WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Technical Report

No. TR-2025-2
July 2025

WIDA Alternate Screener Technical Brief

School of Education
UNIVERSITY OF WISCONSIN–MADISON

Contents

1. Introduction

This technical report is intended to provide information regarding the WIDA Alternate Screener assessment released in July 2025.

WIDA Alternate Screener was field tested in 2022–2023 along with WIDA Alternate ACCESS. The WIDA Alternate Screener assessment was developed based on the field test items. This report provides a brief overview of the tool's design and technical properties. We provide information on the purpose of the test, the test design, and the development process, including pilot testing and field testing. Statistical analyses of the field test include details on item performance and scaling. We also provide test reliability statistics. Finally, we examine the interrater agreement rates of the raters who scored field test students' responses on the performance tasks. All statistical results presented in this report are based on data gathered from the 2022–2023 field test.

The technical information herein is intended for use by those who have technical knowledge of test construction and measurement procedures, as stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

2. Purpose of WIDA Alternate Screener for Alternate ACCESS

WIDA Alternate Screener (hereafter, Alternate Screener) is an assessment designed to provide an initial measure of English language proficiency for students with the most significant cognitive disabilities. It is given to incoming students in grades K–12 to help determine whether they qualify for English language support services. The assessment is typically appropriate for English learners (ELs) who participate, or who would be likely to participate, in your state's alternate content assessments. The results of the assessment help educators in making decisions regarding the English language support services a student may need.

Alternate Screener is intended to be used as one element in the decision-making process of identifying a student as an EL. This decision should ideally be supported by additional evidence, such as the language of instruction in previous schooling, recommendations from previous teachers, the child's home language survey, or any of the recommended or required criteria as determined by the state or district.

WIDA recommends using the Overall score for decision-making purposes where all four domains have been administered. Caution is advised when interpreting domain-specific scores on the Alternate Screener due to the limited number of items in each domain. The specific Alternate Screener scores used in identification are determined at the state level.

3. WIDA Alternate Screener Test Design

WIDA Alternate Screener is an English language proficiency screening assessment for students with the most significant cognitive disabilities. It is typically appropriate for potential ELs who participate, or who would be likely to participate, in a state's alternate content assessments.

Alternate Screener assesses the four language domains of Listening, Reading, Speaking, and Writing. Items in each domain address a sampling of each of the five WIDA English Language Development Standards:

1. Language for Social and Instructional Purposes
2. Language for Mathematics
3. Language for Science
4. Language for Language Arts
5. Language for Social Studies

Items target three of the six WIDA alternate English language proficiency levels: 3—Developing, 4—Expanding, and 5—Bridging. Level 6—Reaching is considered proficient in English and therefore not assessed on the Alternate Screener assessment. Because the Alternate Screener is designed to identify students who may be eligible for English learner services, the focus is on flagging those who are likely performing below Proficiency Level (PL) 3. If a student answers the harder items correctly, they are likely at or above PL4, and therefore not in need of further testing via Alternate ACCESS.

3.1 General Principles of Design and Functionality

WIDA Alternate Screener is administered in the following grade-level clusters:

- Grades K–2
- Grades 3–5
- Grades 6–12

Table 3.1.1

Components of WIDA Alternate Screener

Domain	Number of Tasks	Targeted Proficiency Levels
Listening	3	3-5
Reading	3	3-5
Speaking	2	3-5
Writing	1	3-5

Test administrators (TAs) can administer the domain tests in any order, and the different domain tests can be administered on different days, with no minimum or maximum break between the administrations.

Depending on the student's needs, Alternate Screener should take 30 minutes or less to complete. However, due to the nature of the test and the unique abilities and behaviors of individual students, actual test times can vary widely.

4. WIDA Alternate Screener Test Development

Alternate Screener items were selected from those created for the updated Alternate ACCESS assessment. For detailed information regarding the development of the Alternate Screener items, refer to the *Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration*.

4.1 Item Development and Review

For detailed information regarding the item development of the Alternate Screener items, refer to the *Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration*.

4.1.1 Item Writing and Development

For detailed information regarding the item writing and development of the Alternate Screener items, refer to the *Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration*.

