WHITE PAPER

# mCLASS Reading 3D Technical Manual

Amplify.

# Table of Contents

# Introduction

## Overview

The Amplify Oral Language Screener (OL) efficiently identifies students in Grades K–2[6] who may struggle with comprehending the language structures that are foundational to comprehending early reader texts, understanding interpersonal communication, and writing. During the assessment, sentences are read to the student, and the student is instructed to repeat each sentence orally; this is a reliable and valid method for assessing oral language development (Clay, Gill, Glynn, McNaughton, & Salmon, 2007; Romeo, Gentile & Bernhardt, 2008; Menyuk, 1969). When students hear a sentence and accurately repeat it, they use the oral language structures they control, meaning that they understand what is said, even though they may not produce these structures in spontaneous speech (Clay, Gill, Glynn, McNaughton, & Salmon, 1983). Educators use OL for benchmarking or screening to identify students at risk for oral language comprehension difficulties in syntax at three benchmark administration periods during the school year. The data resulting from OL is used to help educators make district-, school-, and student-level instructional decisions.

## Theory

Reading comprehension is the product of both decoding skill and language comprehension (Gough & Tunmer, 1986; Catts, 2009; Catts, Adlof, & Weismer, 2006). Listening comprehension is a precursor to reading comprehension, and understanding and developing a student's knowledge and skill in language in tandem with decoding is critical to ensuring reading success. Language consists of five domains, and knowledge of all five domains is implicated in learning to read: (a) knowledge of phonology, or the sounds of language, allows for the connection of sounds with letters; (b) knowledge of morphology, or the structure of words, aids in

[6] Use of the assessment is permitted through Grade 5, however, research has only been conducted to support the assessment and provide cut points in kindergarten through Grade 2 due to the nature of the task. Results for older students should be interpreted cautiously because these students are more likely than younger students to simply mimic a sentence that they do not comprehend. Also, older English Language Learner (ELL) students may have developed a range of language structures in their primary language but need practice with word order and development of vocabulary in their second language (Clay, Gill, Glynn, McNaughton, & Salmon, 2007; Romeo, Gentile & Bernhardt, 2007). Consider the student's age, first language proficiency, years in English instruction, and percentage of instructional time in English when interpreting results.

decoding and comprehending complex words as well as expanding vocabulary; (c) knowledge of syntax, or the structure of sentences, helps with understanding how word order affects meaning; (d) knowledge of semantics, or the meaning of words and phrases, helps with word- and phrase-level comprehension; and (e) knowledge of pragmatics, or the meaning of language in context, provides a basis for a more nuanced and specific understanding of text. Knowledge of these domains can be further subdivided into receptive and expressive aspects: receptive language pertains to comprehension, while expressive language pertains to production.
Listening and reading are receptive processes, while speaking and writing are productive processes.

The OL measure targets receptive oral syntax, or a student's understanding of the grammatical structures underlying spoken sentences. In English, word order encodes meaning. For instance, English subjects tend to come before verbs, which tend to come before objects: the sentence "John gave the book to Mary" means something different than "Mary gave the book to John" because the

word order is different. The rules determining the meaning of sentences based on word order constitute syntax. Knowledge of syntactic rules, which is typically implicit in the absence of formal instruction in grammar, is essential to comprehension of both spoken and written language.

The purpose of the OL screener is to gauge whether a student's receptive oral syntactic knowledge is sufficient for development of early reading comprehension. In other words, the OL screener helps educators identify students who struggle with the language structures commonly found in early reader texts. OL can be used in the context of other reading and language assessments to determine instructional next steps in both decoding and language to support reading comprehension development. Poor performance on OL alerts educators that further language support is necessary to support reading comprehension; educators may conduct item (i.e., sentence) analysis of OL or administer additional language assessments to determine the targets of instruction. A Scoring Guide[7] to support item-level error pattern analysis is available to facilitate educators' identification of instructional next steps based on student OL performance.

The OL measure assesses receptive syntax using a sentence repetition task: the assessor reads a sentence out loud, and the student is asked to repeat the sentence

[7] The OL Scoring Guide can be found on the Support and Resource Center on mclasshome.com.

verbatim as the assessor notes any errors made. While sentence repetition may appear to be a simple task of mimicking, underlying the ability to repeat sentences is syntactic processing, which entails syntactic comprehension: while very short utterances may be parroted back phonetically without attention to meaning, longer utterances that exceed the limits of working memory must be processed, and therefore comprehended, prior to repetition (Clay, 1971). The successful repetition of an orally presented sentence implies comprehension of the sentence (Van Moere, 2012; Menyuk, 1969). However, the relationship between repetition and comprehension is attenuated with age, as students over the age of approximately 7 years are more capable than younger students of repeating sentences that they do not comprehend (Menyuk, 1969; Clay, 1971); therefore, OL score interpretations have only been provided for students in kindergarten through Grade 2.

# The Amplify Oral Language Screener

## Design

The OL assessment consists of three grade-independent forms to be administered at the beginning-, middle-, and end-of-year benchmark periods. The forms were designed to be of equivalent difficulty across benchmarking periods in order to facilitate growth monitoring. The three OL forms contain 21 sentences each. The 21 sentences are divided into seven categories (A–G) for assessing students' oral language development in Grades K–2 (Clay, Gill, Glynn, McNaughton, & Salmon, 2007; Gentile, 2004). These seven categories include grammatical structures found in everyday speech and early reader texts. Overall, the categories represent a gradient of difficulty; however, some structures are acquired before others in a student's oral language depending on what was heard in early conversation as well as on the student's native language. Each category contains three sentences at two different levels of complexity: there are two Level 1 sentences and one Level 2 sentence. Level 2 sentences are longer than Level 1 sentences and contain similar but more complex structures. If a student can repeat all of the Level 2 sentences with expression and without error, then they likely have the oral language skill necessary to support reading development. The seven categories of sentence structures for Level 1 are shown in Table 1, with examples.

**Table 1. Category A–G sentence structure**

| Category | Sentence Example | Level 1 Structure |
|---|---|---|
| A | My sister's dress is yellow. | Subject + Verb "to be" + Adjective |
| B | Mary is eating a sandwich. | Subject + Verb + Direct Object |
| C | Linda is playing after school. | Subject + Verb + Prepositional Phrase |
| D | Mom is buying me a book. | Subject + Verb + Indirect Object + Direct Object |
| E | I think he's at home. | Subject + Verb + Noun Clause |
| F | Here's another teddy bear. | Adverb or Relative Pronoun + Verb + Subject |
| G | He is not using his computer today. | Subject + Verb + Negation + Direct Object + Adverb |

In addition to the grammatical specifications for the sentences found in Table 1, sentences were constructed with attention to the characteristics listed in Table 2, based on Marie Clay and colleagues' Record of Oral Language (1971; 2007) as well as on lessons learned from the Pilot study (Study A) in which early content was administered to students.

**Table 2. Additional content specifications for OL**

| Specification | Description |
|---|---|
| Word frequency | Sentences contain high-frequency, tier one vocabulary words, or words that typically do not require instruction for students to learn the meanings (Beck, McKeown, & Kucan, 2013), and common English names to avoid confounding syntactic knowledge with semantic knowledge. |
| Phonetic clarity | Phoneme combinations that impede clarity for either the student or the assessor are avoided (e.g., combinations such as "Sam's supper" were avoided due to the consecutive word-final and word-initial /s/ sounds). |
| Syntactic clarity | Sentences with ambiguous syntactic constructions, such as garden-path sentences, which mislead the listener into an incorrect and confusing initial parse, were avoided. An example of a garden-path sentence is "The horse raced past the barn fell" in which the reader at first interprets "raced" as an active verb in the sentence and then realizes that "raced" modifies "horse". The sentence actually means "The horse, who was raced (i.e., made to run) past the barn, fell" and not that the barn fell. These types of sentences can cause confusion and disrupt comprehension. |
| Negation | Each OL form contains one negated sentence to assess student knowledge of this syntactic construction. |
| Indefinite articles | Given the common substitution of a for an even in students without syntactic comprehension difficulties, sentences were constructed to exclude an. All indefinite articles in OL are a and precede a consonant-initial word. |
| Sensitivity | Sentences are free from gender, cultural, or other types of bias. |
| Relevance | Sentence content pertains to everyday situations to promote semantic and pragmatic comprehensibility. |
| Memory burden | To reduce the need to remember sentence elements extraneous to the measurement of syntactic comprehension, pronouns and articles were used consistently within a sentence (e.g., rather than using his, the, and her all in the same sentence, the was used each time). |

## Administration and scoring procedures

OL is administered one-on-one. It is recommended that the assessor sit in a quiet location with the student in order to facilitate mutual comprehension. The assessor explains that the student should repeat exactly what the assessor says and begins by administering one or two practice items to ensure the student understands the task before scoring commences. The 21 items are administered one at a time as the assessor reads them off of a mobile device or a computer. Italics have been added to the written sentences to guide phrasing, facilitating standardized administration. The assessor marks sentences repeated verbatim as correct and categorizes errors as substitutions, omissions, insertions, or repetitions. Credit is given only for entire sentences that are repeated verbatim (self-corrections are allowed and are noted in the software); sentences with any number of errors are considered incorrect. Given the brevity of the measure and the richness of the data gleaned even from incorrectly repeated sentences, there is no discontinue rule for OL.

