4

Development and Standardization

This chapter describes the development of the TOD, beginning with the theoretical and empirical underpinnings of the assessment. The chapter then details the research samples collected to standardize and validate the TOD tests, as well as the methods employed to derive the TOD scores.

Theoretical and Empirical Underpinnings of the TOD

This section defines dyslexia, as well as its etiological underpinnings (e.g., neurological and hereditary influences), typical comorbid conditions and cognitive/linguistic correlates, traditional and current assessment options, and how the characteristics of dyslexia relate to the TOD tests, indexes, and composites.

Definition and Neurobiology of Dyslexia

The word *dyslexia* comes from the Greek words *dys*, meaning impaired, and *lexis*, meaning word. Although variations in definitions exist, the authors of the TOD took into account the following definition that was adopted by the International Dyslexia Association (IDA) Board, November 2002 (Lyon et al., 2003):

> Dyslexia is a specific learning disability that is neurological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge (p. 2).

Within their dyslexia screening legislation, most states acknowledge this definition (Gearin et al., 2021). Although the emphasis is on the phonological component of language in the IDA definition, other linguistic abilities, such as verbal memory and rapid naming, are included in other definitions (e.g., British Dyslexia Association, Dyslexia Association of Ireland). Additionally, as with any learning disorder, dyslexia exists along a continuum, where the level of impact can range from mild to severe (Hasbrouck, 2020). The impact of dyslexia is also influenced by the environment and any effects due to appropriate early intervention and treatment.

Both environmental and genetic factors influence reading development (Little & Hart, 2022).

The neurobiological basis of dyslexia has been supported by research that shows brain functioning in individuals with dyslexia differs from that of typical readers in a few important ways (Shaywitz & Shaywitz, 2020). For example, studies using functional magnetic resonance imaging (fMRI) of individuals with and without dyslexia indicate a "different distribution of metabolic activation" in the two groups (Hudson et al., 2007, p. 510). Seminal studies (e.g., Cohen et al., 2004; Dehaene et al., 2005; Shaywitz et al., 2002) revealed that individuals with dyslexia experience a disruption/decrease in metabolic activity in the posterior part of the left hemisphere (the left visual word form area), which is associated with skilled and fluent reading, and an overactivation of the left and right anterior systems and the right visual word form area. These neural differences create inefficiencies in producing skilled reading. Researchers have hypothesized that these less efficient areas are activated in order to compensate for the disruption in the "reading areas" of the brain.

In addition to research establishing how the brain works during various types of reading tasks, growing evidence indicates that systematic reading instruction can result in changes in the patterns of metabolic activity in the brain. For example, Aylward et al. (2003) found that children identified as having dyslexia who were exposed to 20 hours of reading intervention showed significant differences in activation (observed pre- and post-intervention via fMRI) in areas of the brain important for reading and language while doing a phonological task. In other words, the brain activity of readers with dyslexia post-intervention was similar to that of typical readers. Similar results have been found using tasks involving phonemic awareness, phonics, word recognition, and morphology (Shaywitz et al., 2004; Shaywitz & Shaywitz, 2020; Weiss et al., 2010). Because the use of fMRI is expensive and does not yield specific diagnostic information, it is not practical for use in schools as a diagnostic tool; however, the results from fMRI and related approaches continue to enhance our understanding of the reading process and dyslexia.

Heritability and Comorbidity

Strong evidence supports the familial transmission of dyslexia (Hulme & Snowling, 2009; Lasnick et al., 2022; Moll, 2022; Pennington & Lefty, 2001). Approximately 45% of children who have one parent with dyslexia will develop this reading disability (Gaab et al., 2020; Snowling & Melby-Lervåg, 2016). If both parents have dyslexia, the chance increases to 75% for developing this reading disorder (Gaab et al., 2020). Thus, family history is one of the strongest risk factors for developing dyslexia. When considering familial transmission of dyslexia, it is important to note that parents may have "reading and spelling problems" but not know they have a disability.

Another line of related research has focused on comorbidity, i.e., the extent to which individuals with a particular disorder, in this case dyslexia, will also demonstrate characteristics consistent with the diagnosis of another disorder. Approximately 40% to 60% of children with dyslexia will also meet the diagnostic criteria for another disorder, including attention-deficit/hyperactivity disorder, speech sound disorder, dyscalculia (a learning disability in math), dysgraphia (a learning disability in writing), or developmental language disorder (Moll et al., 2020; Pennington et al., 2019).

Unexpected Underachievement and Dyslexia Identification Under IDEA

Historically, the concept of "unexpected underachievement" has been a central defining characteristic of dyslexia, as has the assumption of specificity (i.e., the underlying deficit is specific to reading and related skills and does not extend significantly into other domains; Hinshelwood, 1902; Tunmer & Greaney, 2010). The disorder is considered "unexpected" in that other abilities are often intact and educational opportunities are assumed to have been sufficient. This concept of unexpected underachievement has perhaps been the most consistent and enduring component of the definition of dyslexia (Shaywitz & Shaywitz, 2020).

Because dyslexia is a type of learning disability, one of the three procedures described in IDEA 2004 must be followed to identify dyslexia in schools:

1. Ability–achievement discrepancy (though this must not be required); 2. Alternative research-based methods, often operationalized as a pattern of strengths and weaknesses (PSW) approach; 3. A student's response to scientific, research-based intervention, often referred to as *response to intervention* (RTI). Chapter 3 describes interpretation within these frameworks and provides an explanation for the rationale behind the methods. The following paragraphs explain how the TOD fits into them.

Ability–Achievement Discrepancy

Prior to the 2004 reauthorization of IDEA, an ability–achievement discrepancy identification procedure was incorporated into the federal law as a criterion for specific learning disability (SLD) identification. This procedure was used as an attempt to capture the difference between one's expected performance (based on criteria such as general intelligence or oral language) and reading. This approach was made optional in IDEA 2004, and although some school districts still use it as part of an evaluation for specific learning disabilities, it has been criticized as being invalid and "archaic and inappropriate" (Siegel & Hurford, 2019, p. 23), as a "wait to fail" strategy that is inefficient (Stanovich, 1991), and as being no longer accepted practice (Snowling, 2014).

Other experts, however, have noted that the existence of an ability–achievement discrepancy can provide useful information for a learning disability (LD) diagnosis because it is consistent with the very essence of its definition, i.e., an unexpected weakness within a profile of salient strengths (e.g., Bell et al., 2015; Hays et al., 2017; Johnson & Myklebust, 1967; McCallum et al., 2013). This type of discrepancy is particularly pertinent to the identification of twice-exceptional students with dyslexia (Shaywitz & Shaywitz, 2020).

The National Joint Committee on Learning Disabilities (2011) explained that "individuals identified as intellectually gifted may also have LD. Although twice-exceptional individuals may appear to be functioning adequately in the classroom, their performance may be far below what they are capable of, given their intellectual ability. … Educators often overlook these students until late in their academic careers" (p. 238). Furthermore, in examining results from their Connecticut Longitudinal Study, Shaywitz and Shaywitz (2020) explained that in typical readers a strong correlation exists between intelligence test scores and reading performance, whereas in readers with dyslexia, there is little to no relationship between intelligence test scores and reading, which validates the unexpected nature of dyslexia (p. 102).

Similar to the logic underlying the consideration of intellectual ability in an ability–achievement discrepancy model, a disparity between oral language abilities and reading can be considered another indicator of dyslexia. Torgesen (2000) explained that "children should be able to comprehend, or construct, the meaning of what is being read at a level consistent with their general verbal ability" (p. 55). Thus, students with dyslexia may have an ability–achievement discrepancy when their intelligence test or oral language scores are higher than their reading scores.

To provide an estimate of oral language abilities, the TOD-S contains the Picture Vocabulary (1S) test, and the TOD-C contains the Listening Vocabulary (22C) test. As noted, although dyslexia and developmental language disorders are two distinct disorders, they often co-occur (Moll et al., 2020; Spencer et al., 2014). Because of relatively high comorbidity between these disabilities, the TOD-C also contains two measures of reasoning ability, the Picture Analogies (10C) and Geometric Analogies (23C) tests. Chapters 2 and 3 describe procedures for comparing the reasoning test scores with other TOD test scores.

Pattern of Strengths and Weaknesses Approach

Some states and districts have adopted what has been referred to as a pattern of strengths and weaknesses (PSW) model. This approach explores strengths and weaknesses within both cognitive and achievement areas. One premise underlying the PSW model is that children with specific learning disabilities will have *specific* cognitive deficits that lead to poorer academic performance than would be expected given their *specific* cognitive assets. This model differentiates children with specific learning disabilities from those with lower global intelligence (Hale et al., 2006).

This identification approach requires that an examiner establish links between specific cognitive strengths and weaknesses and empirically linked strengths and weaknesses in academic areas. A PSW approach is often referred to as a *consistency model*, as the cognitive or linguistic skill deficits are linked to and predict poor reading and spelling performance, resulting in expected underachievement (Flanagan et al., 2018). As with ability–achievement discrepancy, the TOD was designed to be compatible with a PSW approach. The TOD includes tests of several linguistic processing skills that have been identified as predictors of reading and spelling difficulties (phonological awareness, rapid automatized naming [RAN], working memory, and orthographic processing). As examples, a student with poor phonological processing often has difficulty acquiring phonics, whereas a student with slow rapid automatized naming often has a slow reading rate.

Response to Intervention

Since the passage of IDEA 2004, many U.S. school districts have adopted a response to intervention (RTI) model, also referred to as multi-tiered system of support (MTSS), which identifies students within a school with the lowest reading scores, often obtained from brief curriculum-based measures of oral reading fluency. Within this framework, the unexpected nature of the reading problem is determined by the persistence of difficulty despite appropriate instruction.

Although RTI identifies poor readers and may provide students with timely interventions, it has some limitations because there are many possible reasons why a student would not respond to an intervention, including learning English as a second language, attentional difficulties, behavior problems, limited prerequisite skills, low language and reasoning abilities, limited or ineffective instruction, or an intellectual disability. Thus, RTI can provide interventions and demonstrate that a student is not making adequate progress, but additional information is needed to understand why that is the case and which intervention(s) might be most effective (Mather & Kaufman, 2006). An RTI model can, however, provide useful information regarding the effectiveness of applied interventions that are delivered within schools.