4.1.2 Bias, Sensitivity, and Content Review

For detailed information regarding bias, sensitivity, and content reviews of the Alternate Screener items, refer to the *Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration*.

4.2 Pilot Testing (Cog Lab)

For detailed information regarding pilot testing of Alternate Screener items, refer to the *Alternate ACCESS Field Test Technical Brief*.

4.3 Field Testing

The goal of the Alternate ACCESS field test (FT) was to collect the data needed to select items and tasks to update Alternate ACCESS, to develop an Alternate Screener, and to develop sufficient items for the creation of at least one new test form in each of the four domains of Listening, Speaking, Reading, and Writing, and in the grade-level clusters of K–2, 3–5, 6–8, and 9–12.

Due to the number of items included in the FT, as well as the size of the tested student population, WIDA conducted a stand-alone field test in all WIDA states, territories, and agencies using a census-based field test administration. All WIDA Consortium members were asked to administer the FT form between two and four weeks after the operational administration of ACCESS. The testing window for the Alternate ACCESS FT was February 14–

April 17, 2023. A total of 21,551 students in 40 US states, territories, and agencies participated in the Alternate ACCESS FT.

For the FT, five test forms were spirally distributed to all WIDA members at the SEA level. The sampling plan was developed to account for student demographic characteristics and students' average Alternate ACCESS scores across groups so that each FT form had similar test-taker numbers and characteristics. Additionally, each FT form included states with both large and small populations, and similar aggregated composite scores from the prior test administration across the five FT forms. The FT forms consisted of 10 Listening items, 10 Reading items, eight Speaking items, and eight Writing items. The total estimate of participating students per FT form by grade-level cluster was projected to be at least 1,000 students. Each spiral form included both horizontal and vertical linking items. The detailed horizontal and vertical scaling design is described in the *WIDA Alternate ACCESS Field Test Technical Brief.*

4.4 Selection of Final Form

During the WIDA Alternate ACCESS form selection, staff also flagged items meeting the technical requirements for the WIDA Alternate Screener. The selection of Alternate Screener items was finalized with some updates after the Alternate ACCESS series 602 (2023–24 SY) and the Screener validation study. The development team reviewed all available items for the Alternate Screener forms to ensure a range of English Language Development (ELD) standards, and proficiency levels were considered for each grade-level cluster form.

Because the Alternate Screener is designed to identify students who may be eligible for English learner services, the focus is on flagging those who are likely to perform below PL3. If a student answers the harder items correctly, they are likely at or above PL4, and therefore not in need of further testing via Alternate ACCESS.

Therefore, from the available item pool, only items with proficiency levels above PL3 were considered for Alternate Screener. Within that subset, final item selections were guided by how well they matched the ELD standards used in the operational Alternate ACCESS forms, ensuring consistency across domains and grade clusters.

5. Test Administration

Administration of Alternate Screener follows the same procedures used for the WIDA Alternate ACCESS assessment. The only significant difference is that Alternate Screener does not include the stopping criteria indicated in the Alternate ACCESS assessment; students should complete all items in the WIDA Alternate Screener assessment. Test administrators score the items locally as the test is administered to the student.

For additional information regarding the test administration of the Alternate Screener, review the WIDA Alternate Screener Test Administration Manual and the Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration.

6. Reported Score

After completing testing and documenting scores in the Student Response Booklet, test administrators can access the WIDA Alternate Screener Score Calculator to generate a score report.

Test administrators enter the scores from the score sheets completed during test administration into the score calculator. The official score report—not the score sheets—provides meaningful information about the student's performance and skills in terms of the WIDA Alternate English language proficiency levels. Proficiency level scores in the score report are intended to support identification and placement decisions of students.

Alternate Screener reports whole-number scores based on the domains that are administered. Proficiency levels are reported on a scale of 3 through 5. If a student receives a score that is less than a proficiency level (PL) 3, the score report will note "<3." Score calculations are adjusted for different expectations in the 6–8 grade cluster versus the 9–12 grade cluster.

If four domains are administered, students receive scores for each of the four domains, as well as composite scores for Oral Language, Literacy, Comprehension, and Overall.

