and programming relational databases for cross-sectional and longitudinal studies with thousands of variables and millions of records (cases). He was the statistician and test design consultant for the standardization study of the Test of Integrated Language and Literacy Skills™ (TILLS™).

Michele A. Anderson, Ph.D., CCC-SLP, serves as a research professor and grant coordinator at Western Michigan University. She is a speech-language pathologist whose research interests include understanding the role of cognitive functions such as verbal working memory in language assessment and intervention for school-age children. Dr. Anderson coordinated the standardization research of the Test of Integrated Language and Literacy Skills™ (TILLS™). She also worked for 18 years in adult neurorehabilitation in outpatient and home-based settings.



CHAPTER 2

Validity of the TILLS

Evidence of a test's validity relies most critically on the answer to the question "Is the test valid for my specific purpose?" As described in Chapter 1 of the *TILLS Examiner's Manual*, the TILLS is designed to serve three key purposes:

1. To identify language and literacy disorders
2. To document patterns of relative strengths and weaknesses
3. To track changes in language and literacy skills over time

In this chapter, we present evidence that supports the validity of TILLS for its three purposes. We then provide additional evidence of broadly based construct validity. Construct validity is established through evidence that a test measures the abilities (or "constructs") that it is designed to measure. For the TILLS, the key constructs are the oral and written language abilities (at both sound/word and sentence/discourse levels) that school-age students need for academic and social purposes.

The most important forms of test validity are those most directly related to the examiner's purpose for giving the test.

VALIDITY FOR PURPOSE 1: IDENTIFYING LANGUAGE AND LITERACY DISORDERS

The primary purpose of the TILLS is to identify students with language disorders in the oral or written domain. Various diagnostic terms are used to describe these impairments. Oral impairments are commonly called primary language disorders, specific language impairments, and speech-language impairments. Written language impairments are called learning disabilities, dyslexia, and specific reading disabilities. Some professionals use the broader term *language learning disability* to encompass primary disorders of oral and written language. This also reflects the fact that oral and written language disorders rarely occur in isolation of one another (e.g., Kamhi & Catts, 2011).

To provide empirical evidence to support use of the TILLS for identifying primary disorders of language and literacy, we tested students who had previously been diagnosed with language disorders in the oral or written domain (see Table 2.1).

Table 2.1. Selection criteria for students participating in validity studies

Language disorders: Spoken or spoken plus written modalities (N = 191)	Language disorders: Written modality only (N = 88)
Ages 6–18 years Eligibility for special education or private services on the basis of having language impairment or language-based learning disorder that included spoken language only or spoken and written language impairments May have or have had a concomitant speech impairment May have comorbid attention-deficit/hyperactivity disorder (ADHD) May have comorbid impairments in the written modality	Ages 6–18 years Eligibility for special education or private services on the basis of having a learning disability on the basis of reading or writing only No speech impairment or oral language diagnosis now or in the past May have comorbid ADHD

Identification of Language and Literacy Disorders at Different Ages

Students with language disorders may come to the attention of parents, teachers, or other professionals because of difficulty in conversational language or due to their inability to meet the demands of academic language, including speaking, listening, reading, or writing in curricular contexts. The TILLS is designed to detect language impairments, regardless of whether they occur in the oral or written modality.

Language profiles can change over time due to maturation, age expectations for oral and written language performance, language intervention, or special education. As a student's behavioral profile changes with age, the particular subtests most relevant to diagnosis can be expected to change as well. Therefore, we identified the subset of TILLS subtests that most effectively identifies the presence of a language disorder, regardless of modality (oral or written), at different age ranges. The subtests that contribute to the Identification Core score at different ages are listed in Table 2.2. If an examiner's purpose in giving the TILLS is to identify a language disorder, then all of the subtests listed in Table 2.2 for the student's age must be administered.

Clinicians need four pieces of information to use the TILLS Identification Core to identify language and literacy disorders in an evidence-based manner:

1. List of subtests that are *most discriminative* for students at each age (i.e., the Identification Core)
2. Evidence supporting the composite's *sensitivity* for accurately identifying disorders among students known to have them
3. Evidence supporting the composite's *specificity*, indicating that it does not over identify students who do not have a disorder as having one
4. Specific *cut scores* that make accurate identification possible

The cut score is the score below which a diagnosis of disorder is made. Table 2.2 reports the diagnostic accuracy that can be achieved with use of the subtests of the Identification Core. The sensitivity and

Sensitivity is the percentage of students with impairments identified as having impairments.

Specificity is the percentage of typically developing students correctly identified as having normal language.

Cut score is the score that maximizes correct identification of students with and without impairments. The proper cut score should always be used when identifying students with language and literacy disorders.

Table 2.2. TILLS subtests that support diagnosis of language and literacy disorders at different ages

Age range (years)	Identification Core ^a	Sensitivity	Specificity	Cut score ^b
6;0–7;11	1. Vocabulary Awareness (VA) 2. Phonemic Awareness (PA) 4. Nonword Repetition (NWRRep)	84	84	24
8;0–11;11	1. Vocabulary Awareness (VA) 5. Nonword Spelling (NWSpell) 10. Nonword Reading (NWRRead) 12. Written Expression–Discourse Score (WE-Disc)	88	85	34
12;0–18;11	2. Phonemic Awareness (PA) 5. Nonword Spelling (NWSpell) 7. Reading Comprehension (RC) ^c 11. Reading Fluency (RF) 12. Written Expression–Word Score (WE-Word)	86	90	42

^aThe Identification Core score is the sum of the standard scores on the subtests listed for each age group.

^bStudents scoring at or above this score on the TILLS were classified as not having impairments; students scoring below this score were classified as having impairments.

^cNote that the Listening Comprehension (LC) subtest must be administered prior to the Reading Comprehension (RC) subtest, but the LC score is not used in the Identification Core.

Technical notes: We obtained scores from 279 students previously diagnosed with a language or literacy disorder during the standardization phase of TILLS development. Of these, 51 were between the ages of 6 and 7 years, 160 were between the ages of 8 and 11 years, and 68 were between 12 and 18 years of age. All children met the selection criteria provided in Table 2.1 for language disorders in the spoken and written modalities or in the written modality only. These students had received speech-language or other special education services at some point during their time in school. We compared test scores from these children with data from 1,255 students collected during norming of the TILLS (352 between the ages of 6 and 7 years, 461 between the ages of 8 and 11 years, 442 between the ages of 12 and 18 years of age) who represented the same age bands.

We determined the optimal set of subtests for each age band through a multistage procedure. First, we graphically displayed the performance for children in the normative sample and students previously diagnosed with language and literacy disorders on all TILLS subtests to determine age groups in which the subtest standard scores for children with language and literacy disorders were most similar. This step yielded the three age clusters listed in Table 2.2. Second, we identified those subtests showing a mean difference of at least 1 standard deviation between students with and without language/literacy disorders at each of the three age bands. We conducted initial exploratory discriminant analyses using those subtests to determine their accuracy in differentiating the students with and without language and literacy disorders. We conducted additional analyses to determine whether group differentiation could be improved by combining subtest scores. This procedure started with scores from the best performing subtests and systematically added subtest scores until no additional improvement in group differentiation could be obtained. We then repeated the analyses at each normative age group within the age bands displayed in Table 2.2. Finally, we identified a common cut score that maximized sensitivity and specificity at each age. These cut scores are displayed in Tables 2.2 and 2.3.

specificity levels exceed 80% at all ages, which is considered acceptable for clinical use (Plante & Vance, 1994).

To ensure that the accuracy achieved for the three age bands shown in Table 2.2 applies to children at different ages within these age bands, we also provide age-related sensitivity and specificity information in Table 2.3. Note that the age range 14–18 years is treated as a single group because the normative data for those ages were combined into a single group (see Chapter 1).