Another potential limitation is that students who do not meet criteria for "at-risk" status based on some cut score (e.g., lowest performing 25%) may be missed even though they have dyslexia, particularly those who are twice exceptional (gifted, with dyslexia). Despite their advanced cognitive abilities, these students may go undetected because they employ compensatory strategies that mask their reading weaknesses so that their reading skills fall within the average range (Reynolds & Shaywitz, 2009). Reading scores within the average range should not be used to negate a diagnosis of dyslexia, according to Shaywitz and Shaywitz (2020): "There is no one single test score that ensures a diagnosis of dyslexia. It is the overall picture that matters. An extremely bright child who has a reading score in the average range but who struggles and cannot learn to read fluently … has dyslexia" (p. 166).

Dyslexia Identification Under IDEA Summary

All three of the approaches described above have advantages and disadvantages for a dyslexia evaluation. As Kovaleski et al. (2015) note, "knowledgeable practitioners also use clinical judgment to determine which approach is applicable for a given child or in a given school setting. While regulations and policies require school districts to implement a single approach, best practice may reside somewhere in the margins with a hybrid model" (p. 6).

To aid the diagnostic process, the TOD provides global scores that operationalize dyslexia risk status and a dyslexia diagnosis, as described below. Data from any of these three approaches can inform the diagnostic process described in the *Diagnostic and Statistical Manual, Fifth Edition, Text Revision* (*DSM-5-TR*; American Psychiatric Association, 2022). The DSM-5-TR, used by psychologists and related professionals, addresses the criteria required for a specific learning disability diagnosis, of which dyslexia is one type.

Development of the TOD Tests

The TOD was designed to assess the reading and spelling, linguistic processing, and vocabulary and reasoning domains that are most relevant to a diagnosis of dyslexia. In addition, the dyslexia identification models used in schools and clinics were considered in the design of the TOD tests, as were individual differences and environmental factors that influence the trajectory of dyslexia.

The extant literature indicated that the TOD should include measures of regular and irregular word reading (both untimed and timed), reading rate, reading comprehension efficiency (timed), and spelling, the primary areas of difficulty for individuals with dyslexia. In addition, the TOD should incorporate measures of the main linguistic risk factors for dyslexia, i.e., the underlying abilities that can impede the development of reading and spelling skills, such as phonological awareness, rapid automatized naming, and working memory. Finally, the TOD should provide ways to determine the "unexpected" nature of dyslexia by assessing oral vocabulary and reasoning, particularly when dyslexia is not accompanied by other comorbid conditions, such as a developmental language disorder. These skills are described in detail below.

In addition, the TOD should include parent, teacher, and self-rating scales for gathering additional qualitative indicators of dyslexia. The TOD authors adopted as a fundamental guiding principle Wagner's (2018) perspective that the use of multiple indicators improves the reliability, validity, and utility of a diagnosis.

Finally, the authors determined that three separate test batteries were needed. The first would be a screener (TOD-Screener) that could be group administered and identify risk for reading failure. The second (TOD-Early) would be designed to target the essential predictors and skills for early reading (K–Grade 1). The TOD-E is appropriate for children who are prereaders or emerging readers. The third battery (TOD-Comprehensive) would encompass all major elements included in a dyslexia evaluation and would be appropriate for examinees of all ages, beginning with first grade and extending into adulthood.

Primary Areas of Difficulty

As noted, the primary areas of difficulty for individuals with dyslexia are accuracy of word reading, reading rate, and spelling (Fletcher et al., 2019; Wagner et al., 2022); therefore, the TOD includes several measures of each of these abilities.

Phonics Skills and Word Reading

A core problem for individuals with dyslexia is accurate and fluent word reading (Lowell et al., 2014). To assess an individual's ability to read words, an evaluation for dyslexia needs to include measures of context-free word identification skills (Tunmer & Greaney, 2010). Thus, a critical component of a dyslexia evaluation is assessing word-reading accuracy and efficiency with both pseudowords and real words in both a timed and untimed format (Lindstrom, 2019; Siegel & Hurford, 2019).

Pseudoword reading Pseudowords (sometimes called *nonsense words* or *nonwords*) are combinations of letters that conform to English spelling patterns and rules (e.g., *toam, flib*). According to many experts (e.g., Rack et al., 1992; Shaywitz & Shaywitz, 2020), difficulty in reading pseudowords is one of the most important indicators of dyslexia. Pseudowords are hard to pronounce for individuals with dyslexia, who often have difficulty learning the mappings between the speech sounds and the printed letters or acquiring phonics skills. Because pseudowords are not real words, they can be pronounced only through the application of English phonic rules.

Children with severe problems in phonological processing have the most difficulty reading and spelling pseudowords (Hulme & Snowling, 2009), in what has been referred to as *phonological recoding*. Herrmann et al. (2006) explain phonological recoding as follows: "Phonological recoding involves 'sounding out' printed words using knowledge of letter–sound relationships. Pseudoword reading accuracy provides an index of the success with which unfamiliar words can be read aloud using phonological recoding strategies" (p. 196).

The TOD-E contains two tests for assessing pseudoword reading ability and related emerging skills. The Sounds and Pseudowords (4E) test assesses letter–sound knowledge and pseudoword (nonsense word) reading. The Letter and Sound Knowledge (9E) test requires identifying the first, last, or middle sound in a word. The TOD-C also contains two tests for assessing pseudoword reading, the Pseudoword Reading (7C) and Rapid Pseudoword Reading (19C) tests. Test 7C assesses the accuracy of pseudoword reading, whereas Test 19C assesses the efficiency and rate of pseudoword reading, or the examinee's ability to quickly recognize phonic patterns, sometimes referred to as automaticity. These types of measures can help provide insight into an individual's mastery and automaticity with both beginning and more advanced phonics skills.

Irregular word reading Typically, beginning readers develop increased skill in recognizing the orthographic patterns in words based upon the predictable patterns of letter order (Hasbrouck, 2020). Individuals with dyslexia, however, often have difficulty mastering and recalling the orthographic spelling patterns of their language. They have difficulty forming stable word or lexical representations, which then affects the development of both irregular word reading and spelling. To measure these abilities, the TOD-E has the Letter and Sight Word Recognition (7E) test, which includes naming letters and reading both high-frequency and irregular words. The TOD-C has the Irregular Word Reading (11C) and Rapid Irregular Word Reading (20C) tests. Test 11C measures the accuracy of irregular word reading, whereas Test 20C assesses the efficiency and rate of irregular word reading.

Spelling

Difficulty spelling words is an indicator of dyslexia across the life span and, in some cases, can be the only remaining characteristic of dyslexia in adults (Romani et al., 2007). As with word reading, a dyslexia evaluation should include measures of both regular word spelling (words that conform to typical English spelling patterns and rules) and irregular word spelling (words that have one or more parts that do not conform to typical spelling patterns or rules). When learning to spell regular words, students typically first learn to segment the speech sounds and then attach these sounds (phonemes) to the letters (graphemes). However, good phonemic awareness skills are insufficient for spelling both irregular words and homophones (e.g., *there* and *their*; Lowell et al., 2014).

In addition to the ability to segment sounds, spelling requires orthographic processing, often defined as the ability to recall the specific letter and spelling patterns that make up words (Mather & Jaffe, 2021). Some individuals with dyslexia can spell regular words with accuracy but have difficulty spelling irregular words. Most likely, these individuals have good knowledge of sound–symbol correspondences but poor lexical representations that do not develop completely; the errors that they make typically involve regularizing the irregular element of a word, such as spelling the word *said* as *sed* (Romani et al., 2007; Willows & Terepocki, 1993). The TOD-E includes the Sounds and Pseudowords (4E) test, which measures early spelling skills (identifying beginning sounds, providing sounds for letters, recognizing phonically regular pseudowords). The TOD-C includes the Irregular Word Spelling (5C) and Regular Word Spelling (15C) tests, which can help an examiner differentiate between individuals who have poor lexical representations, and those who have difficulty sequencing sounds with their corresponding letters. Additional measures designed to assess basic orthographic processing ability are described in the Orthographic Processing section later in this chapter.

Reading Rate/Reading Comprehension Efficiency

An evaluation for dyslexia should also include timed measures of word recognition to capture the development of reading fluency and automaticity (Tunmer & Greaney, 2010). Individuals with dyslexia are slow reading orally because their automatic decoding skills are weak, so timed measures of word reading, pseudoword reading, and paragraph-level reading should be part of a comprehensive dyslexia evaluation (Pennington et al., 2019).

In addition to the TOD-C tests that assess reading decoding efficiency/rate (i.e., the Rapid Pseudoword Reading [19C] and Rapid Irregular Word Reading [20C] tests), the TOD-S includes the timed Word Reading Fluency (3Sa) test for Grades K–1 (marking the word that goes with a picture) and the timed Question Reading Fluency (3Sb) test for Grade 2–Adult (reading a question and marking the correct answer).

The TOD-C also includes the Oral Reading Efficiency (12C) test (oral reading of a passage for 1 minute) and Silent Reading Efficiency (16C) test (reading paragraphs and answering questions within a time limit). Results from these timed tests can have direct implications for making decisions about the need for the accommodation of extended time, the most common accommodation requested by students with dyslexia as well as the most critical accommodation (Shaywitz & Shaywitz, 2020).

Linguistic Processing

To accurately diagnose an individual with dyslexia, an examiner should assess the various linguistic processes that impact reading and spelling development. Research has suggested that a multiple-deficit model for identifying dyslexia is superior to a single-deficit model. Specifically, a multiple-deficit model is consistent with empirical research results showing multifactorial cognitive and linguistic influences on reading and spelling (Bell et al., 2003; Compton, 2020; Pennington et al., 2019; Perry et al., 2019; Peterson & Pennington, 2015). Reliance on only one factor would exclude a large number of individuals who have dyslexia (Brady, 2019; Pennington et al., 2019).

Four main linguistic processing deficits—phonological processing, working memory, rapid automatized naming (RAN; Callinan et al., 2013; Fletcher et al., 2019; Rose, 2009), and orthographic processing (Georgiou et al., 2021)—have been described as risk factors and marker variables for dyslexia diagnosis. Although other possible correlates have been discussed in the literature (e.g., morphological awareness, processing speed, memory span, visual attention), currently, the most consistent findings support the measurement of these four correlates. In addition, several well-documented preschool measures have been identified as strong predictors of later reading skill (i.e., phoneme awareness, letter name and sound knowledge, and RAN; Peterson & Pennington, 2015).

Recent research also suggests that individuals with dyslexia have difficulty with paired-associate learning tasks that require pairing and recalling speech sounds with symbols. The important point to keep in mind is that multiple differing cognitive profiles have been associated with dyslexia (Bell et al., 2003; Brady, 2019), and that individuals with multiple cognitive deficits are at much higher risk for dyslexia than those with a weakness in only one area (Norton & Wolf, 2012; Pennington et al., 2019).