If only Listening and Speaking domains are administered, students will receive the following scores:

- Listening domain score
- Speaking domain score
- Oral Language composite score

If only Reading and Writing domains are administered, students will receive the following scores:

- Reading Score
- Writing Score
- Literacy composite score

If only Speaking and Writing domains are administered, students will receive the following scores:

- Speaking Score
- Writing Score

If only Reading and Listening domains are administered, students will receive the following scores:

- Reading score
- Listening score
- Comprehension composite score

For any domains not administered and not entered into the WIDA Alternate Screener Score Calculator, the score report will note "Not Tested" for the domain. For any composite scores that include domains not tested, the score box will not display a numerical score.

There are no score caps applied to any domain test on Alternate Screener. Students may score up to proficiency level 5 on all domains and composite scores. The lowest score that students may receive is proficiency level <3.

WIDA recommends using the Overall score for decision-making purposes, where all four domains have been administered. Caution is advised when interpreting domain-specific scores on Alternate Screener due to the limited number of items in each domain. The specific Alternate Screener scores used in identification are determined at the state level.

7. Analysis

7.1 Data

Since Alternate Screener has not yet been administered, the 602 Field Test data administered in 2023, which serves as the basis for the Alternate Screener items, is used to conduct the item statistics and Differential Item Functioning (DIF) analysis.

7.2 Measurement Models

7.2.1 Rasch Model for Scoring

The measurement model that forms the basis of the analysis for the development of Alternate Screener ACCESS is the Rasch measurement model (Wright and Stone, 1979). Additional information on its use in the development of the test is available in the Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration. The test was developed using Rasch measurement principles, and in this sense the Rasch model guided all decisions throughout the development of the assessment and was not merely a tool for the statistical analysis of the data. For example, data based on Rasch fit statistics guided the inclusion, revision, or deletion of items during the development and field testing of the test forms and will continue to guide the refinement and further development of the test. For all domains, a Rasch Rating Scale model was used. Mathematically, this can be represented as

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - F_k$$

where

P_{nik} = probability of person *n* on task *i* receiving a rating at level *k* on the rating scale.

P_{nik-1} = probability of person *n* on task *i* receiving a rating at level *k* - 1 on the rating scale (i.e., the next lowest rating).

B_n = ability of person *n*.

D_i = difficulty of task *i*.

F_k = calibration of step *k* on the rating scale.

All Rasch analyses were conducted using the Rasch measurement software program, Winsteps 3.92.1 (Linacre, 2006). When speaking of the measure of student ability, we use the term "ability measure," rather than "theta," used commonly when discussing models based on item response theory. When speaking of the measure of how difficult an item is, we use the term item "difficulty measure," rather than "b parameter," used commonly when discussing models based on item response theory. Step measures refer to the calibration of the steps in the Rasch

rating scale model previously presented. All three measures (ability, difficulty, and step) are expressed in terms of Rasch logits, which are then converted into scores on the Alternate ACCESS score scale for reporting purposes.

Fit statistics for the Rasch model are calculated by comparing the observed empirical data with the data that the Rasch model would be expected to produce if the data fit the model perfectly. Outfit mean square statistics for items and tasks are influenced by outlier ratings for rater-scored performance tasks. For example, a difficult item that some low-ability students answer correctly, for reasons unknown, will have a high outfit mean-square statistic. Similarly, an easy item that some high-ability students answer incorrectly will also have a high outfit mean-square statistic. Infit mean-square statistics are influenced by unexpected patterns of students' responses and ratings on items and tasks that are roughly targeted for them and generally indicate a more serious measurement problem. The expectation for both statistics is 1.00, and values near 1.00 are not of great concern. Values less than 1.00 indicate that the response and rating patterns are too predictable and thus redundant, or the model is overfitting the data, but are not of great concern. High values are of greater concern.

Linacre (2002) provided more guidance on how to interpret these statistics for dichotomous items:

- Values greater than 2.0 "distort or degrade the measurement system."
- Values between 1.5 and 2.0 are "unproductive for construction of measurement, but not degrading,"
- Values between 0.5 and 1.5 should be considered "productive for measurement."
- Values below 0.5 are considered "less productive for measurement, but not degrading,"

Linacre also stated that infit problems are more serious to the construction of measurement than outfit problems. Because conservative guidelines were followed in the development of Alternate ACCESS, 85% of the test items have infit statistics within the range of 0.5 to 1.5, aligning with the standards for being "productive for measurement" as defined by the guidelines.