Likelihood ratios are an alternate means of expressing diagnostic accuracy. *Positive likelihood ratios* express the odds that a student scoring below the cut score has a disorder (this number should be greater than 1.0; larger numbers are better). *Negative likelihood ratios* express the odds that a student who scores above the cut score actually has a disorder. (This number should be less than 1.0; smaller numbers are better.) These ratios are mathematically derived from sensitivity and specificity data and therefore provide another method of expressing the probability that the TILLS will accurately diagnose a student who has a language or literacy disorder. The positive and negative likelihood ratios

TILLS positive likelihood ratios express high odds that a student who scores below the cut score has a language or literacy disorder.

TILLS negative likelihood ratios express low odds that a student who scores above the cut score has a language or literacy disorder.

Table 2.3. TILLS sensitivity, specificity, and the empirically derived cut score for maximizing identification accuracy

Age	Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio	Cut score ¹
6	84%	82%	4.74	.19	24
7	84%	86%	6.04	.18	24
8	97%	100%	9.70	.03	34
9	83%	81%	4.50	.20	34
10	81%	81%	4.34	.24	34
11	86%	82%	4.86	.17	34
12	83%	100%	8.33	.17	42
13	84%	86%	5.95	.18	42
14-18	87%	87%	6.77	.15	42

¹Students scoring at or above this score on the TILLS Identification Core were classified as not having impairments; students scoring below this score were classified as having impairments.

for the TILLS are reported in Table 2.3. As with sensitivity and specificity data, these positive and negative likelihood ratios can only be achieved with use of the empirically derived age-specific cut score reported in Table 2.3.

The positive likelihood ratios obtained indicate that a student with an Identification Core score below the age-related cut score is 4–6 times more likely to have a language disorder than not. Conversely, a student with a TILLS Identification Core score above the cut score is unlikely to have a language disorder (i.e., the likelihood of a disorder is only 0.20, indicating there is a much greater chance that skills are normal than disordered).

Results for Subgroups of Students with Language Disorders

We included in our analysis students with a previous diagnosis associated with primary impairments of language or literacy (i.e., specific impairments that affect language and literacy in the presence of normal cognitive and sensory function). Information provided by parents on a student information form, and verified by test administrators when possible, was used to classify students into two subgroups. Students whose information implicated spoken language only or spoken plus written language were treated as one group ($N = 200$), and students whose information implicated written language only were treated as a separate group ($N = 91$).

One important question is whether the children in these two groups are fundamentally different, perhaps warranting different Identification Core subtests for diagnostic purposes. Table 2.4 presents the identification rates for the students included in Table 2.2, separated according to the criteria described in Table 2.1. The means for each subgroup are displayed in the graph in Figure 2.1. As this figure shows, the scores for each group are fairly similar, although the subgroup of students who

Table 2.4. Identification rate for subgroups of children with language and literacy disorders

Previous diagnosis	Identification rate
Language disorders: Spoken or spoken plus written modalities	88%
Language disorders: Written modality only	83%

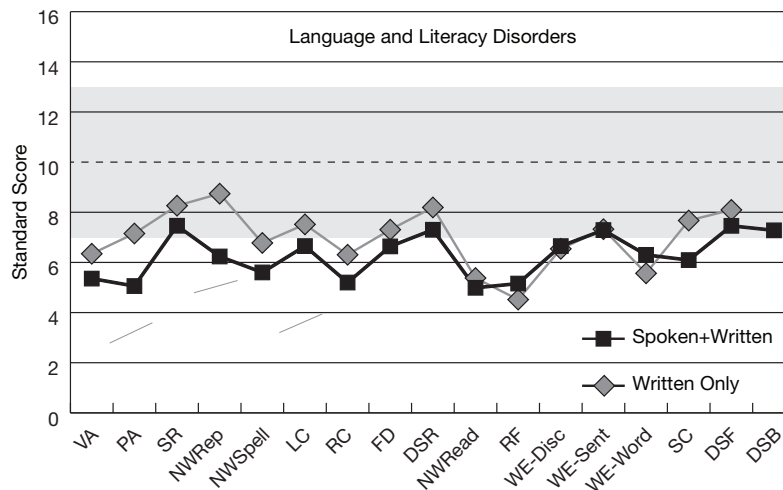


Figure 2.1. Mean scores associated with students previously diagnosed with language and literacy disorders in both the spoken and written modalities and the written modality only. The dashed line indicates the normative group mean, and the gray area indicates ± 1 standard deviation from the normative group mean. (Key: Vocabulary Awareness [VA], Phonemic Awareness [PA], Story Retelling [SR], Nonword Repetition [NWRep], Nonword Spelling [NWSpell], Listening Comprehension [LC], Reading Comprehension [RC], Following Directions [FD], Delayed Story Retelling [DSR], Nonword Reading [NWRRead], Reading Fluency [RF], Written Expression–Discourse Score [WE-Disc], Written Expression–Sentence Score [WE-Sent], Written Expression–Word Score [WE-Word], Social Communication [SC], Digit Span Forward [DSF], Digit Span Backward [DSB].)

Technical notes: We tested a total of 291 students with language-based impairments. We grouped students according to whether their impairments included oral and written language skills or written language skills only, according to the criteria in Table 2.1. We tested 200 students with written and oral language goals and 91 with written language goals only. We evaluated test scores using a mixed ANOVA with subgroup and TILLS subtests as factors. The results indicated a significant effect for subgroup $F(1,272) = 4.75, p = .03$, indicating that the subtest scores for the students with both oral and written impairments were significantly worse than those of the students with only written language impairments. There was also a significant effect for subtest $F(17,4624) = 18.85, p < .0001$, indicating that some subtests were more difficult for both subgroups than others. In addition, there was a significant subtest by subgroup effect $F(17,4624) = 4.58, p < .0001$. Tukey HSD post hoc testing indicated that scores on the Nonword Repetition (NWRep) subtest differed significantly between groups.

were previously diagnosed with both spoken and written impairments had slightly lower mean scores overall, and scored significantly lower on the Nonword Repetition (NWRep) subtest in particular, than students with written impairments only. We calculated the age-appropriate Identification Core score for each student. We then used the age-appropriate cut score (see Table 2.3) to produce the classification rates for each subgroup. Figure 2.1 indicates that the TILLS Identification Core score correctly identified both groups of students at rates that indicate clinical utility for the test (Plante & Vance, 1994). Identification rates for students previously diagnosed with both oral and written language impairments were slightly better than those for students with written language impairments only. This is consistent with the slightly lower overall scores seen in Figure 2.1 for the students with both oral and written language difficulties (lower scores tend to yield better identification rates).

Sensitivity to Students at Risk for but Not Identified as Having Language and Literacy Disorders

Parents or teachers may express concerns about the academic or communication skills of a student who does not have a formal diagnosis. Such students can be considered at risk for a language learning disorder because of these concerns. Criteria for inclusion in the language/literacy risk group during TILLS standardization research are outlined in Table 2.5. We used the scores from this group to assess the extent to which the TILLS is sensitive to the problems of students considered at risk for language and literacy disorders but who have not been diagnosed as having a disorder.

Table 2.5. Selection criteria for students considered at risk for, but not formally diagnosed with, a language or literacy disorder ($N = 198$)

Ages 6–18
No diagnosis of language, reading, and/or language-based learning disorder
Parent answered yes to one or more questionnaire items indicating
• Parent or teacher had concerns about language, reading, and/or language-based learning difficulties • Child's language or literacy skills had been formally tested • Child received support services for language or literacy concerns • Child repeated a grade in school
May or may not have attention-deficit/hyperactivity disorder (ADHD)

As a group, students identified as being at risk performed significantly below the average of the normative group on all subtests of the TILLS. This is displayed graphically in Figure 2.2. However, the range for the risk group also shows considerable variability in the scores, indicating that the scores of some students were clearly comparable to those of the children in the normative group. It is likely that some students who are considered at risk have weak skills that reflect normal population variation but do not merit a formal diagnosis of a language or literacy disorder. Other students may represent cases of undiagnosed disorder and should be considered for special education services or other forms of treatment. We addressed this issue by asking what percentage of children identified as at risk would be diagnosed by the TILLS as having a language or literacy disorder (see Table 2.6). We compared these children's scores to their age-appropriate cut score (reported in Table 2.3) to determine whether the TILLS would identify a language or literacy disorder or not. For purposes of comparison, we also calculated the rates for children in the normative sample. The data in Table 2.6 suggest that there may be value in using the TILLS with students who have been flagged as struggling with aspects of academic language because a sizeable proportion may actually have a language and literacy disorder that requires specialized attention.