Phonological Awareness

Phonological awareness is the ability to distinguish, understand, replicate, and manipulate sounds comprised within a language (Norton & Wolf, 2012). It is a stronger predictor of basic reading skills than of reading fluency and comprehension (Bell et al., 2003; Gray & McCutchen, 2006; Kibby et al., 2014). Children with dyslexia show lower performance on phonological processing and phonemic awareness tasks (manipulating individual phonemes) when compared to peers and children who are matched on reading levels (Kilpatrick, 2015; Melby-Lervåg et al., 2012; Snowling, 2014).

For young readers, the tasks of rhyming words, blending (pushing together speech sounds), and segmenting (breaking apart speech sounds) are important phonemic awareness tasks. Blending is an essential step for acquiring phonics, and segmenting is an essential step for spelling. To assess these abilities, the TOD-E includes two tests: Rhyming (5E) and Early Segmenting (8E). The TOD-C also includes two tests that measure blending and segmenting: Blending (13C) and Segmenting (14C).

For older students, phonemic manipulation tasks, such as substituting sounds in words and deleting sounds from words, are more indicative of continued difficulties with speech sounds; these more difficult phonemic skills contribute to more detailed analysis of the internal structure of words and the acquisition of more fully specified orthographic representations (Ehri, 2007, 2014; Kilpatrick, 2015). The TOD-C includes a test that assesses these skills: Phonological Manipulation (4C). Although the results from many of the TOD tests have direct instructional implications for individuals with dyslexia, the results from Test 4C are particularly important because of the continued relationships of these skills to the most fundamental understanding of speech sounds.

Rapid Automatized Naming

Rapid automatized naming (RAN) has also been identified as a correlate of reading problems (Hulme & Snowling, 2009; Wolf & Bowers, 1999). The relationship of RAN to reading has been explored in a variety of alphabetic languages. For example, in a longitudinal study of four languages (English, Spanish, Slovak, and Czech), Caravolas et al. (2012) found that at the start of literacy instruction, phoneme awareness, letter–sound knowledge, and RAN were the *most* reliable predictors of students' later reading and spelling skills. This study suggests that these abilities are important for reading in all alphabetic orthographies.

RAN has been described as a measure that is a strong predictor of performance on timed reading measures (Norton & Wolf, 2012). In contrast to phonological awareness, RAN is a better predictor of reading fluency than of basic reading skills (Abu-Hamour, 2010; Kibby et al., 2014; Wolf & Bowers, 1999). When slow RAN performance is combined with other deficits, such as poor phonological awareness and working memory, learning to read can be quite difficult (Lowell et al., 2014; Norton & Wolf, 2012).

The TOD includes several measures of RAN. The TOD-E contains the Early Rapid Number and Letter Naming (6E) test. The letters and numbers selected for this test—*A, B, C; 1, 2, 3*—were chosen for young children because they are often the first letters and numbers that children learn. The TOD-C contains the Rapid Letter Naming (6C) test. The lowercase letter pairs used in this test have been identified as the most common confusable ones, such as *b* and *d* and *p* and *q*, which are often more difficult for individuals with dyslexia to master. The TOD-C also contains the Rapid Number and Letter Naming (17C) test. Similarly, some of the uppercase letters and numbers used in this test were selected because of their visual similarity (e.g., *6* and *9*). Continued confusion of these symbols is a symptom of reading difficulties and can occur in older students who have serious delays in reading development (Tunmer & Greaney, 2010). In fact, Al Dahhan et al. (2020) found that, when compared to typical readers on naming speed tasks, adult readers with dyslexia had longer fixation durations, more regressions, and increased neural activity when the letter stimuli were both phonologically and visually similar, such as with the letters *b, p,* and *d.*

Auditory Working Memory

Reading requires working memory and, consequently, an assessment of students who may have dyslexia should include tests of working memory (Dehn, 2008). Although the exact nature of the role of working memory in dyslexia has not been confirmed, numerous studies have indicated that poor memory is a correlate of dyslexia and that auditory working memory predicts many aspects of reading (Kibby et al., 2014; McCallum et al., 2006). For example, McCallum et al. reported significant relationships between a visual (working memory) and an auditory (rote memory) task and several aspects of reading. In addition, results indicated that auditory memory contributed significantly to the prediction of reading decoding and reading fluency ($p < .05$) beyond the power of phonology, orthography, and RAN. Similarly, Dehn (2008) found verbal working memory to be significantly related to word-reading skills, as well as reading comprehension. For most individuals, both reading and working memory capacity improve with age. However, results from a 3-year longitudinal study indicate that skilled readers have stronger growth in working memory than children with dyslexia, and that working memory is most related to growth in reading fluency and reading comprehension (Swanson & Jerman, 2007). The TOD-C contains two measures of working memory: Word Memory (9C) and Letter Memory (18C), which require listening to a string of words or letters and then repeating them in reverse order.

Orthographic Processing

Orthographic processing is the ability to recall the specific letter and spelling patterns that words comprise. To succeed in orthographic processing tasks, individuals must have stored visual images of the correct spellings of words; the tasks cannot be accomplished using only spelling–sound correspondences. These types of tasks involve sensitivity to the order of letters in words (e.g., knowing that English words cannot begin with a *ck* or an *ff*), as well as the ability to recognize and recall specific letters and letter patterns that represent words in print. Knowledge and memory of these lexical representations are often measured with word choice or homophone choice tasks, such as choosing the correct spelling of the word *soap* (e.g., *sope* or *soap*; Olson et al., 1985) or choosing which of two homophones is a flower (e.g., *rows* or *rose*; Stanovich & West, 1989). Mastery of these types of tasks requires specific orthographic knowledge, which is often a weakness in individuals with dyslexia. Findings from a meta-analysis indicated that individuals with dyslexia have a deficit in orthographic knowledge that is as large as their deficits in phonological awareness and RAN (Georgiou et al., 2021).

To measure orthographic processing, the TOD-S includes the Letter and Word Choice (2S) test (included in the TOD-C and TOD-E). This test requires listening to a word presented orally by the examiner and then marking the correctly spelled word from four choices (e.g., *prak, park, karp, rakp*). This type of task is designed to capture the processing efficiency of the orthographic lexicon (Perry et al., 2019), or, in other words, to determine how accurately an individual can recognize the correct spelling of a word from orthographically similar alternatives. Similar to a task developed by Treiman (1993), the TOD-C Word Pattern Choice (8C) test requires choosing the nonword that best conforms to the spelling patterns of English or looks the most like a real English word (e.g., *mpab, pmab, bamp, mpba*). Tests 2S and 8C will be particularly difficult for individuals with dyslexia who have poorly specified lexical representations.

Visual–Verbal Paired-Associate Learning

A critical early stage in learning to read is the process of mapping phonemes to graphemes. This process of mapping phonemes onto the orthographic patterns of words leads eventually to automaticity and immediate recognition of a word (Ehri, 2007, 2014, 2020; Kilpatrick, 2015). An individual with dyslexia has difficulty mapping phonemes to graphemes; consequently, automaticity does not develop with ease (Fletcher et al., 2019). This process of pairing letters with sounds involves what is known as *paired-associate learning* (PAL). PAL involves learning and remembering two stimuli that are artificially associated (e.g., an abstract symbol with a speech sound; Mourgues et al., 2016).

In exploring the specificity and nature of PAL, Litt and Nation (2014) found that children with dyslexia exhibited deficits in visual–verbal and verbal–verbal PAL only (not in visual–visual PAL). They attributed these difficulties to an underlying deficit in “phonological form learning.” Hulme et al. (2007) found in a large sample of typically developing children that both phoneme awareness and PAL were independent predictors of variations in reading skill. Similarly, Warmington and Hulme (2012) found that PAL and RAN were unique predictors of word recognition, whereas PAL, RAN, and phoneme awareness were the best predictors of pseudoword reading. They explain that “the learning of mappings between orthography and phonology is critical for learning to read and likely operates at numerous levels, including the process of learning letter–sound correspondences and the learning of mappings at the level of single letters, letter groups, and whole words when acquiring a word recognition system” (p. 47). The TOD-C Symbol to Sound Learning (21C) test was designed to mimic the initial stages of learning the mappings between speech sounds and symbols, a difficult task for many individuals with dyslexia (Aravena et al., 2013). This test can help an examiner determine whether an individual has trouble recalling sounds with their symbols and then blending these sounds to form words. As with the Aravena study, novel symbols are used in 21C in order to rule out differences in previous exposure.

TOD Pilot Study

A pilot study was conducted to evaluate the functionality and psychometric characteristics of the TOD test items, for both the direct tests and the rating scales.

TOD-C Pilot

The TOD-C was piloted with a sample of 220 individuals in second grade through college. The sample was 20% Hispanic, 2% Asian, 11% Black, 62% White, and 6% Other, with 44% male and 56% female examinees. Thirteen percent of the sample had a clinical diagnosis and/or were in special education, and 12% came from families who did not attend any college. The fact that the sample was more heavily from college-educated families was taken into consideration when evaluating the pilot data. Statistical bias analysis revealed no systematic differences based on parents’ educational attainment (a common proxy for socioeconomic status [SES]).

A total of 24 tests were piloted for the TOD-C. Eleven tests were completed in a response booklet that included four different versions based on grade level (2–5, 6–8, 9–12, college/adult) and could be administered to a group of students all taking the same form. Thirteen tests were individually administered using a stimulus book.

Group-administered tests included the following:

- Letter and Word Choice
- Question Reading Fluency
- Letter–Word Search
- Irregular Word Spelling
- Listening Vocabulary
- Geometric Analogies
- Regular Word Spelling
- Letter Pattern Choice
- Silent Reading Efficiency
- Picture Analogies
- Picture Vocabulary

Individually administered tests included the following:

- Rapid Letter Naming
- Letter Memory
- Blending
- Segmentation
- Pseudoword Reading
- Rapid Pseudoword Reading
- Irregular Word Reading
- Rapid Irregular Word Reading
- Symbol to Sound Learning
- Rapid Number and Letter Naming
- Word Memory
- Substitution
- Deletion

Pilot data were analyzed using passing rates by age and examiner feedback. In addition, items from untimed tests were analyzed using the Rasch one-parameter model (Bond & Fox, 2001; Wright & Stone, 1979). Analyses were conducted using the software program jMetrik (Meyer, 2014). The Rasch model provides a measurement of item difficulty and person ability on a unitary scale. Each item and person in the study receives a Rasch-based numerical value, which can be used to determine a person's probability of success on any given item. The Rasch model specifies that the most precise measurement occurs when a person is tested with items whose difficulties are closely matched to that person's ability, as expressed on the Rasch numerical scale. Therefore, a well-constructed scale must include items with sufficient floors, ceilings, and item gradients, i.e., that span the entire range of abilities in the target population and that spread uniformly enough to provide reasonably precise measurement for all ability levels.