## Interpreting results

Performance on OL results in a raw score out of 21 and a performance interpretation pertaining to risk for language difficulty. Three performance levels result from OL: At or Above Benchmark, Below Benchmark, and Well Below Benchmark. Students performing Below or Well Below Benchmark may need additional assessment and/or instruction in syntactic comprehension.

While OL is a screener rather than a diagnostic measure, conducting an error pattern analysis, as described in the OL Scoring Guide, can facilitate deeper understanding of the syntactic constructions that a student controls (i.e., structures that the student can understand, even if the student cannot yet produce the structures in spontaneous speech) and those that have not yet been acquired. The OL Scoring Guide suggests instructional next steps for educators to implement within the classroom, based on error patterns commonly found among students in kindergarten through Grade 2. The OL Scoring Guide also includes information about the relationship between oral language and reading development, the relationship between oral language and writing development, the research and rationale for using sentence imitation in OL, the types of sentences in OL, and instructional recommendations based on error patterns in OL.

# Overview of Research on the Amplify Oral Language Screener

This chapter describes data and analyses from completed research studies, including the purpose of each study, how participants were recruited, demographics for the participants, the experimental designs, and the descriptive statistics calculated for each study.

## Study A: Pilot study

**Purpose:** The pilot study was designed to examine the feasibility and appropriateness of the instructions, item content, item order, and scoring procedures of OL. Four forms of OL were administered during the pilot study with the intention of narrowing down the content to three final forms based on research results from Studies A–C. Results of the pilot study were used to guide modifications to the assessment procedures and content prior to software implementation and prior to the field studies (Studies B and C).

**Recruitment:** The pilot study was conducted at one of Amplify's "lab schools" (in exchange for product discounts, Amplify employees may periodically visit to observe students and teachers using Amplify products and obtain feedback from users). This school was recruited via the school's account manager. This was the only school contacted, and it was selected based on its convenient location and availability of students who are English Language Learners (ELLs) as well as those who need support in learning some of the vocabulary and sentence structures that are commonly used in instruction, also known as Academic English. Because this measure focuses on oral language development, it was important to pilot the measure with students from a variety of language backgrounds.

**Participants:** The pilot study was conducted during the 2014–2015 beginning-of-year benchmark administration period. In total, 20 students in kindergarten through Grade 2 were assessed by two Amplify researchers (kindergarten: n = 6; Grade 1: n = 6; Grade 2: n = 8).

**Demographic information:** The student sample was 40 percent male and 60 percent female. The sample included both ELLs and Native English speakers. Among the ELLs, home languages included Spanish, Twi (a Niger-Congo language spoken mainly in Ghana), Funali (a Niger-Congo language spoken in Western and Central

Africa), and Garifuna (an Arawakan language spoken in Central America).[6] Several of the students had also been identified by the school as in need of additional instruction in Academic English.

**Study design:** Two Amplify researchers tested students individually in the library of the school. Both testers scored each student's performance; testers alternated who was the primary test administrator and who shadow scored student performance. The four forms of OL were made into eight forms (1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B) by varying the order of the items such that the "A" forms contained items in the order Level 1, Level 2, Level 1 for each alphabetic sentence complexity level, while the "B" forms contained items in the order Level 1, Level 1, Level 2 for each alphabetic sentence complexity level. Students were administered between two and four forms of OL, based on stamina.

OL was administered using paper-and-pencil forms that mimicked the software layout, as software development was still in progress at the time. The testers followed the typical, standardized administration and scoring procedures for OL; in addition, testers asked students questions about the task and the sentences included to gain more information about the appropriateness of the items (e.g., whether a student understood the meaning of a particular word or sentence) and administration and scoring procedures.

**Descriptive statistics:** OL data were reviewed both qualitatively and quantitatively, focusing on five areas: item appropriateness and difficulty, appropriateness of administration and scoring procedures, rater and form equivalence, patterns of overall student performance, and patterns of individual student performance. Some of the content specifications in Table 2 are based on lessons learned from Study A, which guided revisions to the assessment content prior to the field studies (Studies B and C). Descriptive information regarding the pilot study results is presented in Table 3, including the number of students administered each form, the mean score out of 21 on each form, and the standard deviation. Final scores tended to cluster around 15 out of 21 points.

[6] The National Center for Education Statistics (NCES; Seastrom, 2010) suggests suppression of subgroup information and underlying details in the case of small samples in order to protect personally identifiable information; thus, included subgroups will be named without providing demographic counts for Study A.

Table 3. Study A descriptive statistics

| Form | N | Mean | Standard Deviation |
|---|---|---|---|
| 1A | 9 | 15.55 | 0.17 |
| 1B | 7 | 16.50 | 0.10 |
| 2A | 8 | 17.79 | 0.11 |
| 2B | 6 | 14.23 | 0.85 |
| 3A | 9 | 14.81 | 0.44 |
| 3B | 7 | 14.57 | 0.61 |
| 4A | 6 | 13.96 | 0.66 |
| 4B | 7 | 16.00 | 0.20 |
| Overall | 59 | 15.43 | 0.39 |

## Study B: Initial field study

**Purpose:** Study B was designed to collect data necessary to calculate item statistics for OL, which were used to evaluate the need for revisions to the assessment. OL data from Study B were also used to calculate internal consistency reliability and criterion-related validity and set cut points for proficiency. Survey data from Study B were used to inform validity and revisions to the measure.

**Participants:** This field study was conducted during the ’14–’15 beginning- (BOY; August– September), middle- (MOY; January–February), and end-of-year (EOY; May) benchmark administration periods. In total, 2081 students in kindergarten through Grade 2 (kindergarten: n = 715; Grade 1: n = 640; Grade 2: n = 708) in eight schools (referred to as schools A–H) were assessed in one large, urban school district.

**Demographics:** Participants in this field study included students in Grades K–2 and testers who were educators in the schools where the study took place, Amplify researchers, and undergraduate research assistants and their professor. The student sample, described in Table 4, was composed of participants from the following demographic categories: 49.93 percent male, 47.28 percent female, and 2.78 percent unspecified gender; 2.07 percent White, 17.97 percent Black, 74.39 percent Hispanic, and 5.57 percent other or unspecified race. The student sample was 48.38 percent native English speaking, while the remaining 49.68 percent spoke English as a second language (language proficiency was not available for 1.94 percent of the sample); 6.45 percent were students receiving special education services.

Table 4. Study B participants and demographics

| | Kindergarten | Grade 1 | Grade 2 |
|---|---|---|---|
| Sample size (n) | | | |
| Students | 715 | 640 | 708 |
| Gender (n) | | | |
| Male | 359 | 312 | 368 |
| Female | 347 | 311 | 326 |
| Ethnicity (n) | | | |
| White | 21 | 12 | 10 |
| Hispanic | 532 | 474 | 542 |
| Black | 131 | 125 | 118 |
| Native American | 3 | 1 | 1 |
| Asian | 12 | 11 | 18 |
| Language status (n) | | | |
| Native speaker | 347 | 281 | 370 |
| English as a Second Language | 359 | 342 | 324 |
| Other demographics (n) | | | |
| Special education | 52 | 40 | 41 |

**Study Design:** Primary data collection to explore the psychometric functioning of OL was conducted by classroom teachers electronically via handheld device from 2063 K–2 students in Schools A–F during each benchmark administration period (i.e., Form 1 at BOY, Form 2 at MOY, Form 3 EOY).