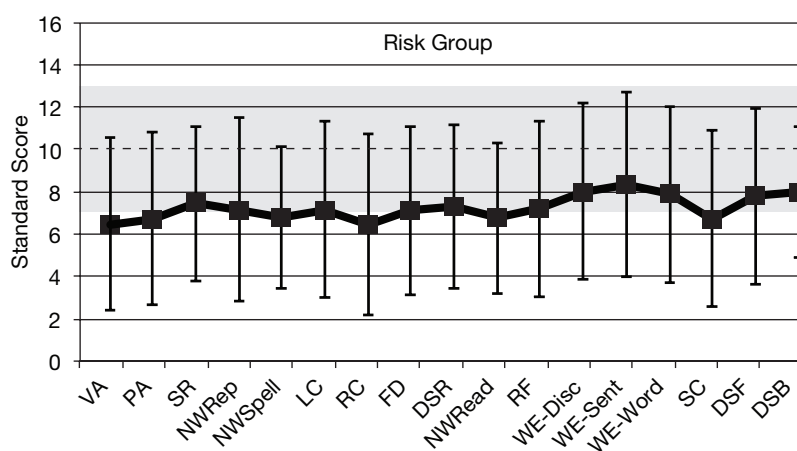


Figure 2.2. Performance by 198 students designated at risk for language or literacy disorders based on parent or teacher report. The dashed line indicates the normative group mean, and the gray area indicates ± 1 standard deviation from the normative group mean. The box and error bars indicate the mean and standard deviation for the performance of the risk group. (Key: Vocabulary Awareness [VA], Phonemic Awareness [PA], Story Retelling [SR], Nonword Repetition [NWRRep], Nonword Spelling [NWSpell], Listening Comprehension [LC], Reading Comprehension [RC], Following Directions [FD], Delayed Story Retelling [DSR], Nonword Reading [NWRRead], Reading Fluency [RF], Written Expression–Discourse Score [WE-Disc], Written Expression–Sentence Score [WE-Sent], Written Expression–Word Score [WE-Word], Social Communication [SC], Digit Span Forward [DSF], Digit Span Backward [DSB].)

Technical notes: A total of 198 students met the “at-risk” criteria provided in Table 2.5. We compared TILLS test scores from these students with the scores of 1,267 typically developing students who were tested during TILLS standardization. We evaluated test scores using a mixed ANOVA with subgroup and TILLS subtests as factors. The results indicated a significant effect for subgroup $F(1,1411) = 238.75, p < .0001$, indicating that the students at risk scored significantly lower on TILLS subtests overall than typically developing students. There was also a significant effect for subtest $F(17,23987) = 6.14, p < .0001$, indicating that some subtests were more difficult for students than others. In addition, there was a significant group by subtest effect $F(17,23987) = 6.01, p < .0001$. Tukey HSD post hoc testing indicated that all of the individual subtest scores were significantly lower for the risk group than the typically developing group.

Table 2.6. Identification rates for children judged at risk for a language or literacy disorder compared with typically developing children from the normative sample

	Identification by the TILLS	
	Language disorder is present	Language disorder is not present
Students at risk ($N = 198$)	59%	41%
Students with no known risk ($N = 1,267$)	13%	87%

Technical notes: We obtained scores from 198 students identified by a parent or teacher as at risk for a language or literacy disorder during the standardization phase of TILLS development. All at-risk students met the selection criteria provided in Table 2.5. These students had not received speech-language or other special education services. “Students with no known risk” included 1,267 students tested during the norming phase of the TILLS. We used the Identification Core score (reported in Table 2.2) to determine the percentage of students who would be identified as having a language or literacy disorder by the TILLS. We based this designation on whether the students’ age-appropriate Identification Core score fell below the age-appropriate cut score (indicating that a language disorder was present) or at or above the age-appropriate cut score (indicating that a language disorder is not present). The age-appropriate cut scores are reported in Table 2.3.

VALIDITY FOR PURPOSE 2: PROFILING STRENGTHS AND WEAKNESSES

The second purpose of the TILLS is to provide information about patterns of relative strengths and weaknesses. Support for this diagnostic purpose rests on two pieces of evidence. First, students who have diagnoses that implicate language and literacy skills should score differently on the TILLS compared to their typically developing peers in ways that are consistent with each diagnosis. Second, the subtest scores of these students should produce profiles of strengths and weaknesses that differ from each other.

Students with Language and Literacy Disorders versus Typically Developing Peers

Students diagnosed with disorders known to affect language and literacy should perform differently on the TILLS than their typically developing peers. Furthermore, their profiles of relative strengths and weaknesses on the TILLS should vary with their diagnoses. In Tables 2.7–2.10, we present standard score data from students with four diagnoses commonly represented on special education caseloads—primary language literacy disorder, autism spectrum disorder, deaf or hard of hearing, and intellectual disability. These tables provide evidence that the TILLS is sensitive to the significantly lower language and literacy abilities of all four diagnostic groups when compared with peers with typical language development.

Patterns of Strengths and Weaknesses for Different Diagnostic Groups

Not only should students with disabilities differ from their typically developing peers, but they should also show patterns of strengths and weaknesses that are consistent with their diagnoses. It is clear from the scoring profiles for students with different disabilities in Figure 2.3 that each diagnostic group has its own pattern of relative strengths and weaknesses.

Students previously diagnosed with language and literacy disorders show weaknesses in multiple areas of both spoken and written language (see Figure 2.3a). In particular, Vocabulary Awareness (VA), Phonemic Awareness (PA), Nonword Spelling (NWSPELL), Nonword Reading (NWRD), Reading Comprehension (RC), Reading Fluency (RF), Written Expression (WE), and Social Communication (SC) are all below -1 standard deviation on average.

In contrast to the pattern for students with language and literacy disorders, the pattern for students with autism spectrum disorder (see Figure 2.3b) indicates particularly low skills in the area of social communication, a hallmark impairment for this disorder. These students also have particular difficulty with the Vocabulary