The Rasch model also provides information about the goodness-of-fit of each item to the scale, as well as differential item functioning between groups of interest (e.g., gender, race/ethnicity). Items that displayed any evidence of bias or problems with model fit, as determined by the Rasch analysis, were deleted. Final item order was determined using the Rasch estimate of item difficulty and selected for standardization and validation in a nationwide study.

Based on review of passing rates, examiner feedback, and item-response theory statistics, the following changes were made prior to the standardization study:

- Some artwork was redone or replaced to improve functionality.
- Items were changed, added, deleted, and rearranged to ensure better measurement, and some easier items were added to provide a better floor for the TOD-C for first graders.
- The Letter–Word Search test was eliminated because it did not discriminate well across the age range of the test.
- Oral Reading Efficiency (a 1-minute timed test) was added to the TOD-C to fill a gap in measuring oral reading ability.
- Three tests were identified to function as a group-administered screener (TOD-S) and also as part of the TOD-C:

    - Picture Vocabulary
    - Word Choice (renamed Letter and Word Choice)
    - Question Reading Fluency
        - Word Reading Fluency, an alternate version of Question Reading Fluency, was created for examinees in first grade with beginning reading ability.
- Start and stop rules were developed for individually administered tests.

Following these changes, the standardization version of the TOD-C consisted of 23 tests, the first three of which were designated as TOD-S tests.

TOD-E Pilot

The TOD-E was piloted with a sample of 66 individuals in prekindergarten through first grade. Thirty-eight percent of the sample were in prekindergarten, 29% in kindergarten, and 33% in first grade. The sample was 15% Hispanic, 11% Asian, 11% Black, 54% White, and 9% Other, with 50% male and 50% female examinees. Four percent of the sample had a clinical diagnosis and/or were in special education, and 8% came from families who did not attend any college. The fact that the sample was more heavily from college-educated families was taken into consideration when evaluating the pilot data. Statistical bias analysis revealed no systematic differences between lower and higher SES families.

A total of eight tests were piloted for the TOD-E. They were all individually administered using a stimulus book.

- Letter and Sound Knowledge
- Letter and Sight Word Recognition
- Rhyming
- Beginning Sounds
- Nonsense Word Repetition
- Early Rapid Letter and Number Naming
- Word Knowledge
- Early Segmenting

Based on review of passing rates, examiner feedback, and item-response theory statistics, the following changes were made prior to the standardization study:

- Some artwork was redone or replaced to improve functionality.
- Some easy items were eliminated, and more difficult ones were added to better cover the intended skill range.
- Three tests—Beginning Sounds, Nonsense Word Repetition, and Word Knowledge—were eliminated because they did not discriminate well.
- Sounds and Pseudowords was added as a replacement for Nonsense Word Repetition.
- Prekindergarten (age 4) was dropped because many of the tests did not function well at that age, and the decision was made to standardize the TOD-E tests through second grade to expand the age range of the tests.
- Oral Reading Efficiency (a 1-minute timed test) was added to fill a gap in measuring oral reading ability.
- The three TOD-C tests designated for use as a screener were added to the TOD-E:
    - Picture Vocabulary
    - Word Choice (renamed Letter and Word Choice)
    - Question Reading Fluency
        - Word Reading Fluency, an alternate version of Question Reading Fluency, was created for examinees in kindergarten and first grade with beginning reading ability.

Following these changes, the standardization version of the TOD-E consisted of 10 tests, the first three of which were designated as TOD-S tests.

Standardization and Validation Studies

Data to support publication of the TOD were collected from 2019 through 2021. The occurrence of the Covid-19 pandemic during this time impacted the latter portion of the data collection. Although in most cases the tests were administered in person under pre-pandemic conditions (i.e., face-to-face), some alternate administration procedures were used during the pandemic. Some cases were administered in person using personal protective equipment (most frequently this included masks worn by both the examiner and examinee). When compared, the different administrations produced no evidence to indicate different clinical interpretations (this equivalency study is described later in the chapter), and thus data collected during the pandemic were included in the standardization sample. Additionally, a small number of cases were administered remotely using digital easels presented through the Presence® Platform (www.presence.com). This sample included 19 individuals within a restricted age range (6–33 years) and thus was too small for a true equivalency study. Review of raw score mean differences between these individuals and the rest of the sample by age year showed no meaningful differences and therefore these individuals were also included in the standardization sample. Thus, the TOD scores are representative of administration formats found in remote and masked mid-pandemic administration, as well as traditional administration procedures. The standardization data set is considered to be robust and appropriate for deriving standard scores and related metrics (e.g., percentiles).

Several data sets were collected to support the publication of the TOD (each is described separately later in this chapter). One hundred and six data collectors from 39 states administered the TOD to examinees accessed through schools, neighborhoods, or community organizations. The goal was to collect normative reference samples that were representative of the U.S. population in terms of gender, race/ethnicity, and parental educational level (a well-established index of socioeconomic status). Due to oversampling that occurred during data collection, some participants were deleted to better represent the population parameters. The cases that were desampled were not meaningfully different on other demographics than the cases that were retained.

In addition to the standardization data collection, data from several clinical samples were collected to support the validity of the TOD (each is described separately later in this chapter). For example, individuals with a clinical or learning disability diagnosis were also included in the standardization data collection. To ensure adequate representation of the U.S. school-based population, those with high-incidence disabilities (for whom the TOD is likely to be used) were included within the standardization sample, based on age and primary diagnosis, so that the proportions of these individuals in the standardization sample would approximate their prevalence in the typical school-based population. Individuals with low-incidence disabilities (e.g., moderate ID, autism spectrum disorder) were excluded from the standardization sample and used for validation purposes only. A total of 2,518 examinees ranging in age from 5 to 89 years were included in the TOD standardization and validation samples.

Standardization Samples

TOD-S Standardization Samples

All examinees who were administered the TOD-C or TOD-E were also administered the TOD-S. The TOD-S standardization sample consisted of 2,070 examinees and was broken down into child and adult subsamples for the development of scores. The TOD-S child sample consisted of all individuals from kindergarten through 12th grade who took either the TOD-C or the TOD-E. Of the 1,723 individuals in the TOD-S child sample, 337 had a high-incidence clinical diagnosis and/or a reading learning disability. (Most of these individuals had a reading disability; the percentage is consistent with the population.) Table 4.1 details the demographic characteristics of the TOD-S child standardization sample with regard to gender, race/ethnicity, parental education level, and region, along with corresponding percentages from the U.S. Census for comparison (Bernan Press & ProQuest, 2020). Most demographic categories closely match the proportions of the U.S. Census figures, exceeding the guideline that they be within 5% of the population at the time the normative data are collected (Andersson, 2005). Geographic region showed some variance; the South was slightly overrepresented while the Northeast was slightly underrepresented.

Table 4.1. Demographic Characteristics of the Standardization Sample: TOD-S Child

Characteristic	*n*	% of sample	U.S. Census %[a]
Gender			
Male	845	49.0	51.1
Female	877	50.9	48.9
Other	1	0.1	
Parents' educational level			
No high school diploma	146	8.5	11.5
High school graduate	477	27.7	26.1
Some college	485	28.2	30.3
Bachelor's degree or higher	615	35.7	32.2
Race/Ethnicity[b]			
Asian	83	4.8	4.7
Black/African American	247	14.3	13.6
White	865	50.2	52.1
American Indian/Alaska Native	25	1.5	0.7
Native Hawaiian/Pacific Islander	16	0.9	0.2
Other/Multiracial	57	3.3	4.6
Hispanic Origin	430	25.0	24.1
U.S. geographic region			
Northeast	177	10.3	16.3
Midwest	330	19.2	21.4
South	814	47.2	38.3
West	402	23.3	24.1

Note. N = 1,723. Due to rounding, total percentages may not equal 100.0%.

[a] Bernan Press & ProQuest (2020). Gender, race/ethnicity, and region are based on ages 5–18 years; parents' educational level is based on ages 25–64 years (those most likely to have children ages 5–18 years).

[b] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Tables 4.2 and 4.3 delineate the stratification of the child sample by age year and grade. The sample is most heavily concentrated at the younger ages/grades, which reflects the need for narrower normative age groups when development of skills measured by the TOD is most rapid, as well as the overlap of TOD-C and TOD-E in Grades 1 and 2.

The TOD-S adult sample consisted of 347 individuals, 64 of whom had a clinical diagnosis and/or reading disability. This same sample provided the basis for the TOD-C adult scores. Table 4.4 details the demographic characteristics of the TOD-S adult standardization sample with regard to gender, race/ethnicity, education level, and region, along with corresponding percentages from the U.S. Census for comparison (Bernan Press & ProQuest, 2020). This sample reflected a slight overrepresentation of the Midwest and underrepresentation of the Northeast; those lacking a high school diploma were slightly underrepresented. Table 4.5 presents the age groups within the sample, demonstrating good coverage across the age range of the test.

Table 4.2. Age Breakdown of the Standardization Sample: TOD-S Child

Age (years)	*n*
5	71
6	170
7	185
8	210
9	164
10	132
11	137
12	143
13	118
14	95
15	104
16	81
17	76
18	37

Note. N = 1,723.

Table 4.3. Grade Breakdown of the Standardization Sample: TOD-S Child

Grade	*n*
K	121
1	199
2	221
3	170
4	146
5	140
6	145
7	128
8	103
9	101
10	88
11	87
12	74

Note. N = 1,723.

Table 4.4. Demographic Characteristics of the Standardization Sample: TOD-S/TOD-C Adult

Characteristic	*n*	% of sample	U.S. Census %[a]
Gender			
Male	162	46.7	49.1
Female	183	52.7	50.9
Other	2	0.6	
Educational level			
No high school diploma	25	7.2	12.5
High school graduate	99	28.5	27.3
Some college	112	32.3	29.3
Bachelor's degree or higher	111	32.0	31.0
Race/Ethnicity[b]			
Asian	24	6.9	5.3
Black/African American	47	13.5	12.4
White	207	59.7	63.1
American Indian/Alaska Native	1	0.3	0.7
Native Hawaiian/Pacific Islander	1	0.3	0.2
Other/Multiracial	7	2.0	2.0
Hispanic Origin	60	17.3	16.3
U.S. geographic region			
Northeast	34	9.8	17.9
Midwest	92	26.5	21.3
South	137	39.5	37.4
West	84	24.2	23.4

Note. N = 347. Due to rounding, total percentages may not equal 100.0%.