7.3 Item Analyses

Section 7.3 offers a comprehensive summary of the item analyses, which are divided into two parts: 1) the Step Value Summary and 2) the Detailed Item Summary.

In Part 1, the Step Value Summary provides an overview of the items on the test form. The first column lists the possible raw scores, followed by the frequency of each score derived from FT data, in the second column. The third column presents the step value measure, while the fourth and fifth columns show the fit indices. For step value estimates, the fit indices typically hover around 1.0 for both infit and outfit statistics across all domains, indicating good model-data fit.

Part 2 contains a detailed table summarizing the analyses of all items or tasks on the test form. The first column provides descriptive names for each item, and the second column indicates the cluster to which the item belongs. The third column describes each item in greater detail, with names that include characters representing the domain (e.g., R for Reading), the target language proficiency level (e.g., P2), the language standard (e.g., LA), a keyword for the item's theme (e.g., Ball), and a numeric item ID (e.g., 21603). The fourth column lists the expected Proficiency Level (PL), and the fifth column specifies the corresponding language standard. The sixth column shows the item's difficulty measure in logits. In the seventh column, the average raw score for polytomous items is reported, reflecting task difficulty, with higher values indicating easier tasks. The eighth and ninth columns present the Rasch item fit statistics, including infit and outfit measures, which are used to assess model-data alignment. The final column presents the point measure correlation for each item, a statistic that evaluates how effectively an item differentiates between high- and low-performing test-takers. This correlation is a key indicator of the item's contribution to the overall reliability and validity of the test.

The results indicate that nearly all items and tasks (96.5%) exhibit infit mean square statistics below 1.0 across all grade-level clusters and domains. This suggests that these items and tasks consistently measure ability within the targeted ability range. As noted earlier, the outfit mean-square statistic is particularly sensitive to outlier responses or scores that fall outside the targeted ability range. Notably, no items have outfit mean-square statistics exceeding 2.0, confirming that no significant outliers were present in the data.

7.3.1 Listening

Table 7.3.1.1

Threshold Summary: List

Raw Score	Freq	Threshold	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics

Information withheld due to confidentiality requirements.

Table 7.3.1.2

Complete Item Analysis: List

Item #	Cluster	UIN	Expected PL	Standard	Item Difficulty (in logits)	P-value	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics	Point Measure

Information withheld due to confidentiality requirements.

7.3.2 Reading

Table 7.3.2.1

Threshold Summary: Read

Raw Score	Freq	Threshold	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics

Information withheld due to confidentiality requirements.

Table 7.3.2.2

Complete Item Analysis: Read

Item #	Cluster	UIN	Expected PL	Standard	Item Difficulty (in logits)	P-value	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics	Point Measure

Information withheld due to confidentiality requirements.

7.3.3 Speaking

Table 7.3.3.1

Threshold Summary: Spek

Raw Score	Freq	Threshold	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics

Information withheld due to confidentiality requirements.

Table 7.3.3.2

Complete Item Analysis: Spek

Item #	Cluster	UIN	Expected PL	Standard	Item Difficulty (in logits)	P-value	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics	Point Measure

Information withheld due to confidentiality requirements.

7.3.4 Writing

Table 7.3.4.1

Threshold Summary: Writ

Raw Score	Freq	Threshold	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics

Information withheld due to confidentiality requirements.

Table 7.3.4.2

Complete Item Analysis: Writ

Item #	Cluster	UIN	Expected PL	Standard	Item Difficulty (in logits)	P-value	Infit Mnsq Fit Statistics	Outfit Mnsq Fit Statistics	Point Measure

Information withheld due to confidentiality requirements.

7.4 Scaling

The table below provides the scaling equation for each domain. This equation is used to convert an examinee's ability measure into the scale score. Each equation is used across all grade-level clusters within each domain. For detailed scaling procedures, refer to section 4.4, Scaling, in the *Annual Technical Report for the WIDA Alternate Access English Language Proficiency Test, Series 602, 2023–2024 Administration*.