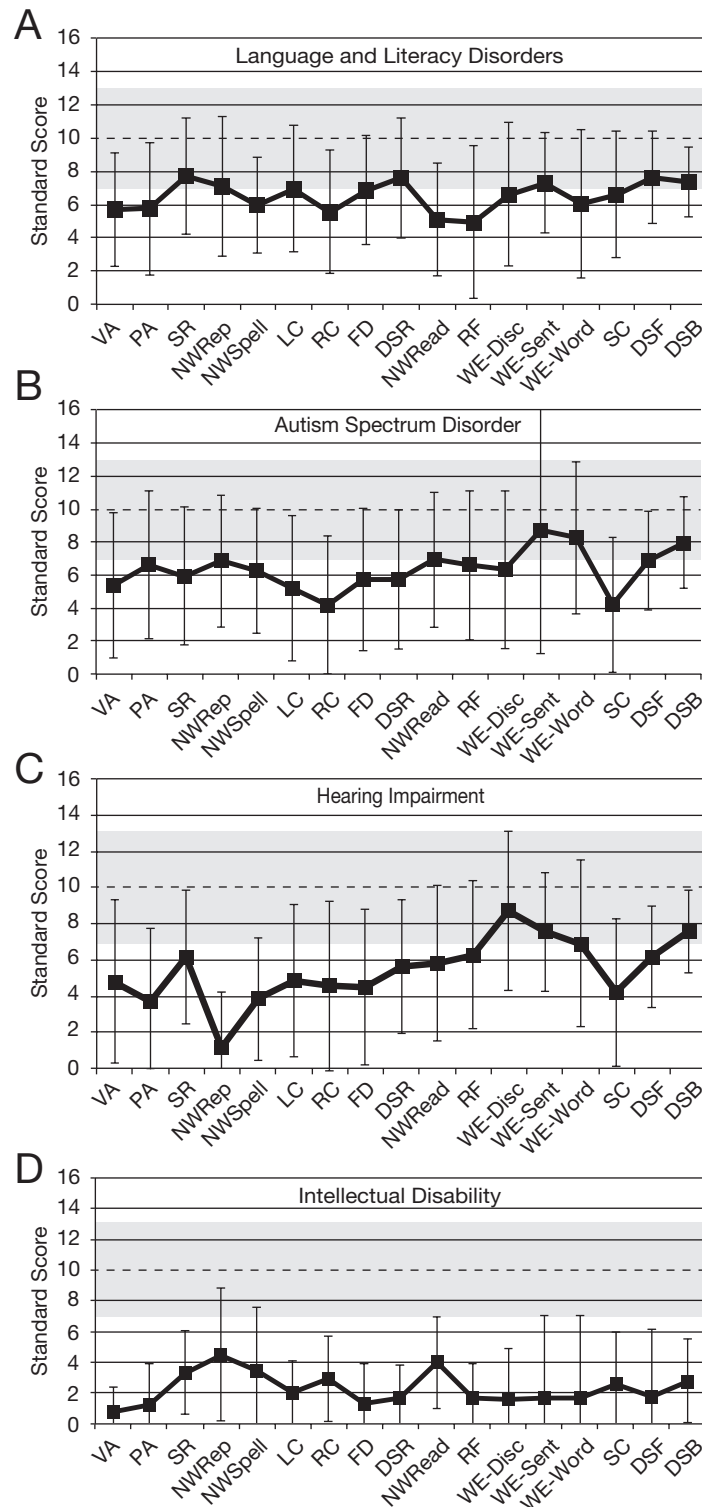


Figure 2.3. Score profiles for students with a) language and literacy disorders, b) autism spectrum disorder, c) hearing impairment, and d) intellectual disability. The dashed line indicates the normative group mean, and the gray area indicates ± 1 standard deviation from the normative group mean. The box and error bars indicate the mean and standard deviation for the performance of the diagnostic group. (Key: Vocabulary Awareness [VA], Phonemic Awareness [PA], Story Retelling [SR], Nonword Repetition [NWRRep], Nonword Spelling [NWSpell], Listening Comprehension [LC], Reading Comprehension [RC], Following Directions [FD], Delayed Story Retelling [DSR], Nonword Reading [NWRRead], Reading Fluency [RF], Written Expression–Discourse Score [WE-Disc], Written Expression–Sentence Score [WE-Sent], Written Expression–Word Score [WE-Word], Social Communication [SC], Digit Span Forward [DSF], Digit Span Backward [DSB].)

Table 2.7. TILLS performance of students previously diagnosed with language disorder

	Students with language and literacy disorder (<i>N</i> = 248)		Typically developing students (<i>N</i> = 248)	
	Mean	<i>SD</i>	Mean	<i>SD</i>
1. Vocabulary Awareness (VA)	5.62*	3.41	9.51	3.26
2. Phonemic Awareness (PA)	5.76*	4.04	9.56	3.24
3. Story Retelling (SR)	7.71*	3.50	9.75	3.12
4. Nonword Repetition (NWRep)	7.05*	4.22	10.09	2.96
5. Nonword Spelling (NWSpell)	5.99*	2.98	9.55	3.40
6. Listening Comprehension (LC)	6.94*	3.83	9.55	3.29
7. Reading Comprehension (RC)	5.54*	3.70	9.45	3.09
8. Following Directions (FD)	6.90*	3.29	9.70	3.09
9. Delayed Story Retelling (DSR)	7.57*	3.65	9.71	3.22
10. Nonword Reading (NWRead)	5.15*	3.46	9.67	3.17
11. Reading Fluency (RF)	4.94*	4.58	9.60	2.85
12. Written Expression–Discourse Score (WE-Disc)	6.65*	4.31	9.53	2.86
12. Written Expression–Sentence Score (WE-Sent)	7.31*	3.02	9.53	3.15
12. Written Expression–Word Score (WE-Word)	6.09*	4.50	9.89	3.01
13. Social Communication (SC)	6.23*	3.89	9.39	3.21
14. Digit Span Forward (DSF)	7.69*	2.83	9.34	3.30
15. Digit Span Backward (DSB)	7.43*	2.17	9.75	3.05

*Statistically significant group differences at $p < .01$.

Subtest scores reported are subtest standard scores (mean = 10, *SD* = 3).

Technical notes: Data were from 496 children, including 248 children and adolescents (mean age = 10;7, age range = 6;0–18;9) previously diagnosed with a language disorder and 248 children and adolescents selected as matched controls. The selection criteria for students with previously diagnosed language disorders (spoken, written, or both) are provided in Table 2.1. We compared these students' TILLS subtest scores with TILLS subtest scores for 248 students (mean age = 10;7, range = 6;0–18;10) from the normative sample who were selected to match the students with language disorder for age, sex (106 girls, 142 boys in each group), racial background (146 white, 102 minority in each group), maternal education (median education level of 1–3 years of college, range of less than a high school diploma to 4 or more years of college). We tested group differences with a mixed ANOVA, which indicated a significant group effect $F(1,462) = 319.77$, $p < .001$ and a significant group by subtest interaction $F(16,7392) = 10.17$, $p < .001$. We documented significant group differences for each subtest ($p < .01$) using a Tukey HSD post hoc analysis.

Awareness (VA) subtest, which requires them to associate words in multiple ways. This finding is consistent with these students' lack of mental flexibility. Likewise, Listening Comprehension (LC) and Reading Comprehension (RC), both of which require inferential ability as well as factual recall, are particularly impaired.

Students who are deaf or hard of hearing (see Figure 2.3c) show particular problems with subtests requiring sound-level skills, including Phonemic Awareness (PA), Nonword Repetition (NWRep), and Nonword Spelling (NWSpell). These students also show impairments in most subtests that involve listening or reading, which is consistent with the literature associating hearing loss with general language and literacy difficulties. Unlike the students who have language and literacy disorders in the absence of a hearing loss, however, these students are relatively strong in the Written Expression (WE) subtest. The availability of visual stimuli during this subtest likely assisted their performance.

Finally, as may be expected, students with intellectual disability showed the lowest average scores of any of the groups tested (see Figure 2.3d). The group mean for

Table 2.8. TILLS performance of students diagnosed with autism spectrum disorder

	Students with autism spectrum disorder (<i>N</i> = 79)		Typically developing students (<i>N</i> = 79)	
	Mean	<i>SD</i>	Mean	<i>SD</i>
1. Vocabulary Awareness (VA)	5.42*	4.44	10.03	2.82
2. Phonemic Awareness (PA)	6.56*	4.52	9.93	2.91
3. Story Retelling (SR)	5.87*	4.12	9.35	3.29
4. Nonword Repetition (NWRRep)	6.81*	3.97	9.81	2.97
5. Nonword Spelling (NWSpell)	6.23*	3.90	9.59	3.24
6. Listening Comprehension (LC)	5.23*	4.45	9.99	3.22
7. Reading Comprehension (RC)	4.11*	4.27	9.92	2.89
8. Following Directions (FD)	5.63*	4.28	10.10	2.74
9. Delayed Story Retelling (DSR)	5.58*	4.17	9.19	3.51
10. Nonword Reading (NWRead)	6.95*	4.21	9.77	3.03
11. Reading Fluency (RF)	6.56*	4.61	9.37	3.26
12. Written Expression–Discourse Score (WE-Disc)	6.17*	4.77	9.70	3.15
12. Written Expression–Sentence Score (WE-Sent)	8.78	7.62	9.76	2.55
12. Written Expression–Word Score (WE-Word)	8.19	4.72	9.72	3.22
13. Social Communication (SC)	4.08*	4.10	10.01	3.20
14. Digit Span Forward (DSF)	6.81*	3.03	9.68	3.03
15. Digit Span Backward (DSB)	7.94	2.76	9.79	2.63

*Statistically significant group differences at $p < .01$.