Initial web-based training was provided to district leaders by members of the Amplify research team involved in the design of the assessment within a train-the-trainer model in which district leaders were expected to disseminate training information to classroom teachers. Topics covered in training included the purpose of OL, sentence types included in OL, OL administration procedures, OL scoring procedures, using the OL software, practice administering OL, interpreting OL results, and using OL results to guide instruction.

Student performance data collected at BOY, MOY, and EOY were used to examine difficulty, discrimination, and reliability of the items and test forms to ensure appropriate psychometric properties are demonstrated. Note that the software included an optional discontinue rule, which was triggered after a student scored incorrect on the first five items. At this point, the test administrator had the option to discontinue the assessment, and the student received a score of 0.

Around the EOY administration period (May 2015), a subset of students from schools D, E, and F that were administered OL at BOY, MOY, and EOY participated in additional testing to support assessment validity and establish cut points and performance levels. At schools D, E, and F, 551 students (kindergarten: n = 186: Grade 1: n = 170; Grade 2: n = 195) were administered the TOLD- P:4 Sentence Imitation subtest (Test of Language Development – Primary, Fourth Edition; Newcomer & Hammill, 2008)[6] by Amplify researchers as well as undergraduate research assistants and their professor. Students were selected for this analysis based on teacher perception of language proficiency such that a range of abilities was represented.

The TOLD-P:4 is a well-established test battery of oral language development that contains nine subtests covering syntax, semantics, and phonology within the speaking, listening, and organizing domains. The TOLD-P:4's Sentence Imitation (SI) subtest is a sentence imitation measure similar to OL that consists of 36 items of varying syntactic complexity. Credit is received only for sentences that are repeated verbatim. The TOLD-P:4 SI also has a discontinue rule: students who score incorrect on five sentences in a row are discontinued, and their score up to that point is entered. Final raw scores can be converted to age equivalents, percentile ranks, and scaled scores with seven associated descriptive terms ranging from Very Poor to Very Superior, based on the student's age. The psychometric properties of the TOLD-P:4 SI demonstrate strong internal consistency reliability (Cronbach's alpha = 0.93), test-retest reliability (r = 0.87), and inter-rater reliability (r = 0.99) across age groups. The TOLD-P:4 SI was validated against the Pragmatic Language Observation Scale, with a correlation of r = 0.78.

A paper-and-pencil feedback survey was administered at EOY to all educators who administered OL throughout the '14–'15 school year; participation in the survey was voluntary. This survey took approximately 10 minutes to complete and gathered reactions to the measure overall, its administration guidelines, its content, and its accompanying performance description, as well as participant demographics. These data were used to refine assessment content for the 2015–2016 school year.

[6] The original TOLD-P:4 SI sample included 625 students, but 74 students were removed from the analysis due to improper use of the discontinue rule. The assessment should be discontinued after a student makes errors on five items in a row, and the score up to that point is entered. Testers who improperly used the discontinue rule stopped assessing students when any five items were incorrect, rather than five consecutive items.

**Descriptive statistics:** Descriptive information regarding the OL results for Study B is presented in Table 5 by grade and form. Assessments in which the discontinue rule was triggered were removed from the analyses because in discontinued assessments, we cannot be sure of how a student would have performed past the fifth item; thus, 122 assessments were excluded in kindergarten, 32 in Grade 1, and 9 in Grade 2. Results from the remaining sample show an increase in scores and a decrease in standard deviation as grade increases, indicating that older students consistently perform better on this measure than younger students.

Table 5. Study B descriptive statistics

| Grade | Form | Sample Size (n) | Average score | Minimum score | Maximum score | Standard Deviation |
|---|---|---|---|---|---|---|
| Kindergarten | 1 | 715 | 14.00 | 1 | 21 | 4.57 |
| | 2 | 715 | 15.36 | 1 | 21 | 4.63 |
| | 3 | 715 | 16.34 | 2 | 21 | 4.06 |
| 1 | 1 | 640 | 16.39 | 4 | 21 | 3.63 |
| | 2 | 640 | 17.02 | 1 | 21 | 3.41 |
| | 3 | 640 | 18.02 | 1 | 21 | 3.13 |
| 2 | 1 | 708 | 17.60 | 2 | 21 | 2.85 |
| | 2 | 708 | 18.46 | 3 | 21 | 2.66 |
| | 3 | 708 | 19.11 | 6 | 21 | 2.38 |

## Study C: Inter-rater reliability study and final field study

**Purpose:** Study C was designed to evaluate the inter-rater reliability of OL to ensure that student OL results do not vary based on characteristics of trained assessment administrators. Additionally, Study C was intended to guide revisions to the final forms. Initially four forms of OL were developed, with the intention of narrowing down the content to three final forms based on data from Studies A–C. Thus, additional data were collected on Form 4 of OL in Study C to support item analysis for this alternate form.

**Participants:** This field study was conducted during the '14–'15 MOY benchmark administration period at schools E, F, G, and H. In total, 190 students in kindergarten through Grade 2 (kindergarten: n = 57; Grade 1: n = 66; Grade 2: n = 67) participated in the study. Students were assessed with OL by classroom teachers and Amplify consultants.

**Demographics:** Participants in this field study included students in Grades K–2 and testers who were educators at the schools where the study took place as well as Amplify consultants. Students were from the Pacific geographic division of the United States (U.S. Census Bureau, n.d.). The student sample, described in Table 6, was composed of participants from the following demographic categories: 33.12 percent male and 66.84 percent female; 3.68 percent White, 4.21 percent Black, 88.95 percent Hispanic, and 1.58 percent Asian. The student sample was 52.11 percent native English speaking, while the remaining 47.89 percent spoke English as a second language; 6.32 percent were students receiving special education services.

**Table 6. Study C participants and demographics**

| | Kindergarten | Grade 1 | Grade 2 |
|---|---|---|---|
| Sample size (n) | | | |
| Students | 57 | 66 | 67 |
| Gender (n) | | | |
| Male | 24 | 15 | 24 |
| Female | 33 | 51 | 43 |
| Ethnicity (n) | | | |
| White | 0 | 4 | 3 |
| Hispanic | 50 | 62 | 57 |
| Black | 4 | 0 | 4 |
| Native American | | | |
| Asian | 0 | 0 | 3 |
| Language status (n) | | | |
| Native speaker | 22 | 32 | 45 |
| English as a Second Language | 35 | 34 | 22 |
| Other demographics (n) | | | |
| Special education | 0 | 4 | 8 |

**Study design:** Data to support inter-rater reliability (IRR) analyses as well as examine psychometric properties (e.g., difficulty and discrimination) for Forms 1, 2, 3, and 4 were collected at MOY; students in the IRR analyses were assessed by two raters at the same time; one rater interacted directly with the student while scoring and the other shadow scored the student's performance. At schools E and F, 18 of the 437 students who took Form 2 at MOY (kindergarten: n = 6; Grade 1: n = 5; Grade 2: n =
7) were shadow scored by a second Amplify consultant on paper and pencil alongside the classroom teacher, who assessed students using a device. There were also 17 students (kindergarten: n = 6; Grade 1: n = 4; Grade 2: n = 7) administered Form 4 at MOY by two Amplify consultants on paper and pencil. At schools G and H, 67 students (kindergarten: n = 19; Grade 1: n = 26; Grade 2: n = 22) were assessed with either Forms 1 and 4 or Forms 3 and 4 by two Amplify consultants on paper and pencil. Again, the students who participated in Study C were selected based on teacher judgment of language proficiency to ensure a range of abilities was represented.

**Descriptive statistics:** Descriptive information regarding the Study C results is presented in Table 7 by grade and form. Again, results show a general increase in scores as grade increases, indicating that older students perform better on this measure than younger students. Larger differences in scores between forms are also observed in kindergarten with respect to Grades 1-2, in which score differences between forms decrease, indicating higher and more consistent performance as grade increases.