Table 7.4.1

Scaling equation by domain

Domain	Scale Score
Listening	(Ability Measure in Logits × 7.948) + 942.606
Reading	(Ability Measure in Logits × 7.495) + 940.879
Speaking	(Ability Measure in Logits × 7.678) + 941.392
Writing	(Ability Measure in Logits × 7.297) + 943.625

7.5 Differential Item Functioning

Differential Item Functioning (DIF) analysis aims to determine whether item or task performance is influenced by factors unrelated to English language proficiency, the construct being measured by the test. Essentially, DIF analysis seeks to identify items that may function differently for various groups due to irrelevant characteristics. For the Alternate ACCESS, student performance was compared across four groupings: (1) males versus females, (2) Hispanic versus non-Hispanic ethnic backgrounds, (3) race (Hispanic versus individual ethnic groups), and (4) primary disabilities. Students with missing test scores, gender, or ethnicity were excluded from the analysis. For gender and Hispanic vs. non-Hispanic and race DIF analysis, male and Hispanic groups serve as reference groups. For disability DIF analysis, multiple group performances were compared against the overall performance simultaneously, rather than setting one group as a reference group and conducting multiple pairwise comparisons.

To ensure sufficient sample sizes within racial groups and disability categories, the analysis included the four largest racial groups: Hispanic, White, Black, and Asian. For disability categories, groups with fewer than 100 students were aggregated, while those with 100 or more students were analyzed separately. This approach ensured robust and reliable detection of DIF across all examined subgroups.

A multiple-group analysis was used for DIF detection within the context of rating scale models, which Alternate ACCESS employs. This approach is an extension of the item response theory (IRT) model to multiple groups and is preferred due to its flexibility in assessing the invariance

of item properties such as discrimination and difficulty (Tay et al., 2015). For DIF detection, rating scale models are estimated separately for each group with constraints. To identify DIF, one item difficulty of one group (the focal group) is compared to that of the reference group, while keeping all other difficulties consistent across groups. If the difference is statistically significant, that item exhibits DIF for the corresponding source.

Winsteps provides two types of DIF contrasts: (1) a paired DIF effect between two specific groups, with the hypothesis that an item has the same difficulty across the groups, and (2) a contrast between a specific group and the overall average difficulty across all groups, with the hypothesis that an item's difficulty is equal to its average difficulty across groups. For gender and ethnicity, the first type was used, with the male and Hispanic groups as the references. The five racial groups—White (W), Black (B), Asian (AS), American Indian/Alaskan Native (AI), and Pacific Islander/Hawaiian (PI)—are compared to the Hispanic group, which serves as the reference group. For types of disabilities, the second type was employed since there was no specific reference group; instead, the item difficulty for each disability group was compared against the overall average difficulty for each item.[1]

Following guidelines by ETS for NAEP assessment (Allen, Carlson, & Zalanak, 1999), Alternate ACCESS tasks are classified into three DIF levels:

- AA (no DIF), when the Rasch-Welch Chi-square statistic is not significant or when it is significant and |DIF| is less than 0.43 logits
- BB (weak DIF), when the Rasch-Welch Chi-square statistic is significant and |DIF| is greater than or equal to 0.43 but less than 0.64 logits
- CC (strong DIF), when the Rasch-Welch Chi-square statistic is significant and |DIF| is greater than or equal to 0.64 logits

Note: ETS uses Delta units, where 1 Delta unit is equivalent to 0.426 logits.

The following tables are organized into four sections, divided by domains and clusters:

1. Overall DIF Summary. This section provides a summary of the number of items identified with DIF across the three levels (AA, BB, or CC) for gender, ethnicity, and disabilities. For disabilities, each item may exhibit at least five DIF effects due to

[1] For DIF analysis, students are categorized by their primary disabilities as defined under IDEA (https://sites.ed.gov/idea/regs/b/a/300.8). These categories include Autism Spectrum Disorder (AS), Deaf-blindness (DB), Developmental Delay (DD), Hearing Impairment including Deafness (HI), Infant/Toddler with a Disability (ITD), Intellectual Disability (ID), Multiple Disabilities (MD), Orthopedic Impairment (OI), Other Health Impairment (OHI), Serious Emotional Disability (SED), Specific Learning Disability (SLD), Speech or Language Impairment (SLI), Traumatic Brain Injury (TBI), and Visual Impairment including Blindness (VI). Students who do not report a disability type are classified as having "No Primary Disability recorded" (NPD).

multiple comparisons among disability groups. This highlights the complexity of DIF analysis for this population, given the variety of group comparisons involved.