[a] Bernan Press & ProQuest (2020). Demographic characteristics are based on the general adult population.

[b] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.5. Age Breakdown of the Standardization Sample: TOD-S/TOD-C Adult

Age (years)	*n*
18–23	113
24–39	64
40–49	40
50–59	54
60–69	37
70–89	39

Note. N = 347.

TOD-C Standardization Samples

Participants within the TOD-C sample were divided into two subsamples to support the school age versus adult use of the test. The resulting TOD-C samples consisted of 1,401 individuals in Grades 1 through 12 (referred to as the *child sample*, 272 of whom had a high-incidence clinical diagnosis and/or reading disability) and 347 individuals of post-high-school age, some of whom were in college (referred to as the *adult sample* and described in the previous section and Tables 4.4 and 4.5). Table 4.6 presents the demographic characteristics of the TOD-C child sample, along with corresponding percentages from the U.S. Census (Bernan Press & ProQuest, 2020). These closely match the U.S. Census figures, with a few exceptions. For example, within this sample the South was slightly overrepresented and the Northeast underrepresented.

Table 4.6. Demographic Characteristics of the Standardization Sample: TOD-C Child

Characteristic	*n*	% of sample	U.S. Census %[a]
Gender			
Male	687	49.0	51.1
Female	713	50.9	48.9
Other	1	0.1	
Parents' educational level			
No high school diploma	118	8.4	11.5
High school graduate	388	27.7	26.1
Some college	391	27.9	30.3
Bachelor's degree or higher	504	36.0	32.2
Race/Ethnicity[b]			
Asian	56	4.0	4.7
Black/African American	195	13.9	13.6
White	727	51.9	52.1
American Indian/Alaska Native	17	1.2	0.7
Native Hawaiian/Pacific Islander	13	0.9	0.2
Other/Multiracial	50	3.6	4.6
Hispanic Origin	343	24.5	24.1
U.S. geographic region			
Northeast	122	8.7	16.3
Midwest	253	18.1	21.4
South	699	49.9	38.3
West	327	23.3	24.1

Note. N = 1,401. Due to rounding, total percentages may not equal 100.0%.

[a] Bernan Press & ProQuest (2020). Gender, race/ethnicity, and region are based on ages 6–18 years; parents' educational level is based on ages 25–64 years (those most likely to have children ages 6–18 years).

[b] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Tables 4.7 and 4.8 delineate the stratification of the child normative sample by age year and grade in school. The sample is most heavily concentrated in the elementary and middle school years, when skills measured by the TOD develop more rapidly and students are most likely to be identified.

Subsamples of both the TOD-C child sample and the TOD-C adult sample completed one or more of the three TOD-C Rating Scales: Self-Rating, Teacher Rating, and Parent/Caregiver Rating. Tables 4.9 and 4.10 illustrate the demographic characteristics of the subsamples upon which the Rating Scale *T*-scores are based. The proportions of gender, race/ethnicity, parent education, and region resemble those found in the larger TOD-C samples. Two hundred and eighty-one participants were rated by individuals across all three scales: Self, Parent/Caregiver, and Teacher.

Table 4.7. Age Breakdown of the Standardization Sample: TOD-C Child

Age (years)	*n*
6	48
7	81
8	168
9	164
10	132
11	139
12	146
13	121
14	98
15	106
16	82
17	78
18	38

Note. N = 1,401.

Table 4.8. Grade Breakdown of the Standardization Sample: TOD-C Child

Grade	*n*
1	81
2	120
3	171
4	146
5	140
6	147
7	131
8	106
9	104
10	91
11	89
12	75

Note. N = 1,401.

Table 4.9. Demographic Characteristics of the Standardization Sample: TOD-C Child Rating Scale

Characteristic	*n*	% of sample
Gender		
Male	584	48.1
Female	630	51.9
Other	1	0.1
Parents' educational level		
No high school diploma	81	6.7
High school graduate	296	24.4
Some college	316	26.0
Bachelor's degree or higher	522	43.0
Race/Ethnicity[a]		
Asian	72	5.9
Black/African American	148	12.2
White	625	51.4
American Indian/Alaska Native	15	1.2
Native Hawaiian/Pacific Islander	6	0.5
Other/Multiracial	43	3.5
Hispanic Origin	306	25.2
U.S. geographic region		
Northeast	100	8.2
Midwest	258	21.2
South	585	48.2
West	272	22.4
Age (years)		
6	31	2.6
7	56	4.6
8	130	10.7
9	138	11.4
10	122	10.0
11	131	10.8
12	136	11.2
13	115	9.5
14	83	6.8
15	100	8.2
16	74	6.1
17	62	5.1
18	37	3.0
Grade		
1	50	4.1
2	92	7.6
3	142	11.7
4	129	10.6
5	114	9.4
6	143	11.8
7	127	10.5
8	96	7.9
9	96	7.9
10	82	6.8
11	74	6.1
12	70	5.8

Continued in next column

Note. N = 1,215. Parent/Caregiver Rating Scale *n* = 997; Teacher Rating Scale *n* = 448; Self-Rating Scale *n* = 1,066. Due to rounding, total percentages may not equal 100.0%.

[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.10. Demographic Characteristics of the Standardization Sample: TOD-C Adult Rating Scale

Characteristic	*n*	% of sample
Gender		
Male	118	44.2
Female	148	55.4
Other	1	0.1
Educational level		
No high school diploma	13	4.9
High school graduate	75	28.1
Some college	78	29.2
Bachelor's degree or higher	101	37.8
Race/Ethnicity[a]		
Asian	22	8.2
Black/African American	30	11.2
White	158	59.2
American Indian/Alaska Native	1	0.4
Native Hawaiian/Pacific Islander	1	0.4
Other/Multiracial	6	2.2
Hispanic Origin	49	18.4
U.S. geographic region		
Northeast	22	8.2
Midwest	80	30.0
South	116	43.4
West	49	18.4
Age (years)		
18–23	86	32.2
24–39	49	18.4
40–49	36	13.5
50–59	33	12.4
60–69	30	11.2
70–89	33	12.4

Note. N = 267. Due to rounding, total percentages may not equal 100.0%.

[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

TOD-E Standardization Sample

The TOD-E sample consisted of 342 individuals in kindergarten through second grade, 70 of whom had a clinical diagnosis (e.g., reading disability). Table 4.11 illustrates the demographic characteristics of this sample, along with corresponding percentages from the U.S. Census (Bernan Press & ProQuest, 2020). The TOD-E sample closely matched the U.S. Census with the exception of race/ethnicity; Whites were slightly underrepresented. Tables 4.12 and 4.13 delineate the stratification of the normative sample by age year and grade in school and illustrate good representation across the age range of the test.

A TOD-E subsample completed one or two of the TOD-E Rating Scales, the Teacher Rating and Parent/Caregiver Rating. Although a Self-Rating Scale was administered for the TOD-E sample, the results were unreliable, so the scale was dropped. Table 4.14 illustrates the demographic characteristics of the subsample upon which the Rating Scale *T*-scores are based. The proportions of gender, race/ethnicity, parent education, and region resemble those found in the larger TOD-E sample. Eighty-five participants were rated by both parent and teacher.

Table 4.11. Demographic Characteristics of the Standardization Sample: TOD-E

Characteristic	*n*	% of sample	U.S. Census %[a]
Gender			
Male	170	49.7	51.0
Female	172	50.3	49.0
Parents' educational level			
No high school diploma	30	8.8	11.5
High school graduate	92	26.9	26.1
Some college	103	30.1	30.3
Bachelor's degree or higher	117	34.2	32.2
Race/Ethnicity[b]			
Asian	30	8.8	4.8
Black/African American	54	15.8	13.5
White	145	42.4	50.2
American Indian/Alaska Native	8	2.3	0.7
Native Hawaiian/Pacific Islander	3	0.9	0.2
Other/Multiracial	8	2.3	5.2
Hispanic Origin	94	27.5	25.4
U.S. geographic region			
Northeast	56	16.4	15.8
Midwest	78	22.8	21.2
South	118	34.5	38.6
West	90	26.3	24.4

Note. N = 342. Due to rounding, total percentages may not equal 100.0%.

[a] Bernan Press & ProQuest (2020). Gender, race/ethnicity, and region are based on ages 5 years–9 years, 3 months; parents' educational level is based on ages 25–64 years (those most likely to have children ages 5 years–9 years, 3 months).

[b] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.12. Age Breakdown of the Standardization Sample: TOD-E

Age (years)[a]	*n*
5	72
6	122
7	104
8–9:3	44

Note. N = 342.

[a]8-year normative group extends through age 9 years, 3 months.

Table 4.13. Grade Breakdown of the Standardization Sample: TOD-E

Grade	*n*
K	122
1	118
2	102

Note. N = 342.

Table 4.14. Demographic Characteristics of the Standardization Sample: TOD-E Rating Scale

Characteristic	*n*	% of sample
Gender		
Male	103	48.8
Female	108	51.2
Parents' educational level		
No high school diploma	11	5.2
High school graduate	54	25.6
Some college	70	33.2
Bachelor's degree or higher	76	36.0
Race/Ethnicity[a]		
Asian	25	11.8
Black/African American	38	18.0
White	81	38.4
American Indian/Alaska Native	5	2.4
Native Hawaiian/Pacific Islander	3	1.4
Other/Multiracial	4	1.9
Hispanic Origin	55	26.1
U.S. geographic region		
Northeast	56	16.4
Midwest	78	22.8
South	118	34.5
West	90	26.3
Age (years)[b]		
5	37	17.5
6	54	25.6
7	81	38.4
8–9:3	39	18.5
Grade		
K	74	35.1
1	66	31.3
2	71	33.6

Note. N = 211. Parent/Caregiver Rating Scale *n* = 154; Teacher Rating Scale *n* = 142. Due to rounding, total percentages may not equal 100.0%.

[a]Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

[b]8-year normative group extends through age 9 years, 3 months.