**Table 7. Study C descriptive statistics**

| Grade | Form | Sample Size (n) | Average score | Minimum score | Maximum score | Standard Deviation |
|---|---|---|---|---|---|---|
| Kindergarten | 1 | 7 | 8.43 | 1 | 15 | 5.22 |
| | 2 | 6 | 11.83 | 2 | 19 | 6.91 |
| | 3 | 12 | 14.08 | 8 | 19 | 3.65 |
| | 4 | 32 | 11.69 | 1 | 20 | 5.24 |
| 1 | 1 | 12 | 15.92 | 14 | 18 | 1.24 |
| | 2 | 5 | 14.40 | 6 | 19 | 4.98 |
| | 3 | 12 | 13.25 | 0 | 20 | 6.62 |
| | 4 | 37 | 14.51 | 0 | 20 | 4.86 |
| 2 | 1 | 12 | 13.83 | 4 | 21 | 6.38 |
| | 2 | 9 | 15.44 | 10 | 20 | 4.10 |
| | 3 | 10 | 13.20 | 0 | 19 | 6.89 |
| | 4 | 36 | 14.86 | 0 | 21 | 5.94 |

# Final Form Analysis and Construction

Data collected for Forms 1–4 in Studies B and C were used to conduct item analysis, with a focus on item difficulty and discrimination as well as on determining the need and potential location for a discontinue rule. Recall that OL consists of three benchmark forms, and that a fourth form was developed for field testing so that the best items could be selected for the final version of the measure after completion of the research. In addition to item-level analysis, results from the educator survey administered in Study B were used to determine the need for modifications to the final version of the measure.

## Item analysis

Item analysis results were examined in order to identify whether any item revisions were necessary. The items on each form were examined according to an Item Response Theory (IRT; Embretson & Reise, 2000) framework. IRT attempts to quantitatively model the likelihood that a student with a specific level of ability will answer a specific question correctly. Calibration of an IRT model results in parameter estimates of difficulty for each test item as well as estimates of student ability, placing both on a common scale that enables direct comparisons. The analyses were conducted using Winsteps Rasch calibration software (Linacre, 2014). Estimated item-difficulty values, item-fit indices, and point- biserial correlations (an indicator of item discrimination) are presented in Appendix A.

Item difficulties from Forms 1 to 3 suggest the three forms are of equivalent difficulty (mean difficulties of Forms 1 to 3 range from −2.17 to −1.86). Item infit and outfit statistics are within the range of 0.50 to 1.50, indicating that the items generally perform as expected across the range of student ability, except one item. It is possible that this item (the first item on Form 1) demonstrates both poor infit and outfit due to unfamiliarity with the assessment: sentence repetition is likely a novel form of assessment, and as students become accustomed to the procedure, their performance is more likely to reflect their actual level of syntactic comprehension. The point-biserial correlations for all the items are above 0.20, indicating good item discrimination. In sum, all the items in Forms 1 to 3 demonstrate good item difficulty, item fit, and item discrimination, and no items were replaced in Forms 1 to 3 based on item analysis.

## Survey analysis

The educator survey, described in further detail in the "Additional Validity Evidence" section, included questions about the difficulty, cultural and gender sensitivity, and appropriateness of the items in OL. Respondents were given the opportunity to call out items that they found problematic. These responses were reviewed, and two items from Forms 1 to 3 were identified for replacement. One item was replaced to address concerns about articulation clarity, and another item was replaced on the basis of cultural sensitivity.

## Summary

As a result of the item and survey analyses, two items were revised in Forms 1–3. One item from Form 4 was used to replace an item from Form 3, and one item from Form 2 was replaced with a novel item. The remaining items from Form 4 were unused and unnecessary for the final OL assessment, which consists of three forms of 21 items each. Item analysis also examined the need and potential location of a discontinue rule, but there was not a clear point in the assessment at which students had been performing poorly and would not succeed on further items. Furthermore, the measure is brief and even incorrectly repeated sentences provide rich data for item analysis. Thus it was determined that there should be no discontinue rule for OL.

# Reliability

Reliability is generally described as the consistency of a measuring instrument: reliability statistics present information about the precision of an instrument, expressed as a ratio. A test with perfect score precision has a reliability coefficient equal to 1, meaning that 100 percent of the variation among persons' scores is attributable to variation in the trait or skill the test measures, and none of the variation is attributable to error. Perfect reliability is unattainable in educational measurement; a test with a reliability coefficient of 0.90 is more likely. On such a test, 90 percent of the variation among students' scores is attributable to the trait or skill being measured, and 10 percent is attributable to errors of measurement. If the trait or skill were measured a second time, students' scores would fluctuate to some degree; that is, scores on the second test would not be perfectly consistent with the same students' initial scores.

Further, reliability is an essential characteristic of interim and formative assessments that are used for instructional decision-making; if results are spurious and unreliable, inappropriate decisions might be made. Salvia, Ysseldyke, and Bolt's (2013) standards for reliability were used to evaluate the reliability data for OL. According to these standards, a minimum reliability of 0.60 is required to make educational decisions about groups of students, a minimum of 0.70 suggests adequate reliability generally, a minimum of 0.80 is required for screening decisions, and a minimum of 0.90 is required for important educational decisions concerning an individual student. Decisions made from early identification or screening measures, such as OL, typically do not involve a high-stakes decision to change an individual student's placement or educational classification (Kaminski & Good, 1996).

This chapter provides details on three types of reliability evidence for OL: internal consistency reliability, inter-rater reliability, and alternate-form reliability.

- Internal consistency reliability refers to the degree of confidence in the precision of scores from a single measurement.
- Inter-rater reliability refers to the degree to which different raters consistently estimate the same student's performance.
- Alternate-form reliability refers to the extent to which test results generalize to different forms. Alternate forms of a test with different items should yield the same approximate scores.

## Internal consistency reliability

Internal consistency reliability of OL was estimated using Cronbach's alpha, based on classical test theory. Cronbach's alpha is the most widely used reliability coefficient that measures the degree of internal consistency/homogeneity between variables measuring one construct/concept, i.e., the degree to which different items measuring the same variable provide consistent results (Crocker & Algina, 1986). Study B provides data for internal consistency reliability analyses for Forms 1, 2, and 3.[6] Table 8 provides Cronbach's alpha and sample sizes for each OL form. Overall, Cronbach's alpha ranges from 0.85 to 0.86, which is above Salvia, Ysseldyke, and Bolt's (2013) criteria for making screening decisions based on assessment results. The highest levels of reliability are found in kindergarten and Grade 1, ranging from 0.82 to 0.87; in Grade 2, values are slightly lower, ranging from 0.76–0.78, likely because many Grade 2 students approach the score ceiling on the measure (average scores range from 17.60 to 19.11 out of 21 in Grade 2).

**Table 8. Internal consistency reliability of OL**

| Grade | Form | Cronbach's Alpha |
|---|---|---|
| All | 1 | 0.86 |
| | 2 | 0.86 |
| | 3 | 0.85 |
| Kindergarten | 1 | 0.87 |
| | 2 | 0.88 |
| | 3 | 0.86 |
| 1 | 1 | 0.84 |
| | 2 | 0.82 |
| | 3 | 0.82 |
| 2 | 1 | 0.78 |
| | 2 | 0.77 |
| | 3 | 0.76 |

[6] Reliability and validity analyses are based on the original versions of Forms 1–3; the updated versions include a total of two new items across the three forms; these items contain the same syntactic structure as the original items.

## Inter-rater reliability

Inter-rater reliability (IRR) indicates the extent to which test results generalize across assessors. IRR is important for screening assessments such as OL because student performance should be scored in the same manner by any trained administrator, leading to the same outcome. Score fluctuations are attributable to sources of error via the assessors, including scoring mistakes and differing interpretations of scoring procedures and student responses. The IRR estimates reported here are based on two independent assessors simultaneously scoring, or shadow scoring, student performance during a single test administration.