Subtest scores reported are subtest standard scores (mean = 1, *SD* = 3).

Technical notes: Data are from 79 students (mean age = 12;3, range = 6;2–18;8) diagnosed with autism spectrum disorder (ASD). We compared these students with 79 students from the normative sample (mean age = 12;3, range = 6;2–18;8). We selected these typically developing students to match the students with autism spectrum disorder for age, sex (12 girls, 67 boys in each group), racial background (62 white, 17 minority in the ASD group; 63 white, 16 minority in the typically developing group), maternal education (median 1–3 years of college, range of less than a high school diploma to 4 or more years of college for each group). We tested group differences with a mixed ANOVA, which indicated a significant group effect $F(1,145) = 84.37$, $p < .001$ and a significant group by subtest interaction $F(16,2320) = 6.00$, $p < .001$. We documented significant group differences for 14 of the subtest scores ($p < .01$) using a Tukey HSD post hoc analysis.

each subtest was at or below 2 standard deviations in all cases. The skills tested by the TILLS were very difficult for these students; however, as the error bars show, there was a noteworthy range of performance across subtests. Very low scores cannot be assumed for all students with intellectual disabilities, and the TILLS can detect skills that are relative strengths for some of these students.

VALIDITY FOR PURPOSE 3: TRACKING CHANGE OVER TIME

The third major purpose of the TILLS is to track changes over time. Tracking change involves determining when two standard scores obtained at different points in time indicate true change. This determination rests on the amount of variance in the individual subtest scores over time. The True Change Interval for each TILLS subtest indicates the number of standard score points that could reflect random variability rather than true change in test scores over time. This measure is related to test–retest reliability (see Chapter 3). When the True Change Interval for a subtest is small, it is

Table 2.9. TILLS performance of students with hearing impairment

	Deaf or hard-of-hearing students (<i>N</i> = 40)		Typically developing students (<i>N</i> = 40)	
	Mean	<i>SD</i>	Mean	<i>SD</i>
1. Vocabulary Awareness (VA)	4.70*	4.39	9.71	2.86
2. Phonemic Awareness (PA)	3.48*	4.24	9.43	3.58
3. Story Retelling (SR)	6.08*	3.76	10.18	3.30
4. Nonword Repetition (NWRRep)	1.23*	3.25	10.64	2.65
5. Nonword Spelling (NWSpell)	3.65*	3.25	9.72	3.17
6. Listening Comprehension (LC)	4.73*	4.25	10.08	2.36
7. Reading Comprehension (RC)	4.30*	4.68	9.97	2.69
8. Following Directions (FD)	4.30*	4.23	9.85	2.38
9. Delayed Story Retelling (DSR)	5.50*	3.67	10.13	2.81
10. Nonword Reading (NWRRead)	5.70*	4.39	9.44	3.02
11. Reading Fluency (RF)	6.24*	4.09	10.08	2.12
12. Written Expression–Discourse Score (WE-Disc)	8.68	4.55	10.58	2.10
12. Written Expression–Sentence Score (WE-Sent)	7.55	3.31	10.26	2.97
12. Written Expression–Word Score (WE-Word)	6.90*	4.67	10.63	2.57
13. Social Communication (SC)	4.23*	4.28	9.41	2.99
14. Digit Span Forward (DSF)	6.05*	2.82	9.28	2.71
15. Digit Span Backward (DSB)	7.78	2.35	9.82	2.79

*Statistically significant group differences at $p < .01$.

Subtest scores reported are subtest standard scores (mean = 1, *SD* = 3).

Technical notes: Data are from 40 students who were deaf or hard of hearing (mean age = 11;3, range = 6;0–15;10). All students with hearing loss were aided or had cochlear implants. We compared these students with 40 students from the normative sample (mean age = 11;3, range = 6;0–15;10). We selected these students to match the students with hearing loss for age, sex (18 girls, 22 boys in each group), racial background (28 white, 12 minority in each group), and maternal education (median of 1–3 years of college, range of less than a high school diploma to 4 or more years of college). We tested group differences with a mixed ANOVA, which indicated a significant group effect $F(1,73) = 84.67, p < .001$ and a significant group by subtest interaction $F(16,1168) = 7.73, p < .001$. We documented significant group differences for 13 of the subtest scores ($p < .01$) using a Tukey HSD post hoc analysis.

relatively easy to track changes over time. Note, however, as the True Change Interval becomes larger, it becomes increasingly difficult to detect true change.

Determining True Gains or Losses

Table 2.11 presents the True Change Interval for each subtest for each age group. These values correspond to 68% and 90% confidence intervals around a student's score. The 68% confidence band values are also provided on page 39 of the *Examiner Record Form*. These values reflect the standard error of the estimate (SEE) for each subtest. We used the test–retest correlations (reported in Chapter 3) to calculate the SEE so that the True Change Intervals account for the amount of score variation that might be expected simply by testing students more than once. The values in Table 2.11 indicate the minimum amount of change in the standard score that would have to occur from the first to second testing session for the clinician to conclude that true score change had occurred.

Clinicians should keep in mind that subtest and composite scores show developmental change (see Figure 2.4) that is independent of gains accrued via special

Table 2.10. TILLS performance of students diagnosed with intellectual disability

	Students with intellectual disability (N = 14)		Typically developing students (N = 14)	
	Mean	SD	Mean	SD
1. Vocabulary Awareness (VA)	0.78*	1.63	9.5	3.13
2. Phonemic Awareness (PA)	1.21*	2.69	10.14	2.96
3. Story Retelling (SR)	2.86*	2.73	9.14	3.63
4. Nonword Repetition (NWRRep)	3.43*	4.24	11.14	2.32
5. Nonword Spelling (NWSpell)	2.00*	2.08	9.86	2.45
6. Listening Comprehension (LC)	2.93*	2.76	9.64	3.32
7. Reading Comprehension (RC)	1.29*	2.67	9.43	4.11
8. Following Directions (FD)	1.71*	2.05	9.43	3.16
9. Delayed Story Retelling (DSR)	3.43*	3.08	9.57	3.94
10. Nonword Reading (NWRRead)	1.64*	2.31	8.93	2.73
11. Reading Fluency (RF)	1.57*	3.03	10.86	1.46
12. Written Expression–Discourse Score (WE-Disc)	1.71*	4.38	9.86	2.85
12. Written Expression–Sentence Score (WE-Sent)	2.57*	3.32	9.79	3.66
12. Written Expression–Word Score (WE-Word)	2.57*	3.32	9.79	3.66
13. Social Communication (SC)	2.79*	2.67	9.00	3.90
14. Digit Span Forward (DSF)	3.93*	1.82	8.50	3.78
15. Digit Span Backward (DSB)	5.50*	2.77	8.79	2.23

*Statistically significant group differences at $p < .01$.

Subtest scores reported are subtest standard scores (mean = 10, SD = 3).

Technical notes: Data are from 14 students diagnosed with intellectual disability (mean age = 13;2, range = 8;2–17;9). We compared these students with 14 students from the normative sample (mean age = 13;0, range = 8;2–17;8). We selected these students to match the students with intellectual disability for age, sex (6 girls, 8 boys in each group), racial background (9 white, 5 minority in the intellectual disability group; 10 white, 4 minority in the typical group), and maternal education (median of 4 or more years of college, range of less than a high school diploma to 4 or more years of college). Because many of the students with intellectual disability received scores of 0 on many of the subtests, the data were not normally distributed. Therefore, we used the nonparametric Mann-Whitney U test to determine group differences that exceeded $p < .001$. Group differences on all subtests met this criterion.

education or other treatment. The interpretation of change is complicated when a student moves from one normative age group to another from the first to the second testing. For example, a standard score of 10 at age 7 requires fewer correct answers than a standard score of 10 at age 8. Therefore, even when the student answers more questions correctly during the second test session, this difference may be masked by the age-related change between one normative age group and another. This issue is discussed more fully in Chapter 3 of the *TILLS Examiner's Manual*.