Clinical Samples

As described earlier in this chapter, clinical cases were collected as part of the standardization data collection, and individuals with high-incidence diagnoses were included in the normative sample. These individuals are also part of the clinical sample, along with individuals who showed other clinical symptoms. To be included in the clinical sample, the examinee required a diagnosis of reading disability or other clinical diagnosis (e.g., autism, ADHD). Participation in special education was not a requirement, although most of the individuals in the clinical samples did receive some form of intervention services. Individuals who were visually impaired or deaf/hard of hearing all had interventions (e.g., special glasses, hearing aids) that allowed them to complete the TOD tests with the standardized administration procedures.

Individuals with multiple diagnoses were categorized by primary diagnosis; however, individuals with both a reading disability and another clinical diagnosis were included in the reading disability group as well as the group with their primary clinical diagnosis. Within the TOD-C clinical sample, a comparison was made of the Dyslexia Diagnostic Index (DDI) mean of two groups: individuals with only a reading disability diagnosis and those with a comorbid diagnosis. To compare these groups, the effect size was calculated as the difference between the mean standard scores of the two groups, divided by the pooled standard deviation. By this method, an effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 is considered large (Cohen, 1992). The effect size of the difference between the means was small. Thus, the individuals with reading disability have equivalent severity of impairment in variables of interest, regardless of the presence of additional conditions. Validity analyses based on the clinical samples are detailed in Chapter 5.

TOD-S Clinical Samples

Tables 4.15 and 4.16 show the demographic characteristics and diagnostic composition of the TOD-S child clinical sample. Because of the inclusion criteria for these samples, the demographics were not expected to replicate the U.S. Census demographic distribution. However, the samples offer a high level of diversity. Males outnumbered females, as is often the case in clinical samples. Tables 4.17 and 4.18 show the same information for the TOD-S/TOD-C adult clinical sample, illustrating a high level of diversity.

Table 4.15. Demographic Characteristics of the Clinical Sample: TOD-S Child

Characteristic	*n*	% of sample
Gender		
Male	370	58.6
Female	260	41.2
Other	1	0.2
Parents' educational level		
No high school diploma	79	12.5
High school graduate	160	25.4
Some college	135	21.4
Bachelor's degree or higher	257	40.7
Race/Ethnicity[a]		
Asian	48	7.6
Black/African American	109	17.3
White	298	47.2
American Indian/Alaska Native	10	1.6
Native Hawaiian/Pacific Islander	3	0.5
Other/Multiracial	19	3.0
Hispanic Origin	144	22.8
U.S. geographic region		
Northeast	99	15.7
Midwest	127	20.1
South	277	43.9
West	128	20.3
Age (years)		
5	33	5.2
6	54	8.6
7	62	9.8
8	93	14.7
9	60	9.5
10	69	10.9
11	66	10.5
12	51	8.1
13	57	9.0
14	27	4.3
15	24	3.8
16	13	2.1
17	14	2.2
18	8	1.3

Note. N = 631. Due to rounding, total percentages may not equal 100.0%.

[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.15 continued on next page

Table 4.15. Demographic Characteristics of the Clinical Sample: TOD-S Child *(continued)*

Characteristic	*n*	% of sample
Grade		
K	58	9.2
1	59	9.4
2	92	14.6
3	75	11.9
4	64	10.1
5	69	10.9
6	57	9.0
7	52	8.2
8	31	4.9
9	26	4.1
10	21	3.3
11	14	2.2
12	13	2.1

Note. N = 631. Due to rounding, total percentages may not equal 100.0%.

[a]Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.16. Diagnostic Composition of the Clinical Sample: TOD-S Child

Primary diagnosis	*n*	% of sample
Intellectual disability	21	4.6
Developmental delay	51	11.1
Autism spectrum disorder	47	10.3
Attention-deficit/hyperactivity disorder	128	28.0
Language disorder	57	12.5
Speech disorder	87	19.0
Emotional or behavioral disorder	19	4.2
Visually impaired	22	4.8
Deaf/Hard of hearing	11	2.4
Other health impairment	15	3.3
Reading disability	298	73.6

Note. N = 631. Due to rounding, total percentages may not equal 100.0%. One hundred twenty-seven of the individuals with a reading disability also reported another diagnosis.

Table 4.17. Demographic Characteristics of the Clinical Sample: TOD-S/TOD-C Adult

Characteristic	*n*	% of sample
Gender		
Male	27	38.0
Female	42	59.1
Other	2	2.8
Educational level		
No high school diploma	6	8.4
High school graduate	8	11.3
Some college	32	45.1
Bachelor's degree or higher	25	35.2
Race/Ethnicity[a]		
Asian	8	11.3
Black/African American	7	9.9
White	40	56.3
American Indian/Alaska Native	1	1.4
Native Hawaiian/Pacific Islander	15	21.1
Other/Multiracial	8	11.3
Hispanic Origin	7	9.9
U.S. geographic region		
Northeast	3	4.2
Midwest	9	12.7
South	43	60.6
West	16	22.5
Age (years)		
18–23	26	36.6
24–39	18	25.4
40–49	11	15.5
50–59	8	11.3
60–69	4	5.6
70–89	4	5.6

Note. N = 71. Due to rounding, total percentages may not equal 100.0%.

[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.18. Diagnostic Composition of the Clinical Sample: TOD-S/TOD-C Adult

Primary diagnosis	*n*	% of sample
Intellectual disability	5	8.2
Attention-deficit/hyperactivity disorder	22	36.1
Language disorder	3	4.9
Speech disorder	3	4.9
Emotional or behavioral disorder	4	6.6
Visually impaired	16	26.2
Deaf/Hard of hearing	7	11.5
Other health impairment	1	1.6
Reading disability	16	22.5

Note. N = 71. Due to rounding, total percentages may not equal 100.0%. Six of the individuals with a reading disability also reported another diagnosis.

TOD-C Clinical Samples

Tables 4.19 and 4.20 show the demographic characteristics and diagnostic composition of the TOD-C child clinical sample. The demographics show considerable diversity and a higher number of males than females. The TOD-C adult clinical sample is described in the previous section (Tables 4.17 and 4.18), as the TOD-S and TOD-C adult samples are identical.

Table 4.19. Demographic Characteristics of the Clinical Sample: TOD-C Child

Characteristic	*n*	% of sample
Gender		
Male	299	58.5
Female	211	41.3
Other	1	0.2
Parents' educational level		
No high school diploma	61	12.0
High school graduate	121	23.7
Some college	97	19.0
Bachelor's degree or higher	232	45.4
Race/Ethnicity[a]		
Asian	36	7.1
Black/African American	80	15.7
White	264	51.7
American Indian/Alaska Native	7	1.4
Native Hawaiian/Pacific Islander	2	0.4
Other/Multiracial	16	3.1
Hispanic Origin	106	20.7
U.S. geographic region		
Northeast	68	13.3
Midwest	95	18.6
South	226	44.2
West	122	23.9
Age (years)		
6	6	1.2
7	28	5.5
8	71	13.9
9	61	11.9
10	69	13.5
11	69	13.5
12	51	10.0
13	59	11.6
14	29	5.7
15	28	5.5
16	15	2.9
17	16	3.1
18	9	1.8
Grade		
1	19	3.7
2	53	10.4
3	76	14.9
4	64	12.5
5	70	13.7
6	59	11.6
7	52	10.2
8	33	6.5
9	30	5.9
10	25	4.9
11	15	2.9
12	15	2.9

Continued in next column

Note. N = 511. Due to rounding, total percentages may not equal 100.0%.
[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.19. Demographic Characteristics of the Clinical Sample: TOD-C Child *(continued)*

Note. N = 511. Due to rounding, total percentages may not equal 100.0%.
[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

Table 4.20. Diagnostic Composition of the Clinical Sample: TOD-C Child

Primary diagnosis	*n*	% of sample
Intellectual disability	21	6.0
Developmental delay	13	3.7
Autism spectrum disorder	49	14.0
Attention-deficit/hyperactivity disorder	118	33.8
Language disorder	33	9.5
Speech disorder	62	17.8
Emotional or behavioral disorder	17	4.9
Visually impaired	17	4.9
Deaf/Hard of hearing	10	2.9
Other health impairment	9	2.6
Reading disability	278	54.4

Note. N = 511. Due to rounding, total percentages may not equal 100.0%. One hundred sixteen of the individuals with a reading disability also reported another diagnosis.

TOD-E Clinical Sample

Tables 4.21 and 4.22 show the demographic characteristics and diagnostic composition of the TOD-E clinical sample. Similar to the TOD-C sample demographics, the TOD-E demographics show considerable diversity and a higher number of males than females.

Table 4.21. Demographic Characteristics of the Clinical Sample: TOD-E

Characteristic	*n*	% of sample
Gender		
Male	83	60.1
Female	55	39.9
Parents' educational level		
No high school diploma	19	13.8
High school graduate	40	29.0
Some college	41	29.7
Bachelor's degree or higher	38	27.5
Race/Ethnicity[a]		
Asian	16	11.6
Black/African American	29	21.0
White	44	31.9
American Indian/Alaska Native	3	2.2
Native Hawaiian/Pacific Islander	1	0.7
Other/Multiracial	6	4.4
Hispanic Origin	39	28.3
U.S. geographic region		
Northeast	32	23.2
Midwest	34	24.6
South	53	38.4
West	19	13.8
Age (years)[b]		
5	33	23.9
6	49	35.5
7	34	24.6
8–9:3	22	15.9
Grade		
K	59	42.8
1	40	29.0
2	39	28.3

Note. N = 138. Due to rounding, total percentages may not equal 100.0%.

[a] Individuals of Hispanic origin are included in the race/ethnicity category under Hispanic Origin; remaining categories include only individuals of non-Hispanic origin.

[b] 8-year clinical group extends through age 9 years, 3 months.

Table 4.22. Diagnostic Composition of the Clinical Sample: TOD-E

Primary diagnosis	*n*	% of sample
Intellectual disability	1	0.8
Developmental delay	38	30.9
Autism spectrum disorder	1	0.8
Attention-deficit/hyperactivity disorder	16	13.0
Language disorder	24	19.5
Speech disorder	29	23.6
Emotional or behavioral disorder	2	1.6
Visually impaired	5	4.1
Deaf/Hard of hearing	1	0.8
Other health impairment	6	4.9
Reading disability	31	22.5

Note. N = 138. Due to rounding, total percentages may not equal 100.0%. Eighteen of the individuals with a reading disability also reported another diagnosis.