Agreement between raters is typically evaluated using either Cohen's kappa (for nominal variables) or intra-class correlations (for ordinal, interval, or ratio variables; Hallgren, 2012). Because OL scores are ordinal, we use the intra-class correlation (ICC) to evaluate IRR. ICC is one of the most  commonly used statistics for assessing IRR for ordinal, interval, or ratio variables and is suitable for studies with two or more raters (Hallgren, 2012). Cicchetti (1994) provides commonly cited interpretations of agreement based on ICC values: ICC values less than 0.40 are poor, values between 0.40 and 0.59 are fair, values between 0.60 and 0.74 are good, and values between 0.75 and 1.00 are excellent. In addition, percent agreement within a reasonable range is provided for each measure. Percent agreement is calculated by dividing the number of instances of score agreement across raters by the total number of scores; thus, percent agreement can vary between 0 and 100 percent. For OL, percent agreement within 1 point is provided (fluctuations within 1 point are reasonable for OL, which has a total score of 21 points). The higher the percent agreement, the stronger the evidence for IRR.

Study C provides data for the IRR analysis. Results of the IRR analysis for OL are reported in Table 9, including the number of assessments included in each analysis, the number of raters, ICCs, and percent agreement. Sample sizes are provided for the number of assessments rather than the number of unique students; because students were assessed multiple times by different raters with different forms, the number of unique students does not match the number of assessments. Overall and by grade, intra-class correlations are high (ICC = 0.75–0.85) and constitute excellent agreement according to the standards established by Cicchetti (1994). Overall, percent agreement is 71.57 percent, and by grade it ranges from 58.06 to 80.56 percent. We note that agreement is highest for the oldest and highest performing students, likely because consistent scoring is easiest when fewer mistakes are made. Taken together, the intra-class correlations and percent agreement results provide strong support for the inter-rater reliability of OL.

Table 9. Inter-rater reliability of OL

| Grade | Kindergarten | Raters (n) | Intra-class correlation | Percent agreement within 1 point |
|---|---|---|---|---|
| All | 102 | 9 | 0.82 | 71.57% |
| Kindergarten | 30 | 9 | 0.82 | 58.06% |
| 1 | 34 | 9 | 0.75 | 74.29% |
| 2 | 35 | 9 | 0.85 | 80.56% |

**Note:** There is one more assessment included in the overall IRR analysis than there are in the total of the three grade-specific analyses because the grade of one included student was unknown.

## Alternate-form reliability

Alternate-form reliability indicates the extent to which test results generalize to different item samples. Students are tested with two different but equivalent (i.e., alternate) forms of the test, and scores from these two forms are correlated. Alternate forms of a test should yield approximately equivalent scores. Administering alternate forms of the same measure may lead to practice effects due to the similarity of the items and administration procedures, but to a lesser degree than in test-retest reliability studies in which the same form is administered to students twice.

Study B provided data for item analysis for Forms 1 to 3. The distribution of item difficulty on each form should be similar; item difficulties should not be significantly different for different forms. In Appendix A, the detailed item analysis results are presented (described in the previous “Item and Survey Analysis” section). ANOVA results suggest that the item difficulties are not significantly different across the three different forms [F (2, 60) = 0.21, n.s.]. The figure in Appendix A also indicates that item difficulty distributions are very similar across the three forms. Lastly, the Kolmogorov-Smirnov test suggests that the item difficulty distributions for the three forms are not significantly different (D = 0.14, n.s.), further providing evidence for the alternate-form reliability of OL.

# Validity

Validity refers to the degree to which a test measures the construct that it claims or was intended to measure. Formally, validity is defined as the degree to which evidence and theory support the interpretation of test scores according to test usage (American Educational Research Association, 1999). In other words, validity represents our degree of confidence that interpretations of test scores accurately represent what we believe they do (e.g., high scores on a language assessment actually represent high language skill). In this sense, validity is a way to describe a test's accuracy or utility.

Validity is not "proven" but rather evidence is collected to strengthen the assertion that a test accurately measures the desired construct(s). Validity was traditionally considered a property that assessments themselves possessed; it was categorized as content-, construct-, and criterion validity. The current view, however, considers a more unified treatment under which validity evidence is collected to support test score interpretations (Kane, 2001; Messick, 1989) and may be captured under a more general heading of evidence for construct validity. Assessing the validity of a test involves the use of data and other information both internal and external to the test instrument itself.

To facilitate discussion and demonstration, evidence for the construct validity and criterion validity of OL is presented via concurrent and prediction results. Criterion-related validity is the extent to which student performance on the assessment procedure being validated can estimate student performance on a criterion measure (Salvia, Ysseldyke, & Bolt, 2013). Criterion-related validity includes concurrent and predictive validity. Evidence for the concurrent or predictive validity of an assessment refers to the degree to which current outcomes are associated with outcomes on an external, conceptually related instrument administered near-concurrently (concurrent validity evidence) or subsequently (predictive validity evidence). Predictive correlations are attenuated by time due to language growth that occurs in the interim between testing occasions; both predictive and concurrent correlations are attenuated by differences in test content specifications. Concurrent and predictive validity of OL were evaluated against the TOLD-P:4 SI. These assessments both use sentence imitation tasks to measure receptive oral syntax. One of the major differences between OL and the TOLD-P:4 SI is that OL is designed for students in kindergarten through Grade 2, while the TOLD-P:4 SI is designed for children age 4–9 and therefore includes a broader range of syntactic structures than OL.

Concurrent validity was evaluated for a subset of students in Study B who were administered both OL and the TOLD-P:4 SI within one month of each other at EOY. Predictive validity was evaluated for students administered OL followed by the TOLD-P:4 SI; specifically, students in the predictive validity analyses took OL at BOY and MOY and TOLD-P:4 SI at EOY approximately 4–7 months apart. These analyses provide an estimate of the linear relationship between OL scores and scores on the TOLD P:4 SI, which covers the same linguistic domain using the same type of task.

Raw scores from the TOLD P:4 SI resulting from Study B were converted into percentile ranks and scaled scores based on student age, using lookup tables provided in the TOLD-P:4 Examiner's Manual (Newcomer & Hammill, 2008). Student performance on the TOLD-P:4 SI was markedly lower in Grade 1 than kindergarten and Grade 2 largely due to the scaled score cut point jump from 11 to 14 points for the "Average" performance category for students who are 6.0 to 6.5 years of age versus 6.5 to 7.0 years of age. As most Grade 1 students in Study B were between 6.5 and 7.5 years of age, and therefore susceptible to this cut point jump, a raw score transformation was applied to yield an appropriate distribution for subsequent analysis. Specifically, scaled scores for Grade 1 students in this analysis were increased by a single point. The figure in Appendix B shows the distribution of scaled scores by grade, after the transformation. The table in Appendix B provides the distributions of performance categories on the TOLD-P:4 SI subtest by grade.

## Concurrent validity

Pearson correlations were used to characterize the concurrent relationship of OL with the TOLD-P:4 SI at EOY using data provided by Study B. Pearson correlation coefficients (after removing multivariate outliers) and sample sizes are provided for each grade in Table 10.

**Table 10. Concurrent validity of OL with TOLD-P:4 SI**

| Grade | Correlations of OL at EOY with TOLD-P4 SI at EOY | Sample size (n) |
|---|---|---|
| Kindergarten | 0.48 | 186 |
| 1 | 0.50 | 170 |
| 2 | 0.56 | 195 |

The concurrent correlations of OL with the TOLD-P:4 SI at EOY were 0.48 in kindergarten, 0.50 in Grade 1, and 0.55 in Grade 2. The results indicate a moderate positive concurrent relationship between OL and the TOLD-P:4 SI and provide adequate concurrent validity evidence for OL. The magnitude of these correlations, however, may be due to differences in the score distributions of the sample on the two measures: the sample was generally high performing on OL, while lower performance was observed on the TOLD-P:4 SI. Differences in the syntactic complexity of the measures and the training for the data collectors who administered them may explain these differences in score distributions. While both measures are sentence imitation tasks, the target age range of the TOLD-P:4 SI is 4 to 9 years, while OL targets students in kindergarten through Grade 2 (students in these grades are typically 5 to 8 years old). The TOLD-P:4 SI therefore implies a broader range of syntactic complexity, making it more difficult for students to approach the maximum score on the TOLD-P:4 SI than on OL. Furthermore, as noted in the "Additional Validity Evidence" section, the amount and type of training for educators who collected the OL data was variable, while the TOLD- P:4 SI data collectors received consistent training; this difference in training may have led to stricter scoring of the TOLD-P:4 SI and, therefore, lower score distributions.