OTHER ASPECTS OF CONSTRUCT VALIDITY

Once a clinician is satisfied that there is evidence to support his or her primary purpose for administering the TILLS (for example, to identify a disorder or to develop a profile of strengths and weaknesses), he or she may also wish to consider further evidence of test validity. Evidence supporting that a test measures the skills it claims to measure is known as

Construct validity assesses whether the intended model and specific abilities of the test are supported by empirical evidence.

Table 2.11. Standard error of the estimate (SEE) values for TILLS subtests

	Subtests																
	VA	PA	SR	NWRep	NWSpell	LC	RC	FD	DSR	NWRead	RF	WE-Disc	WE-Sent	WE-Word	SC	DSF	DSR
Confidence Interval																	
68%	1	2	2	2	1	2	2	2	2	1	2	1	1	2	2	2	2
90%	2	3	4	4	2	4	3	3	4	2	3	2	2	3	3	3	3

Key: Vocabulary Awareness (VA), Phonemic Awareness (PA), Story Retelling (SR), Nonword Repetition (NWRep), Nonword Spelling (NWSpell), Listening Comprehension (LC), Reading Comprehension (RC), Following Directions (FD), Delayed Story Retelling (DSR), Nonword Reading (NWRead), Reading Fluency (RF), Written Expression–Discourse Score (WE-Disc), Written Expression–Sentence Score (WE-Sent), Written Expression–Word Score (WE-Word), Social Communication (SC), Digit Span Forward (DSF), Digit Span Backward (DSB).

construct validity. The TILLS is intended to measure curriculum-relevant language and literacy skills at the sound/word and sentence/discourse levels. The test also includes basic memory measures, which may be useful for interpreting performance across subtests (see Chapter 2 of the *Examiner's Manual*). Next, we present several forms of evidence that support construct validity of the TILLS.

Evidence that TILLS Scores Reflect Age-Related Change

Raw scores from a test that measures developmental skills should show age-related change over the age range during which the skill is developing. Figure 2.4 presents changes in raw scores across ages 6–18 years. As this figure shows, all of the TILLS subtests show some degree of age-related change. Ages 6;0–7;11 years were broken into 6-month intervals in the normative data because students at the youngest ages show the most age-related change. One-year normative intervals are justified from 8 to 13 years of age. From 14 to 18 years of age, little differences in raw scores occur. This is not to say that students from 14 to 18 years of age do not differ from each other in other important ways, but performance is no longer age-dependent with respect to the skills measured with the TILLS. For this reason, we collapsed these ages into one category for the normative sample.

Evidence that TILLS Measures Skills at the Sound/Word and Sentence/Discourse Levels

If the TILLS measures multiple skills, then the statistical structure of students' scores on the different subtests within the test should reflect this. Otherwise, a student's score on any one subtest is a fair representation of what he or she is likely to score on the remaining subtests, so there is little reason to administer additional subtests. Furthermore, when a clinician wishes to make broad statements that go beyond individual subtest scores, such as conclusions concerning skills at the sound/word level versus the sentence/discourse level, there should be evidence that such overarching interpretations are valid. We used a factor analysis to assess whether the TILLS can be interpreted as reflecting separate skills (also known as latent constructs; see Table 2.12).

We conducted exploratory factor analysis to investigate the dimensional structure of the 16 scores for the 14 TILLS subtests listed in Table 2.12 (note that the Written Expression [WE] subtest yields 3 scores, accounting for there being more scores than subtests). A two-factor solution was supported by the data (see technical notes with Table 2.12). Factor 1 is defined by the Phonemic Awareness (PA), Nonword Spelling (NWSpell), Nonword Reading (NWRead), and Reading Fluency (RF) subtests and the Written

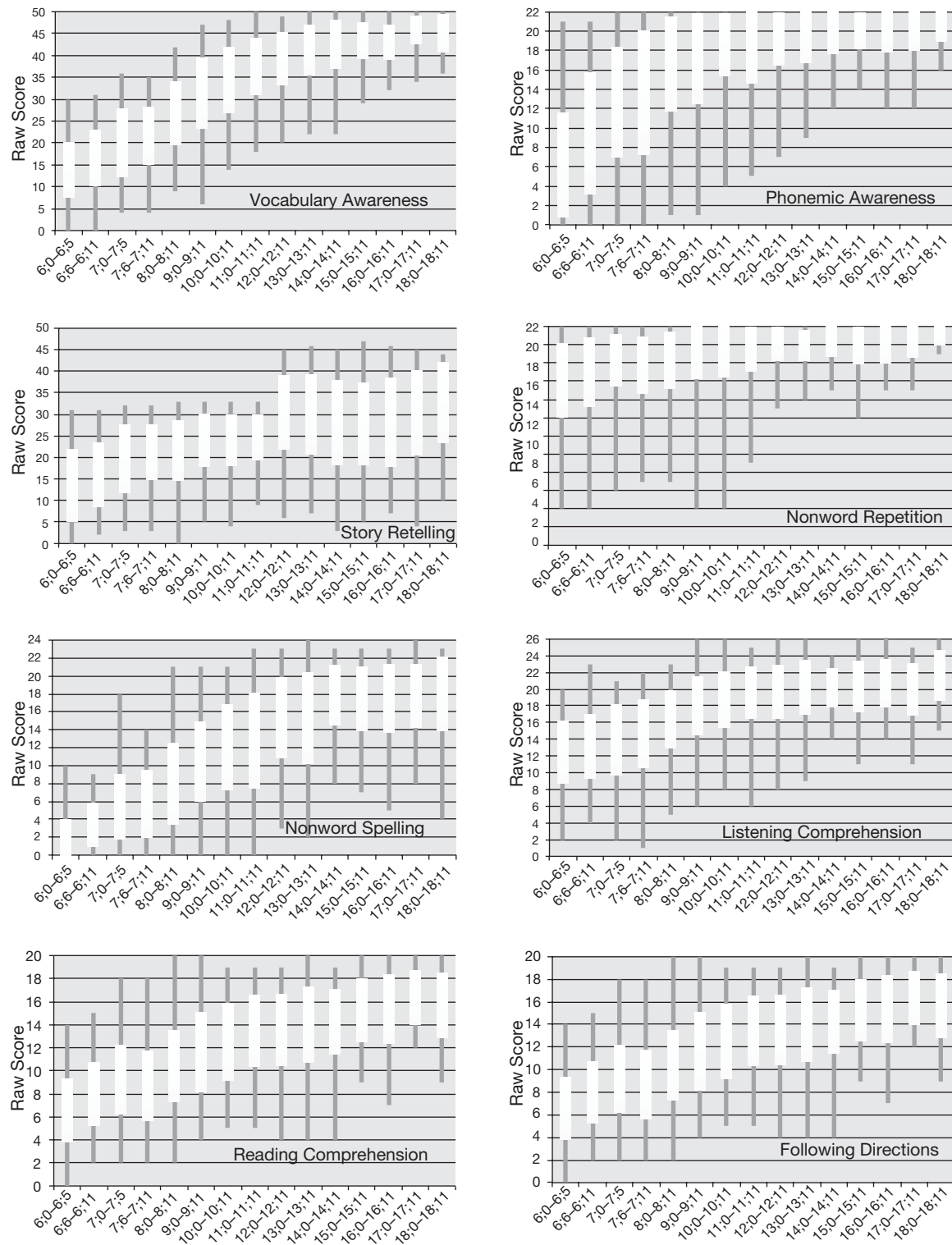


Figure 2.4. Age-related change in TILLS subtests. The bars indicate ± 1 standard deviation around the mean, and the lines indicate the minimum to maximum scores at each age.