Final Item Selection

For each untimed TOD test, the standardization study responses were analyzed using the Rasch one-parameter model, as detailed in the TOD Pilot Study section earlier in this chapter. Analyses were conducted using the software program jMetrik (Meyer, 2014). Based on the Rasch analysis, as well as conventional analyses (e.g., review of passing rates), examiner feedback, and raw score ranges for each test, a few items were deleted from many of the tests. This was done to eliminate items that did not fit well with the measurement model or that demonstrated bias, or whose elimination would not compromise precision of measurement. Final item order was determined using the Rasch estimate of item difficulty, such that items progressed from easiest to most difficult without any large measurement gaps between items. Rasch item difficulty and person ability analysis was used to ensure that each test contained items fully covering the intended age/grade range with a relatively consistent item gradient based on the increase in difficulty from one item to the next. Most timed tests retained all items, with a few exceptions based on examiner feedback. All items selected for final publication demonstrated no systematic bias by gender, race/ethnicity, or socioeconomic status (SES), based on item analyses *and* on an in-depth analysis of items by bias experts. Review of standardization data also indicated the need to combine the TOD-C Substitution and Deletion tests into a single test of Phonological Manipulation; when combined, the tests provided stronger measurement precision and better floor and ceiling data. In addition, Oral Reading Efficiency was removed from the TOD-E because it was too difficult for many of the examinees in kindergarten and for some in first grade. Final test order was determined by which tests loaded onto index scores and to ensure task type varied from one test to the next. A later section describes the determination of item sets for the Picture Vocabulary (1S) and Letter and Word Choice (2S) tests.

The Rating Scale scores correlated with the direct test scores at moderate-to-high levels, indicating a strong relationship between behavior ratings and direct assessment of skills. The exception to this pattern was the TOD-E Self-Rating form. The correlations between this Rating Scale score and the TOD-E results were low, indicating that children's perceptions of their skills differed from their actual performance. Thus, the TOD-E Self-Rating form was dropped from the TOD. Table 3.9 in Chapter 3, however, provides examples of questions that examiners can ask young children.

Establishing Basal and Stop Rules

In developing the final basal and stop rules for the TOD tests, the goal was to ensure a reliable, accurate, and efficient administration for each test and to have consistent rules across tests, if possible. Lengthier rules (five consecutive correct and five consecutive incorrect responses) were used for most tests during the standardization study to optimize data collection for item analysis by obtaining many item responses from each participant. However, such rules are impractical in a clinical setting, where the aim is to minimize the burden on the examiner and examinee by collecting only the amount of data required for accurate assessment. The aim was to find a single rule that could be applied across as many TOD tests as possible.

Final rules were determined by examining whether implementing a rule of four would make any appreciable difference in raw scores and standard score transformations. Analyses indicated that decreasing the basal and stop rules from five incorrect items to four had virtually no impact on scores; thus, the rule of four was applied to most TOD tests. Some tests of word reading and phonological awareness retained their standardization basal rules of only one or two items because the earlier items on those tests were of prerequisite skills (e.g., reading letters), rather than simply easier items of the same type.

Creation of Item Sets and Derivation of Ability Scores for Picture Vocabulary and Letter and Word Choice

Two tests, Picture Vocabulary (1S) and Letter and Word Choice (2S), needed to be treated differently than other TOD tests because the method of administration did not require the examiner to apply start and stop rules. For this reason, item sets were created for each of the three TOD-S forms (Grades K–1, 2–5, 6–Adult) with the goal that each examinee be presented with some items that were too easy, some that were near their level, and some that were too difficult. The determination of item sets was informed by person ability and Rasch item difficulty such that individuals would encounter both easy and difficult items within their grade-appropriate item set.

When a test uses item sets, the raw scores have meaning only within the context of the particular set of items from which they were obtained. Because examinees take different sets of items, their raw scores cannot be converted directly to standard scores. However, Rasch ability is scaled identically across all TOD-S item sets, and thus raw scores can be transformed into Rasch ability scores, which can in turn be converted into standard scores.

Rasch ability scores were derived iteratively from the Rasch difficulties of the items using the Newton-Raphson procedure (Wright & Stone, 1979). Appendix Tables A.1 and A.2 provide an ability score for the total raw scores representing failure on all items (raw score of 0) or success on all items in the item set. By definition, a score of 0 and a maximum score contain no variance; consequently, corresponding person abilities cannot be calculated directly using the Rasch model. The tabled values for these scores are thus extrapolated estimates. The differences between the ability scores for the nearest pair of adjacent total raw scores were used to extrapolate tabled values for the total raw scores of 0 and the maximum for each item set. For example, the Picture Vocabulary ability score of 30 on the Grade 6–Adult form is the sum of the ability score for 29 and the difference between the scores for the ability scores of 28 and 29. The ability score for 0 was derived in an analogous manner.

Derivation of Standard Scores

TOD-S, TOD-C, and TOD-E Test Scores

Age- and grade-stratified norms for the directly administered TOD tests were created using cNORM (Lenhard et al., 2018), a package for the R statistical platform. cNORM has two primary features:

1. It is a *continuous*, regression-based modeling process, which uses the variance of the entire normative sample to correct for age-specific distributional anomalies and sampling error.
2. It is a *distribution-free* method, meaning that the modeling process does not directly model age-group distribution parameters (e.g., mean, variance) and makes no assumptions about these parameters. As a result, cNORM can generate useful normative models, even when processing non-normally distributed input samples.

cNORM operates by modeling the *raw score* (*r*) as a function of person location (*l*, expressed as a percentile rank or other normative score) and the explanatory variable of age (*a*, expressed as a continuous age variable, or a discrete variable such as age-group membership or grade level). Age is "explanatory" in the sense that the latent ability for which *r* is an indicator increases with age, and that increase is presumably caused by the developmental changes that accompany the passage of time and increasing age. The functional relationship among these variables can be expressed as:

$$E(r) = f(l,a)$$

To create the raw-to-standard-score mapping required for clinical applications of the TOD tests, this functional relationship must be operationalized as a multiple regression equation. To determine the optimal regression equation, cNORM employs the mathematical methods of the Taylor series. Strictly speaking, the Taylor series is an infinite polynomial expansion, but for practical purposes, much of the variance in the functional relationship can be estimated with a finite expansion by reducing the polynomial to the degree *k*, which is 4 by default.

The Taylor series is thus simplified to a Taylor polynomial and consequently reduced to a model selection question for the following regression function:

$$r = \sum_{i,j=0}^{k} c_{i,j} \, l^i a^j$$

The predictors in the regression function include all powers of age and location and their linear combinations up to power *k*. cNORM selects the most relevant predictors and estimates all constants *c* to approximate *r* with the desired precision and as few predictors as possible. Usually, regression functions with five terms or less suffice to explain a large proportion of the variance in the normative sample (i.e., $R^2 \geq .95$).

The analysis proceeds in steps:

1. The normative sample is partitioned into roughly equal-sized age groups (or, alternatively, is grouped by grade levels).
2. Within these groups, each person is assigned a percentile rank as their value for *l*.
3. Powers of *a* and *l* are computed, up to the value of *k* (e.g., a^2, a^3, … , a^k). Products of these powers are computed (e.g., a^2l^2, a^2l^3, … , a^kl^k).
4. The powers and products are entered as predictors in a best-subset regression analysis, with *r* as the outcome variable.
5. The expansion of the Taylor function is determined by estimating regression coefficients of the most relevant predictors from the regression analysis. The most parsimonious regression function, meeting predefined fit criteria, is selected.
6. The regression equation resulting from the previous step is used to determine the normative score (e.g., IQ-type standard score) associated with each possible raw score on the test, either by directly computing the raw scores associated with a normative score at a specific age, or by determining the zero-crossings of the inverse function of the regression model to retrieve the normative score corresponding to a specific raw score.

cNORM has the additional advantage of allowing the specification of a post hoc age-stratification scheme (i.e., one that is independent of the age groups that were used in the modeling process). Because the modeled relationship between age, location, and raw score is a continuous function over chronological age, the raw-to-standard-score mapping can be generated at any point along the age continuum, theoretically with any level of precision, even down to a single day. This permits the test developer to impose a stratification scheme on the published raw-to-standard-score lookup tables (e.g., at three-month intervals within each age year) that best suits the intended clinical application of the test and reflects the progression of scores in the raw data.

Rating Scale Scores

Whereas standard scores with a mean of 100 are typically used for direct performance tests, *T*-scores with a mean of 50 typically are used for rating scales. These two types of scores are mathematically equivalent as they are both based on the normal distribution. The initial step of creating rating scale *T*-scores involved evaluating mean differences by age and grade. When no significant differences were found, it was determined that a single raw-to-*T*-score lookup table could be used for each rating scale sample. This process involved transforming the raw score distribution to approximate a normal distribution. The normalized raw scores were transformed into *z*-scores, which were then converted to *T*-scores, which have a mean of 50 and a standard deviation of 10. The use of normalized *T*-scores means that a given *T*-score value corresponds to the same percentile rank for all scales.

Derivation of Risk and Diagnostic Indexes

The TOD direct assessment batteries (TOD-S, TOD-C, and TOD-E) contain operationalizations of key measures of reading and spelling, as well as the linguistic processing abilities that underlie and predict them (i.e., phonological awareness, rapid automatized naming, orthographic processing, and auditory working memory), according to the literature (e.g., Bell et al., 2003; Kilpatrick, 2015; Mather & Jaffe, 2021; Mather & Wendling, in press; McCallum et al., 2006; Shaywitz & Shaywitz, 2020). This literature informed creation and development of all the TOD tests, described earlier in this chapter, as well as the various composites in the reading and spelling, linguistic processing, and vocabulary and reasoning domains. Three separate global index scores defining risk and/or diagnostic probability were created, one for each of the three TOD direct assessment batteries: the Dyslexia Risk Index (DRI) for the TOD-S; the Dyslexia Diagnostic Index (DDI) for the TOD-C; and the Early Dyslexia Diagnostic Index (EDDI) for the TOD-E. The DRI is a robust predictor of risk, and the DDI and EDDI are the best scores to predict a diagnosis of dyslexia. In the process of creating the DDI and EDDI, two additional indexes were derived for each assessment: the Reading and Spelling Index (RSI) and Linguistic Processing Index (LPI) for the TOD-C; and the Early Reading and Spelling Index (ERSI) and Early Linguistic Processing Index (ELPI) for the TOD-E, as described below.

Research to Determine Composition of Risk and Diagnostic Indexes

TOD-S

The TOD-S Dyslexia Risk Index (DRI) was created by averaging the tests assessing letter/word identification/spelling (Letter and Word Choice [2S]) and fluency (either Word Reading Fluency [3Sa] or Question Reading Fluency [3Sb], depending on the grade level of the examinee). The skills measured by these two TOD-S tests are considered appropriate as screening tests because they assess key literacy skills.