## Predictive validity

Pearson correlations were used to characterize the predictive relationship between OL at BOY and MOY and the TOLD-P:4 SI at EOY. Pearson correlation coefficients (after removing multivariate outliers) and the sample sizes are provided for each grade in Table 11.

**Table 11. Predictive validity of OL with TOLD-P:4 SI**

| Grade | Correlations between OL at BOY and TOLD- P:4 SI at EOY | BOY-EOY sample size (n) | Correlations between OL at MOY and TOLD- P:4 SI at EOY | MOY-EOY sample size (n) |
|---|---|---|---|---|
| Kindergarten | 0.45 | 186 | 0.46 | 186 |
| 1 | 0.61 | 170 | 0.62 | 170 |
| 2 | 0.62 | 195 | 0.60 | 195 |

OL had positive predictive correlations with the EOY administration of the TOLD-P:4 SI (kindergarten: r = 0.45–0.46; Grade 1: r = 0.61–0.62; Grade 2: r = 0.60–0.62). These results indicate a moderate positive predictive relationship between OL and the TOLD-P:4 SI and provide adequate predictive validity evidence for OL. As mentioned earlier, the magnitude of these correlations may be due to differences in the score distributions of the sample on the two measures.

## Additional validity evidence

A survey was designed to collect teacher feedback about the content and administration of OL and participation in OL research (Studies B and C). The survey was paper-based, and all educators who administered OL within Studies B and C were asked to participate, although participation was voluntary. The survey collected information about educators' training with, use of, and level of satisfaction with OL and the Scoring Guide, teacher demographic information, and perceptions of OL. The survey was composed of 58 questions divided into seven thematic domains: (1) Educator demographics; (2) Educator use of OL; (3) OL instructions and scoring procedures; (4) OL items; (5) OL training; (6) OL Scoring Guide; and (7) Overall satisfaction with OL. There were 38 structured survey items (e.g., multiple choice), and survey participants could provide open-ended feedback through 20 open-ended questions.

Completed surveys were submitted confidentially to Amplify consultants who worked in the schools as reading coaches. As described previously, survey results informed changes to the original OL materials as necessary. Particular attention was given to feedback specific to assessment items; this feedback was reviewed internally to determine whether a revision was warranted. Based on survey responses, revisions were made to two items. Survey results were also provided to Amplify Account Managers to further address any issues with the school district.

## Survey results

The number of completed responses totaled 60, representing six schools in one large, urban district.

**Educator Demographics.** All respondents reported using the OL during the field study. All respondents were classroom teachers who taught Grades K–3 with a median of six years' experience teaching their current grade. All respondents held a bachelor's degree or higher, and the majority were female.

**Educator use of OL.** Respondents reported assessing both their own students and other teachers' students with OL, including students classified as Native English speakers, ELLs, and English as a Second Language (ESL)[6] speakers. The median assessment time reported was 6 minutes per student. Respondents reported using the OL data for identifying students for oral language instruction, monitoring growth, forming intervention groups, guiding small-group and whole-class instructional content in speaking, listening, reading, and writing, and making special education referrals.

**OL use.** The majority of respondents found the OL administration instructions to be clear to students (87%) and teachers (77%). The majority of respondents also found the scoring procedures to be clear (77%) and fair (67%).

**OL items.** The majority of respondents (77%) found the difficulty of the items on the OL to be just right. The majority found the items to be culturally sensitive (58%), free from gender bias (92%), free of inappropriate content (90%), and age-appropriate for kindergarten through Grade 2 (83%). Comments regarding issues with specific items were reviewed and taken into account when making revisions to the items.

**OL training.** The majority of respondents (67%) reported that they had not received OL training during the '14–'15 school year, while a little over half (55%) reported having been trained on a similar measure in previous years; overall, 63 percent had been trained to use either OL or a similar measure (or both). All trained respondents felt prepared to administer OL as a result of training, and the majority (65%) reported feeling prepared to use OL to guide instruction.

**OL scoring guide.** The majority of respondents (82%) reported that they had not accessed the Oral Language Scoring Guide. Half of the respondents who read the document had conducted an error pattern analysis, and a quarter reported using the instructional recommendations in the document.

**Overall satisfaction with OL.** The majority of respondents (55%) were satisfied with OL overall, while 32 percent felt that OL provides an accurate representation of student oral language comprehension. Given the variability in the amount and type of training reported by educators, these results highlight the importance of providing training that includes a rationale for the measure's design, instructions for administering and scoring the measure, assessment practice, and guidance on how to use the data. These findings will be used to encourage greater attention to training and dissemination of supporting materials in the future.

[6] Native English speakers speak English as a first language. English Language Learners (ELLs) are non-native speakers of English who have not reached fluency (according to district assessments). English as a Second Language (ESL) speakers are non-native speakers of English who have already become fluent or have reached a level of proficiency such that they are no longer considered ELLs.

# Cut Points

Cut points for determining risk in oral syntactic comprehension development based on OL results were set for each grade and benchmarking period to help educators track growth against performance standards over time. Cut points were set using data from the students in Study B who were administered OL at BOY, MOY, and EOY and the TOLD-P:4 SI at EOY of the ’14–’15 school year. Transformed TOLD scaled scores and interpretations were used for Grade 1 (see the “Validity” section); kindergarten and Grade 2 scales scores were not transformed.

The primary consideration in setting the cut points was the results of the contrasting group analysis (Cizek & Bunch, 2007). Under this method, mean and median OL performance is calculated for three TOLD-P:4 SI performance levels defined based on percentile ranks: below the 25th percentile (“Below Average”, “Poor”, and “Very Poor” performance according to the TOLD-P:4 SI labels), between the 25th and 37th percentile (the lower part of “Average” performance), and above the 37th percentile (the higher part of “Average” performance plus “Above Average”, “Superior”, and “Very Superior” performance). These groups classify student performance in terms of the normative sample, which in turn helps educators identify students who require intervention to accelerate growth and move toward proficiency. This way of classifying performance aligns with the purpose of OL. The midpoint of these mean and median values is next calculated to provide tentative thresholds between the three TOLD-P:4 SI performance levels, used to set the tentative cut points for the Well Below Benchmark, Below Benchmark, and At or Above Benchmark (i.e., Red, Yellow, and Green, respectively) performance levels on OL.

The secondary consideration was the distribution of students across the TOLD-P:4 SI performance levels. We tried to keep the percentages of students in each OL performance level (i.e., Well Below Benchmark, Below Benchmark, and At or Above Benchmark) consistent from predictor to criterion. For example, 50 percent of students in our kindergarten EOY sample scored below the 25th percentile on the TOLD-P:4 SI; therefore, we set the kindergarten EOY Well-Below Benchmark/Below Benchmark (i.e., Red/Yellow) cut point so that 50 percent of students also had interpretations of Well Below Benchmark (i.e., Red).

Based on these two considerations, the range of tentative cut points was submitted to examination of classification accuracy, specificity, sensitivity, logistic regression analyses results (i.e., likelihood of being at or above the 25th percentile on the TOLD-P:4 SI), negative prediction value (i.e., the probability of being at or above the 25th percentile on the TOLD-P:4 SI given At or Above Benchmark status on OL), marginal percentages (i.e., the percentages of students at or above the 25th percentile on the TOLD-P:4 SI for a specific OL score), and the score distributions of the three TOLD-P:4 SI performance levels.

Additional considerations in examining the possible sets of cut points across grades and benchmarking periods as a system included the following:

1. As the OL forms are of equivalent difficulty across all benchmarking periods in kindergarten through Grade 2 and should allow for demonstration of growth, the cut points should be monotonically non-decreasing as grade and/or benchmark period increases;
2. Allowing for minor student errors, cut points should fall below the maximum score for OL (i.e., 21);
3. Theoretical knowledge about the pace and trajectory of language and reading development was considered when choosing between multiple possible cut points.

Based on these procedures and considerations, cut points were set for OL that identify students as Well-Below Benchmark, Below Benchmark, and At or Above Benchmark in kindergarten through Grade 2. The final cut points are presented in Table 12.