2. DIF analysis for gender and ethnicity. This section details the DIF results for individual items. The second and fourth columns indicate the DIF level (AA, BB, or CC) for gender and ethnicity, respectively. The third and fifth columns identify items that favor one group over the other at each DIF level. Ideally, even when all items fall into the AA category, there should be a relatively even distribution of items favoring each group to ensure there is no systematic bias in the test.
3. DIF analysis for race. This section presents a breakdown of the DIF results across different ethnic groups. It provides detailed insights into item performance and potential bias related to individual ethnicity.
4. DIF analysis for disability. This section focuses on DIF results for different disability categories. It provides a closer examination of item performance across various disability groups, ensuring that the test is equitable and free from bias across these sub-populations.

The DIF analysis spans four key domains, (Listening, Reading, Speaking, and Writing) and investigates potential bias across gender, ethnicity, race, and disability groups. DIF levels are categorized as A (negligible), B (moderate), and C (large).

No B- or C-level DIF was detected for gender or ethnicity. However, several items flagged B- and C-level DIF for racial and disability groups across all domains.

Among racial groups, students identifying as American Indian or Pacific Islander exhibited several items with B- and C-level DIF. In the Listening domain, two C-level DIFs were identified: Item 3 (K2) flagged for American Indian participants and Item 6 (35) for Pacific Islander participants. The Reading domain presented three C-level DIFs, all within the American Indian group, appearing in Items 3 (K2), 5 (35), and 9 (612). In the Speaking domain, three C-level DIFs were noted: two for American Indian participants (Items 4 [35] and 5 [612]) and one for Pacific Islander participants (Item 4 [35]). The Writing domain showed two additional C-level DIFs: one for the American Indian group and one for the Pacific Islander group, both associated with K2 items. Overall, the American Indian group had the highest number of flagged items. Notably, the sample sizes for these racial subgroups were relatively small, fewer than 100 participants each. As a result, the stability and reliability of these findings may be limited. This caveat should be kept in mind when interpreting the results.

For disability groups, several C-level DIFs were identified across the domains. In the Listening domain, one C-level DIF was observed for both the "Other" and Speech or Language Impairment (SLI) groups in Item 1 (K2), with both items favoring the baseline group. The Reading domain revealed a single C-level DIF for the No Primary Disability (NPD) group, along with two additional C-level DIFs—one for the Specific Learning Disability (SLD) group and one

for the SLI group—appearing in Item 3 (35) and Item 7 (612), respectively. These two items favored the corresponding disability groups over the baseline. In the Speaking domain, two C-level DIFs were flagged: one for the SLI group in Item 2 (K2), and another for the "Other" group in Item 6 (612); both items favored the baseline group. Lastly, in the Writing domain, two C-level DIFs were detected—one for the Multiple Disabilities (MD) group and one for the SLD group in Item 1 (K2). The item favored the MD group when compared to the baseline, while it favored the baseline group in comparison to the SLD group.

7.5.1 Listening

DIF Analysis: List

Table 7.5.1.1

Overall DIF Summary: List

DIF Level	Gender	Ethnicity	Race	Disability

Information withheld due to confidentiality requirements.

Table 7.5.1.2

DIF Summary for Gender and Ethnicity: List

Item #	M/F DIF Level	M/F Favored Group	H/O DIF Level	H/O Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.1.3

DIF Analysis for Race: List

Item #	W/H DIF Level	W/H Favored Group	B/H DIF Level	B/H Favored Group	A/H DIF Level	A/H Favored Group	AI/H DIF Level	AI/H Favored Group	PS/H DIF Level	PS/H Favored Group

Note: Since there were no students in the group for Item 1, no DIF test was administered.

Information withheld due to confidentiality requirements.