TOD-C and TOD-E

Development of the diagnostic indexes for both the TOD-C and TOD-E (DDI and EDDI, respectively) required using a stepwise process employing multiple and logistic regression analyses. The ultimate goal of the process was to create global scores that could predict significantly the probability of dyslexia. To accomplish this goal, specific reading, spelling, and linguistic processing tests in each battery were chosen and entered into multiple regression analyses, as indicated below.

The operationalization of the TOD-C DDI began by selecting the reading and spelling tests affected most by dyslexia (e.g., those with the lowest means within the TOD reading disability clinical sample). Then, the average of various combinations of these tests became the criterion variables for a series of multiple regression analyses. A subgoal was to maximize efficiency for examiners by limiting the number of criterion tests to four, two of which come from the TOD-S and are highly reliable, and to select tests based on theory and research, representing various reading and spelling skills (e.g., phonics, orthography, fluency, timed and untimed, decoding and encoding).

Ultimately, the four reading and spelling tests operationalizing the criterion variable that yielded the most powerful equation were Letter and Word Choice (2S); Word Reading Fluency (3Sa) or Question Reading Fluency (3Sb), depending on the age of the examinee; Irregular Word Spelling (5C); and Pseudoword Reading (7C). The four linguistic processing tests identified as the best predictors were Phonological Manipulation (4C), Rapid Letter Naming (6C), Word Pattern Choice (8C), and Word Memory (9C).

A similar process was followed in the development of the TOD-E EDDI. The TOD-E contains a total of five reading and spelling tests and three linguistic processing tests. Multiple regression analyses conducted by using 1) all three TOD-E early linguistic processing tests as predictors and 2) the average of the five TOD-E early reading/spelling skills as the criterion produced a strong predictive equation. The three TOD-E linguistic processing tests are Rhyming (5E), Early Rapid Number and Letter Naming (6E), and Early Segmenting (8E). The five reading and spelling tests selected are Letter and Word Choice (2S); Word Reading Fluency (3Sa) or Question Reading Fluency (3Sb), depending on the age of the examinee; Sounds and Pseudowords (4E); Letter and Sight Word Recognition (7E); and Letter and Sound Knowledge (9E).

Reading and Spelling Indexes and Linguistic Processing Indexes

The research completed to create the TOD-C Dyslexia Diagnostic Index (DDI) and the TOD-E Early Dyslexia Diagnostic Index (EDDI) led to the development of two additional indexes for each battery.

The tests of reading and spelling that produced the strongest prediction equation as described above were combined into the Reading and Spelling Index (RSI) for the TOD-C and the Early Reading and Spelling Index (ERSI) for the TOD-E.

The tests of linguistic processing shown to produce the strongest prediction of reading and spelling ability were combined into the Linguistic Processing

Index (LPI) for the TOD-C and the Early Linguistic Processing Index (ELPI) for the TOD-E.

Creation of Risk, Diagnostic, Reading and Spelling, and Linguistic Processing Index Scores

The final step in creating the indexes for the TOD-C and TOD-E required combining the test scores from the research steps described previously. Standard scores for the indexes were calculated by summing the standard scores from the eight tests that make up the most powerful DDI and EDDI regression equations. Then the index scores were created by calculating *z*-scores, which were transformed into standard scores with a mean of 100 and standard deviation of 15. Chapter 3 provides guidance for interpretation of the indexes. Because age- and grade-related differences were accounted for in the individual test scores, only one lookup table was required to convert the sum of standard scores into the index scores for each reference sample.

Once the set of tests used for risk and diagnostic scores was finalized, logistic regression analyses were used to confirm that the indexes successfully predicted membership into either a group of examinees previously identified as having a learning disability in basic reading, or a second group from a demographically matched sample of examinees from the standardization sample.

Derivation of Composite Scores

In addition to the dyslexia risk and diagnostic indexes, composite scores were derived for many of the skills measured by the TOD. The composites contribute to the broad application of the TOD and its flexibility to answer different referral questions. Each composite score was created by adding the standard scores for the tests they comprised, calculating *z*-scores, and then transforming the *z*-scores into standard scores with a mean of 100 and standard deviation of 15. As with the index scores, age- and grade-related differences are already accounted for by the test standard scores and, therefore, a single lookup table was created for each reference sample for the conversion of sums of standard scores into composite scores.

In almost all cases, age-based test scores are combined for age-based composites and grade-based test scores are combined for grade-based composites. However, for the Reading Fluency composite in the TOD-C child sample, the score is obtained by combining the age-based Word Reading Fluency (WRF; 3Sa) or Question Reading Fluency (QRF; 3Sb) score with the grade-based Oral Reading Efficiency (ORE; 12C) score. Since the ORE provides only grade-based scores for the child age group, its combination with age-based WRF or QRF was evaluated. Comparing the age- and grade-based versions of the test scores revealed that differences were almost all less than the *SEM* for the tests. This suggests that any error that results from combining these age- and grade-based scores to create the Reading Fluency composite could be considered test error and is not likely to affect clinical interpretation. This approach is further supported by analysis of age- and grade-based composite score differences, which demonstrated only small effect sizes.

Derivation of Growth Scores

The TOD tests were divided into two groups for deriving growth scores: timed tests and untimed tests. Because untimed tests are measuring one construct (the skill in question) while timed tests are measuring two constructs (the skill in question plus speed), it was necessary to treat these types of tests differently when analyzing the data.

For both types of tests, Winsteps (version 5.2.4.0) was used to generate estimates of item difficulty values and person ability values consistent with Rasch item modeling (Linacre, 2023). For untimed tests, dichotomous data could be input directly into Winsteps; for timed tests, however, items were put into item sets, generally of five consecutive items each. Each item set then received a score that was the sum of the individual item scores, and Winsteps was run for these polytomized data. After this point, all procedures applied identically to timed and untimed tests.

An initial Winsteps run was conducted to generate item difficulty parameter estimates and person abilities. These person abilities were then used to determine a “centering constant” for future runs, whose purpose was to provide comparability between tests. Based on these results, the growth scores are defined as follows:

$$\text{Growth score} = 500 + C_T + 9.1024$$

where 500 is added to each score to eliminate negative values, and C_T is a centering constant C unique to test T. In the TOD, C_T is chosen so that, for a group of examinees centered around age 90 months, their median growth score is 475, and 9.1024 is the scaling constant.

Additional Winsteps runs applied the centering constant and generated item difficulty estimates from the standardization sample, and then generated a set of person ability estimates for all cases, both standardization and clinical.

A final Winsteps run used the established item difficulty values along with dummy response data to generate scoring tables. Dummy data were used to ensure that the full possible range of raw scores was represented in the table. Additionally, the dummy data file allowed for different combinations of items to generate scoring tables based on different grade starting points (e.g., for Picture Vocabulary [1S]).

Growth scores were derived for all TOD tests except for Oral Reading Efficiency (12C). In the case of Phonological Manipulation (4C), separate growth score tables were derived for each subtest, Substitution and Deletion.

Derivation of Age and Grade Equivalent Scores

An age equivalent represents the age, in years and months, at which a particular raw score is the average score. A grade equivalent represents the grade placement, in grade and term, at which a particular raw score is the average score. The age-based norms for each test were used to develop the age equivalents, and the grade-based norms were used to develop the grade equivalents. These scores were developed by determining the raw score that corresponded to a standard score of 100 for each of the age and grade groups, then linking it to the midpoint for that age or grade group and interpolating missing scores when needed.

Equivalency Studies

Several equivalency studies were conducted to ensure that the TOD is appropriate for use under different circumstances and by different groups of individuals. To maximize the power of the analyses, the child and adult samples were analyzed together. All equivalency studies were conducted by comparing the group of interest with another group of individuals from the standardization sample matched based on age, clinical status, gender, parent education, and race/ethnicity.

English Language Status

The TOD requires proficiency in English, and thus it was important to determine the extent to which the scores of monolingual English speakers and non-native, bilingual, or multilingual English speakers might differ on TOD tests.

Across all three samples (TOD-S n = 211; TOD-C n = 184; TOD-E n = 31), when non-native English speakers, bilingual, or multilingual individuals were compared with a matched group, effect sizes ranged from 0.01 to 0.37, with a median of 0.18. As noted previously, effect sizes of 0.2 are considered small, 0.5 are considered medium, and 0.8 are considered large (Cohen, 1992). Across all 34 tests, mean differences were nonsystematic. That is, on some tests the monolingual English group scored higher, and on other tests, the non-native/bilingual/multilingual group scored higher. These results indicate that the scores of non-native, bilingual, or multilingual English speakers are sufficiently equivalent to those of matched controls. Thus, it is valid to use the TOD with non-native, bilingual, or multilingual English speakers, as well as to include them in the standardization sample.

Personal Protective Equipment During Covid-19 Pandemic

As previously mentioned, the Covid-19 pandemic occurred while TOD data collection was in progress, and thus some cases were collected with the use of personal protective equipment (PPE), primarily masks. Because wearing masks has the potential to impact both expressive and receptive language, and because the situation of the pandemic could impact testing behavior, it was important to compare the group of individuals assessed during the pandemic with a group matched on demographic variables who were tested pre-pandemic.

Across all samples (TOD-S n = 149; TOD-C n = 120; TOD-E n = 29), mean difference effect sizes ranged from 0.03 to 0.57, with a median of 0.24. Almost all effect sizes were small; however, in the few cases where moderate effect sizes were observed, there were no systematic differences between groups in terms of which group had a higher mean, nor was there any type of test most likely to evidence higher differences. This suggests that differences were likely due to other idiosyncratic reasons, and therefore it was valid to administer the TOD using masks during the pandemic. Thus, the data collected with masks during the pandemic are considered valid.

TOD-S Digital Administration

The TOD-S was standardized using a print (paper and pencil) administration; however, a sample of 368 individuals took the TOD-S digitally on a computer as well. This sample ranged in age from 5 to 77 years and was 43% male and 57% female. Twenty-five percent of the individuals came from families who did not attend any college. The sample was 16% Hispanic, 6% Asian, 22% Black, 47% White, and 9% Other.

The study was counterbalanced so that half of the participants were administered the TOD-S on paper first and half on a computer first. The two versions were administered one immediately after the other. Effect size results from mean-difference comparisons for the four TOD-S tests ranged from 0.00 to 0.06, all small. These results support the use of norms from the print-based administration for the digitally administered TOD-S and confirm that the TOD-S can be validly administered digitally as well.