**Table 12. Cut points for OL**

| Grade | BOY | | MOY | | EOY | |
|---|---|---|---|---|---|---|
| | Well Below/ Below Benchmark (Red/Yellow) | Below/At or Above Benchmark (Yellow/ Green) | Well Below/ Below Benchmark (Red/Yellow) | Below/At or Above Benchmark (Yellow/ Green) | Well Below/ Below Benchmark (Red/Yellow) | Below/At or Above Benchmark (Yellow/ Green) |
| Kindergarten | 13 | 16 | 14 | 17 | 15 | 17 |
| 1 | 15 | 17 | 16 | 18 | 17 | 18 |
| 2 | 17 | 18 | 18 | 19 | 18 | 19 |

**Note:** The cut points listed mean that a student would need a score greater than or equal to the score listed. For instance, in kindergarten at BOY, a student scoring 0–12 would be classified as Well Below Benchmark, 13–15 would be classified as Below Benchmark, and 16–21 would be classified as At or Above Benchmark.

# Appendix 1. Item Difficulty of OL

Item Difficulty of OL

| Form | Item Number | n | IRT Difficulty | p-value | SE | Infit | Outfit | Point-Biserial Correlation |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1802 | –2.75 | 0.89 | 0.08 | 1.28 | 2.34 | 0.28 |
| 1 | 2 | 1802 | 0.88 | 0.41 | 0.06 | 0.98 | 0.90 | 0.48 |
| 1 | 3 | 1802 | –1.63 | 0.78 | 0.07 | 1.02 | 1.04 | 0.49 |
| 1 | 4 | 1802 | –3.48 | 0.93 | 0.10 | 1.12 | 1.31 | 0.33 |
| 1 | 5 | 1802 | –1.03 | 0.70 | 0.06 | 0.96 | 1.12 | 0.54 |
| 1 | 6 | 1802 | –2.72 | 0.88 | 0.08 | 1.05 | 1.41 | 0.42 |
| 1 | 7 | 1802 | –3.03 | 0.91 | 0.09 | 0.98 | 0.83 | 0.44 |
| 1 | 8 | 1802 | 2.09 | 0.25 | 0.07 | 1.04 | 1.30 | 0.37 |
| 1 | 9 | 1802 | –3.98 | 0.95 | 0.12 | 1.01 | 1.71 | 0.33 |
| 1 | 10 | 1802 | –3.10 | 0.91 | 0.09 | 0.90 | 1.07 | 0.47 |
| 1 | 11 | 1802 | 0.01 | 0.55 | 0.06 | 0.99 | 0.95 | 0.51 |
| 1 | 12 | 1802 | –2.98 | 0.90 | 0.09 | 0.83 | 0.45 | 0.53 |
| 1 | 13 | 1802 | –3.00 | 0.90 | 0.09 | 0.95 | 0.96 | 0.45 |
| 1 | 14 | 1802 | –0.30 | 0.59 | 0.06 | 0.93 | 0.88 | 0.55 |
| 1 | 15 | 1802 | –3.55 | 0.94 | 0.10 | 1.09 | 1.66 | 0.32 |
| 1 | 16 | 1802 | –2.81 | 0.89 | 0.09 | 1.12 | 1.34 | 0.37 |
| 1 | 17 | 1802 | 1.35 | 0.35 | 0.06 | 0.96 | 1.11 | 0.45 |
| 1 | 18 | 1802 | –3.32 | 0.92 | 0.10 | 0.88 | 0.61 | 0.47 |
| 1 | 19 | 1802 | –2.26 | 0.84 | 0.08 | 0.86 | 0.81 | 0.56 |
| 1 | 20 | 1802 | –1.34 | 0.74 | 0.06 | 1.01 | 1.03 | 0.51 |
| 1 | 21 | 1802 | –2.09 | 0.83 | 0.07 | 0.93 | 0.77 | 0.53 |

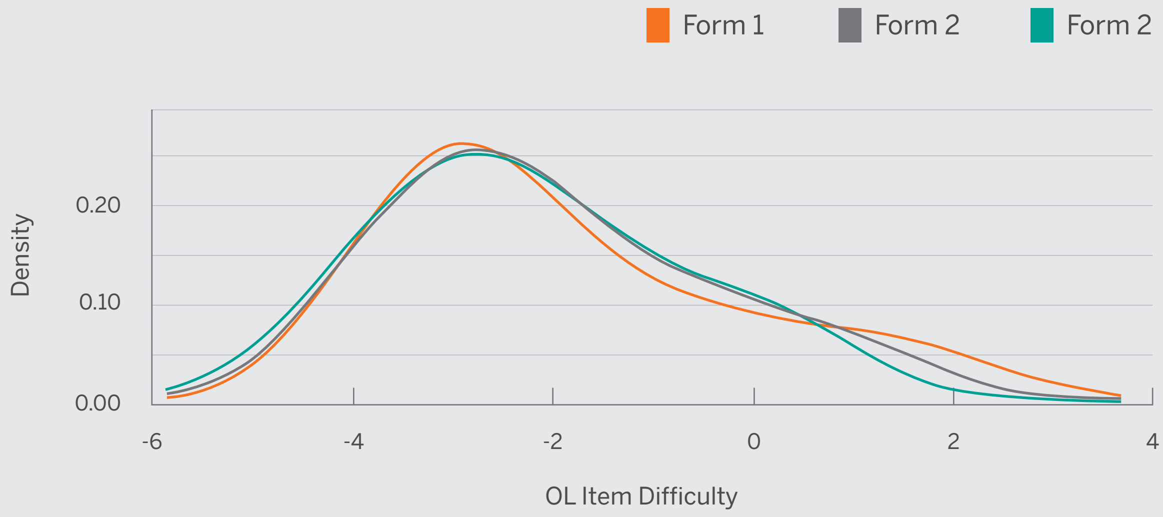
Item Difficulty of OL (cont'd)

| Form | Item Number | n | IRT Difficulty | p-value | SE | Infit | Outfit | Point-Biserial Correlation |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 1876 | –2.69 | 0.90 | 0.09 | 1.24 | 1.55 | 0.33 |
| 2 | 2 | 1876 | 0.94 | 0.44 | 0.06 | 0.96 | 0.92 | 0.47 |
| 2 | 3 | 1876 | –1.72 | 0.82 | 0.07 | 0.97 | 1.02 | 0.52 |
| 2 | 4 | 1876 | –3.90 | 0.96 | 0.12 | 1.10 | 1.38 | 0.32 |
| 2 | 5 | 1876 | –0.31 | 0.64 | 0.06 | 0.94 | 0.90 | 0.53 |
| 2 | 6 | 1876 | –2.38 | 0.88 | 0.08 | 1.04 | 1.16 | 0.45 |
| 2 | 7 | 1876 | –3.45 | 0.94 | 0.11 | 1.00 | 1.06 | 0.39 |
| 2 | 8 | 1876 | 0.07 | 0.58 | 0.06 | 1.01 | 1.01 | 0.48 |
| 2 | 9 | 1876 | –4.02 | 0.96 | 0.13 | 0.97 | 0.90 | 0.35 |
| 2 | 10 | 1876 | –2.79 | 0.91 | 0.09 | 0.92 | 0.95 | 0.49 |
| 2 | 11 | 1876 | –0.67 | 0.69 | 0.06 | 1.04 | 1.09 | 0.49 |
| 2 | 12 | 1876 | –2.80 | 0.91 | 0.09 | 0.89 | 0.74 | 0.51 |
| 2 | 13 | 1876 | –3.02 | 0.92 | 0.10 | 0.97 | 0.95 | 0.44 |
| 2 | 14 | 1876 | –0.95 | 0.73 | 0.06 | 0.97 | 0.93 | 0.52 |
| 2 | 15 | 1876 | –3.66 | 0.95 | 0.12 | 1.01 | 0.72 | 0.38 |
| 2 | 16 | 1876 | –3.26 | 0.93 | 0.10 | 1.08 | 1.13 | 0.37 |
| 2 | 17 | 1876 | 1.23 | 0.40 | 0.06 | 0.99 | 0.95 | 0.44 |
| 2 | 18 | 1876 | –2.35 | 0.88 | 0.08 | 0.97 | 1.00 | 0.49 |
| 2 | 19 | 1876 | –2.37 | 0.88 | 0.08 | 0.81 | 0.66 | 0.58 |
| 2 | 20 | 1876 | –1.50 | 0.79 | 0.07 | 1.19 | 1.37 | 0.40 |
| 2 | 21 | 1876 | –2.32 | 0.87 | 0.08 | 0.92 | 0.94 | 0.52 |