FOR MORE, go to <http://www.tillstest.com>

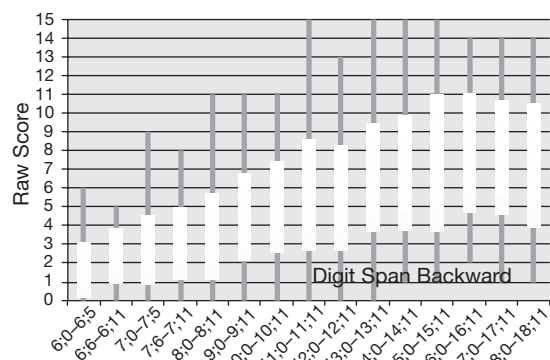
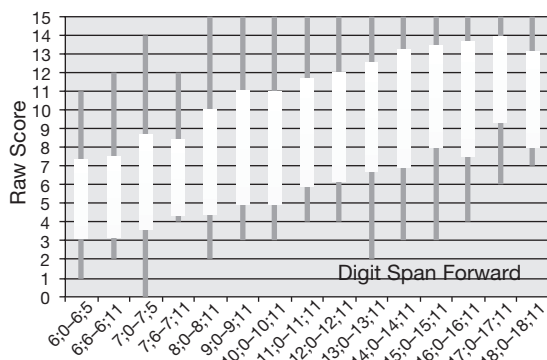
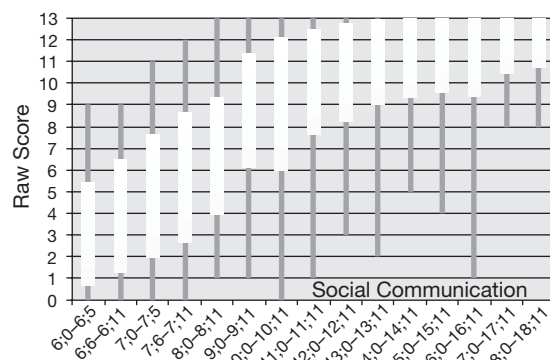
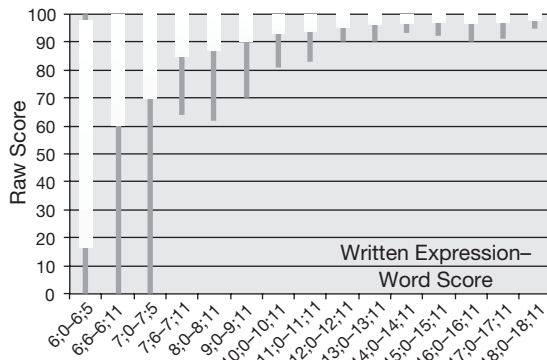
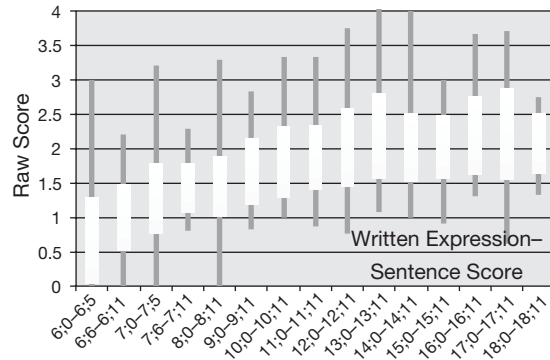
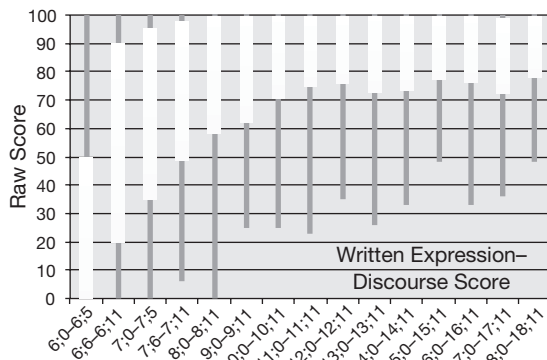
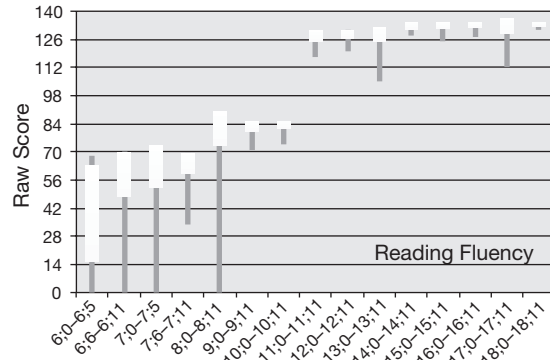
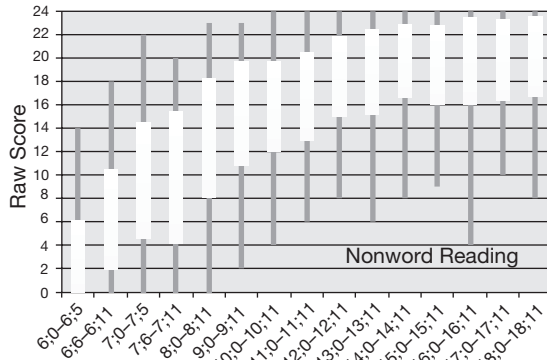
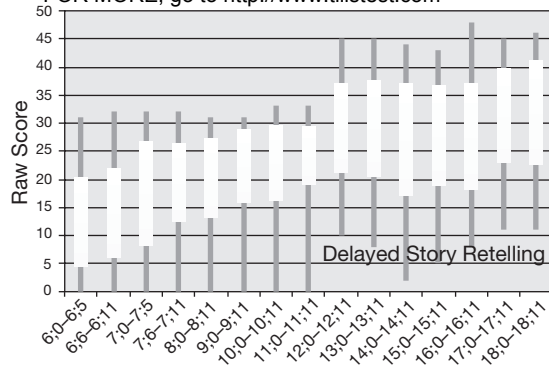


Table 2.12. Two-factor reference structure for TILLS subtests based on exploratory factor analysis

Subtest number and name	Factor reference structure (semipartial correlations)	
	Factor 1 Sound/word-level knowledge	Factor 2 Sentence/discourse-level knowledge
2. Phonemic Awareness (PA)	0.547	0.074
10. Nonword Reading (NWRead)	0.734	–0.058
11. Reading Fluency (RF)	0.406	0.077
5. Nonword Spelling (NWSpell)	0.600	0.067
12. Written Expression–Word Score (WE-Word)	0.409	0.009
1. Vocabulary Awareness (VA)	0.229	0.472
3. Story Retelling (SR)	–0.052	0.500
6. Listening Comprehension (LC)	0.009	0.548
7. Reading Comprehension (RC)	0.264	0.420
8. Following Directions (FD)	0.153	0.409
13. Social Communication (SC)	0.075	0.476
12. Written Expression–Sentence Score (WE-Sent)	0.233	0.258
12. Written Expression–Discourse Score (WE-Disc)	0.220	0.167
4. Nonword Repetition (NWRRep)	0.260	0.204
14. Digit Span Forward (DSF)	0.291	0.175
15. Digit Span Backward (DSB)	0.329	0.119

Technical notes: We based factor analysis (common factor model) on a sample of 1,200 cases drawn from the standardization sample. This sample was drawn from a larger sample of 1,462 cases such that students without disorders (typically developing students) and students with disorders were represented at approximately the levels of their occurrence in the general population. We randomly drew students from the larger sample so that their representation reflected the following proportions: typically developing students = .81, students with language and literacy disorders = .13, students at risk for language and literacy disorders = .05, and students with special education disorders (autism spectrum disorder, deafness or hard of hearing, intellectual disability) = .01.