### Item Difficulty of OL (cont'd)

| Form | Item Number | n | IRT Difficulty | p-value | SE | Infit | Outfit | Point-Biserial Correlation |
|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1897 | –2.28 | 0.89 | 0.08 | 1.00 | 0.98 | 0.45 |
| 3 | 2 | 1897 | 0.02 | 0.63 | 0.06 | 0.97 | 0.98 | 0.49 |
| 3 | 3 | 1897 | –1.72 | 0.84 | 0.07 | 0.98 | 0.92 | 0.49 |
| 3 | 4 | 1897 | –3.24 | 0.94 | 0.11 | 1.17 | 1.73 | 0.27 |
| 3 | 5 | 1897 | –1.06 | 0.77 | 0.06 | 0.90 | 0.89 | 0.55 |
| 3 | 6 | 1897 | –3.00 | 0.93 | 0.10 | 1.00 | 1.26 | 0.39 |
| 3 | 7 | 1897 | –2.73 | 0.92 | 0.09 | 1.09 | 1.33 | 0.36 |
| 3 | 8 | 1897 | 0.27 | 0.59 | 0.06 | 0.94 | 0.95 | 0.50 |
| 3 | 9 | 1897 | –4.20 | 0.97 | 0.15 | 0.96 | 0.86 | 0.30 |
| 3 | 10 | 1897 | –2.94 | 0.93 | 0.10 | 1.01 | 1.06 | 0.39 |
| 3 | 11 | 1897 | –1.61 | 0.83 | 0.07 | 0.97 | 0.99 | 0.50 |
| 3 | 12 | 1897 | –3.73 | 0.96 | 0.13 | 0.84 | 0.60 | 0.44 |
| 3 | 13 | 1897 | –3.06 | 0.94 | 0.10 | 1.01 | 1.31 | 0.38 |
| 3 | 14 | 1897 | –1.43 | 0.81 | 0.07 | 1.02 | 1.03 | 0.47 |
| 3 | 15 | 1897 | –4.25 | 0.98 | 0.16 | 1.07 | 0.92 | 0.25 |
| 3 | 16 | 1897 | –3.60 | 0.96 | 0.12 | 1.12 | 1.20 | 0.28 |
| 3 | 17 | 1897 | 0.60 | 0.54 | 0.06 | 1.04 | 1.03 | 0.46 |
| 3 | 18 | 1897 | –2.17 | 0.88 | 0.08 | 1.01 | 1.11 | 0.45 |
| 3 | 19 | 1897 | –2.84 | 0.93 | 0.10 | 0.84 | 0.61 | 0.51 |
| 3 | 20 | 1897 | –0.29 | 0.67 | 0.06 | 1.03 | 1.03 | 0.48 |
| 3 | 21 | 1897 | –2.30 | 0.89 | 0.08 | 0.99 | 1.03 | 0.46 |

Item Difficulty of OL



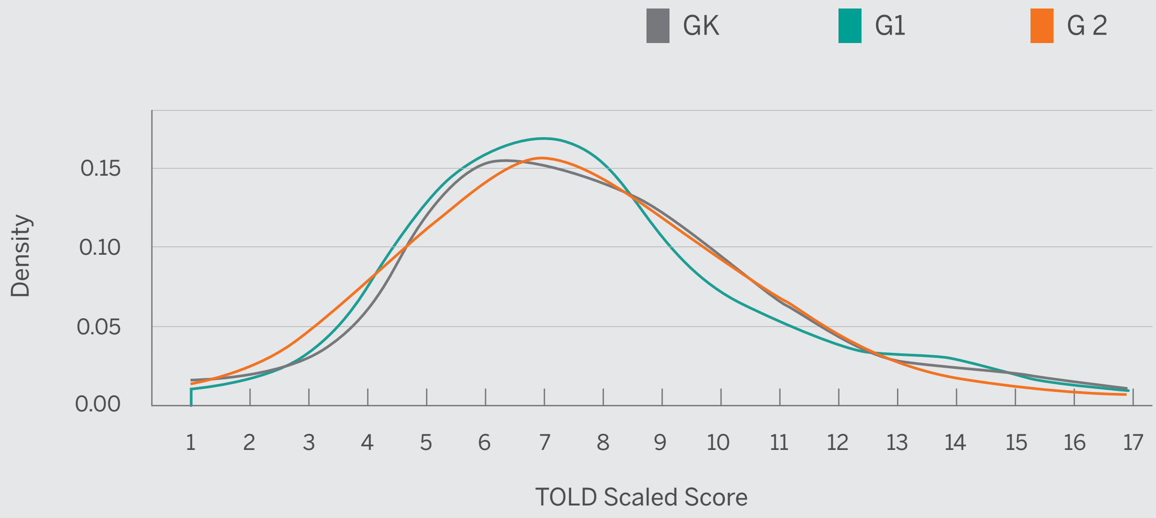
# Appendix 2. TOLD:P4 SI Distribution

TOLD:P-4 SI Distribution

| TOLD Performance Category | Kindergarten | Grade 1 | Grade 2 |
|---|---|---|---|
| Very Superior | 1 (0.54%) | 0 (0.00%) | 0 (0.00%) |
| Superior | 3 (1.61%) | 2 (1.18%) | 0 (0.00%) |
| Above Average | 6 (3.23%) | 10 (5.88%) | 5 (2.56%) |
| Average | 83 (44.62%) | 66 (38.82%) | 83 (42.56%) |
| Below Average | 57 (30.65%) | 53 (31.18%) | 62 (31.79%) |
| Poor | 29 (15.59%) | 35 (20.59%) | 36 (18.46%) |
| Very Poor | 7 (3.76%) | 4 (2.35%) | 9 (4.62%) |

**Note:** Sample sizes and percentages are displayed. Grade 1 scaled scores were transformed by adding 1 point to each student's score

TOLD:P-4 SI Scaled Scores by Grade



# References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: Author.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). Bringing words to life: Robust vocabulary instruction. New York: Guilford Press.

Catts, H. (2009). The narrow view of reading promotes a broad view of comprehension. Language, Speech, and Hearing Services in Schools, 40(2), 178–183.

Catts, H., Adlof, S., and Weismer, S. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. Journal of Speech, Language, and Hearing Research, 49, 278–293.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284.

Clay, M. M. (1971). Sentence repetition: Elicited imitation of a controlled set of syntactic structures by four language groups. Monographs of the Society for Research in Child Development, 1–85.

Clay, M.M., Gill, M., Glynn, T., McNaughton, T., & Salmon, K. (1983). Record of oral language and biks and gutches. Portsmouth, NH: Heinemann.

Clay, M.M., Gill, M., Glynn, T., McNaughton, T., & Salmon, K. (2007). Record of oral language: Observing changes in the acquisition of language structures: A guide for teaching. Auckland, NZ: Heinemann.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart and Winston.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. New York: Psychology Press.

Gough, P. & Tunmer, W. (1986). Decoding, reading, and reading disability. Remedial and Special Education, 7, 6–10.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23.

Kane, M. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38(4), 319–342.

Kaminski, R. & Good, R. (1996). Toward a technology for assessing basic early literacy skills. School Psychology Review, 25(2), 215–227.

Linacre, J. M. (2014). Winsteps® Rasch measurement computer program User's Guide. Beaverton, OR: Winsteps.com

Menyuk, P. (1969). Sentences children use. MIT research monograph, 52.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York: Macmillan.

Newcomer, P. L., & Hammill, D. D. (2008). Test of Language Development: Primary (TOLD-P:4). Austin, TX.

Romeo, K., Gentile, L., & Bernhardt, E. (2008). Sentence repetition and story retelling as indicators of language proficiency in young bilingual children. In 57th Yearbook of the National Reading Conference, 298–310.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). Assessment in special and inclusive education. Cengage Learning.

Seastrom, M. (2010). Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting. Technical Brief. NCES 2011–603. National Center for Education Research.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. Language Testing, 29(3) 325–344.

# For more information visit amplify.com

**Corporate:**
55 Washington Street
Suite 900
Brooklyn, NY 11201-1071
(212) 796-2200

**Sales Inquiries:**
(866) 212-8688 • amplify.com

Amplify.