We subjected subtests scores from the TILLS to an exploratory factor analysis that used a Promax factor rotation. Semipartial correlations are reported in Table 2.12, and values above .30 provide evidence that a subtest is associated with measurement of a particular factor. We considered Delayed Story Retelling (DSR) for inclusion but omitted it from the factor analysis due to its very close relationship to Story Retelling (SR).

The factor analysis suggested that two factors should be retained for rotation. Two forms of evidence supported this conclusion: 1) traditional Eigenvalue analysis showed two Eigenvalues above the conventional 1.0 value, which explained 100% of the common variance in the correlation matrix, and more important, 2) a parallel analysis simulation of 100 iterations clearly showed that two factors should be considered for rotation. Factor extraction was based on maximum likelihood estimation, initial communalities estimates set to squared multiple correlations with oblique (Promax) rotation (power = 2.8).

Note: Subtests that did not meet minimum loading values (< 0.30) on either factor are unshaded.

Expression–Word score (WE-Word). Factor 1 thus is consistent with the sound/word component represented in the theoretical test model for the TILLS. Factor 2 is defined by the Vocabulary Awareness (VA), Story Retelling (SR), Listening Comprehension (LC), Reading Comprehension (RC), Following Directions (FD), and Social Communication (SC) subtests. Factor 2 subtests are consistent with the sentence/discourse component represented in the theoretical model for the TILLS. Note that it makes sense that the Vocabulary Awareness (VA) subtest loads with the sentence/discourse factor because this particular vocabulary subtest assesses semantic knowledge and relationships between word meanings rather than word structure knowledge. Semantic knowledge is associated with the construction of meaning (both receptively and expressively). The VA subtest also requires learners to switch cognitive sets to associate meaning across different word pairs, in addition to understanding the meanings of individual words. Thus, the VA subtest is related more closely to the factor that defines integrated language

comprehension and formulation than to the factor that defines word structure knowledge required for speech perception and production, reading decoding, and spelling.

Subtests that did not meet minimum loading values (<0.30) on either factor included Nonword Repetition (NWRep), Written Expression–Sentence score (WE-Sent), Written Expression–Discourse score (WE-Disc), Digit Span Forward (DSF), and Digit Span Backward (DSB). These subtests remain unshaded in Table 2.12.

The interfactor correlation between the sound/word factor and sentence/discourse factor is 0.598, showing that these are related skills, although they can be viewed as distinctive aspects of language and literacy knowledge. It is important to note that the factor analysis of the TILLS results did *not* support separate factors associated either with receptive and expressive language or with oral and written language. This is consistent with the TILLS theoretical model as well. That is, it supports the concept that oral and written language skills at the same language level have more in common with each other than do language skills that are distinguished by modality (listening, speaking, reading, or writing) or language skills divided in terms of receptive versus expressive modes.

In summary, evidence that the TILLS measures more than a single construct supports the purpose of the TILLS to identify patterns of strengths and weaknesses that are theoretically consistent with the test model. Evidence that language levels (sound/word versus sentence/discourse) are stronger factors than separate modalities (e.g., receptive versus expressive language) supports the model used to construct the TILLS.

CONCURRENT VALIDITY

The TILLS measures skills in a way fundamentally different from most other tests in that it includes both oral and written subtests, often in parallel forms. In addition, it includes subtests that necessarily integrate language skills. However, it does measure skills that are also measured by other separate tests. When tests that measure similar skills correlate, this is taken as evidence of concurrent validity. To conduct this analysis, we asked a subset of students who participated in the TILLS standardization research to take additional tests selected as concurrent measures (i.e., they measured similar constructs). The correlations between scores on these tests and TILLS scores are reported in Table 2.13. These results showed that the TILLS subtest scores correlated significantly with other tests that purport to measure similar skills, despite differences in the specific content and test formats for the TILLS and these alternative language and literacy measures.

FREQUENTLY ASKED QUESTIONS

1. The TILLS manuals emphasize the use of specific cut scores. Why is this important?

Validity information supporting identification no longer applies if the correct cut score is not used to determine the presence or absence of a disorder. Using a higher score than the published cut score will increase the rate of over identification of typically developing students as having disorders. Using a lower score than the published cut score will result in under identification of students who truly have a disorder.

2. Other tests report group differences as evidence that the test can identify disorders. Why does the TILLS present sensitivity and specificity data instead?

Statistical analyses of differences (e.g., t-tests, ANOVA results) can reveal significant group differences even when there is a relatively high proportion of overlap

Table 2.13. Correlation of TILLS subtests with tests measuring similar content

TILLS subtest	Concurrent measure	<i>n</i>	<i>r</i>
1. Vocabulary Awareness (VA)	Word Classes Total subtest, CELF-4	64	.65*
2. Phonemic Awareness (PA)	Elision subtest, CTOPP	77	.51*
3. Story Retelling (SR)	Oral Narration subtest, TNL	41	.71*
4. Nonword Repetition (NWRRep)	The Children's Test of Nonword Repetition	72	.39*
5. Nonword Spelling (NWSpell)	Test of Written Spelling, Fourth Edition (TWS-4)	76	.56*
6. Listening Comprehension (LC)	Narrative Comprehension subtest, TNL	39	.37*
7. Reading Comprehension (RC)	Comprehension subtest, GORT-4	19	.65*
8. Following Directions (FD)	Concepts and Following Directions subtest, CELF-4	41	.38*
10. Nonword Reading (NWRRead)	Word Attack subtest, WRMT-3	24	.50*
11. Reading Fluency (RF)	Fluency subtest, GORT-4	18	.72*

*Statistically significant correlations ($p < .05$).

Note: z-score transformations (raw scores scaled to a mean of 0 and a standard deviation of ± 1) for TILLS subtests were correlated with standard scores from each concurrent measure.

Key: CELF-4, Clinical Evaluation of Language Fundamentals, Fourth Edition (Semel, Wiig, & Secord, 2003); The Children's Test of Nonword Repetition (Gathercole, Willis, Baddeley, & Emslie, 1994); CTOPP, Comprehensive Test of Phonological Processing, First Edition (Wagner, Torgesen, & Rashotte, 1999); GORT-4, Gray Oral Reading Test, Fourth Edition (Wiederholt & Bryant, 2001); TNL, Test of Narrative Language (Gillam & Pearson, 2004); TWS-4, Test of Written Spelling, Fourth Edition (Larsen, Hammill, & Moats, 1999); WRMT-3, Woodcock Reading Mastery Tests, Third Edition (Woodcock, 2011).

Technical notes: All test scores were obtained by examiners whose educational training included test administration (e.g., speech-language pathologists, psychologists, special education teachers) and who received training in the administration of the TILLS. The *r* values are Pearson product-moment correlations, with the exception of Nonword Repetition (NWRRep). The Spearman rho is reported for this subtest due to nonnormal distributions.

between the scores of typically developing students and those with disorders. The statistical procedure used to derive sensitivity and specificity data is more rigorous because it considers the classification of the individual students instead of the group as a whole. This provides the clinician with much stronger information concerning how accurate a diagnosis is likely to be (when the correct cut score is used) than group difference data can provide.

3. If the TILLS scores show age-related change, why doesn't the TILLS report age (or grade) scores?

Although age scores appear to communicate that a student is functioning like typically developing children of a certain age (or at a certain grade level), this is actually not the case. Often there is too much overlap among the scores from typically developing children of different ages to assign a particular score to a particular age. This can be appreciated by noting the overlapping range of scores obtained at different ages for TILLS subtests in Figure 2.4. This overlap not only occurs for the TILLS but for most, if not all, tests that measure skills that develop during childhood. Calculation of age scores necessarily obscures this overlap. As a result, age and grade scores can mislead parents, educators, and the clinician into thinking that a child is showing scores that are either more advanced or more impaired than they truly are. For this reason, age and grade scores have been widely criticized as indefensible on mathematical and conceptual grounds, and therefore, they are not used with the TILLS.