Table 7.5.1.4

DIF Analysis for Disability: List

Item #	AD DIF Level	AD Favored Group	DD DIF Level	DD Favored Group	ID DIF Level	ID Favored Group	MD DIF Level	MD Favored Group	OHI DIF Level	OHI Favored Group	SLD DIF Level	SLD Favored Group	SLI DIF Level	SLI Favored Group	Other DIF Level	Other Favored Group

Note: Groups with sample sizes less than 100 are combined into "Other"; "NPD" is excluded from the table.

Information withheld due to confidentiality requirements.

7.5.2 Reading

DIF Analysis: Read

Table 7.5.2.1

Overall DIF Summary: Read

DIF Level	Gender	Ethnicity	Race	Disability

Information withheld due to confidentiality requirements.

Table 7.5.2.2

DIF Summary for Gender and Ethnicity: Read

Item #	M/F DIF Level	M/F Favored Group	H/O DIF Level	H/O Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.2.3

DIF Analysis for Race: Read

Item #	W/H DIF Level	W/H Favored Group	B/H DIF Level	B/H Favored Group	A/H DIF Level	A/H Favored Group	AI/H DIF Level	AI/H Favored Group	PS/H DIF Level	PS/H Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.2.4

DIF Analysis for Disability: Read

Item #	AD DIF Level	AD Favored Group	DD DIF Level	DD Favored Group	ID DIF Level	ID Favored Group	MD DIF Level	MD Favored Group	OHI DIF Level	OHI Favored Group	SLD DIF Level	SLD Favored Group	SLI DIF Level	SLI Favored Group	Other DIF Level	Other Favored Group

Note: Groups with sample sizes less than 100 are combined into "Other"; "NPD" is excluded from the table.

Information withheld due to confidentiality requirements.

7.5.3 Speaking

DIF Analysis: Spek

Table 7.5.3.1

Overall DIF Summary: Spek

DIF Level	Gender	Ethnicity	Race	Disability

Information withheld due to confidentiality requirements.

Table 7.5.3.2

DIF Summary for Gender and Ethnicity: Spek

Item #	M/F DIF Level	M/F Favored Group	H/O DIF Level	H/O Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.3.3

DIF Analysis for Race: Spek

Item #	W/H DIF Level	W/H Favored Group	B/H DIF Level	B/H Favored Group	A/H DIF Level	A/H Favored Group	AI/H DIF Level	AI/H Favored Group	PS/H DIF Level	PS/H Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.3.4

DIF Analysis for Disability: Spek

Item #	AD DIF Level	AD Favored Group	DD DIF Level	DD Favored Group	ID DIF Level	ID Favored Group	MD DIF Level	MD Favored Group	OHI DIF Level	OHI Favored Group	SLD DIF Level	SLD Favored Group	SLI DIF Level	SLI Favored Group	Other DIF Level	Other Favored Group

Note. Groups with sample sizes less than 100 are combined into "Other"; "NPD" is excluded from the table.

Information withheld due to confidentiality requirements.

7.5.4 Writing

DIF Analysis: Writ

Table 7.5.4.1

Overall DIF Summary: Writ

DIF Level	Gender	Ethnicity	Race	Disability

Information withheld due to confidentiality requirements.

Table 7.5.4.2

DIF Summary for Gender and Ethnicity: Writ

Item #	M/F DIF Level	M/F Favored Group	H/O DIF Level	H/O Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.4.3

DIF Analysis for Race: Writ

Item #	W/H DIF Level	W/H Favored Group	B/H DIF Level	B/H Favored Group	A/H DIF Level	A/H Favored Group	AI/H DIF Level	AI/H Favored Group	PS/H DIF Level	PS/H Favored Group

Information withheld due to confidentiality requirements.

Table 7.5.4.4

DIF Analysis for Disability: Writ

Item #	AD DIF Level	AD Favored Group	DD DIF Level	DD Favored Group	ID DIF Level	ID Favored Group	MD DIF Level	MD Favored Group	OHI DIF Level	OHI Favored Group	SLD DIF Level	SLD Favored Group	SLI DIF Level	SLI Favored Group	Other DIF Level	Other Favored Group

Note. Groups with sample sizes less than 100 are combined into "Other"; "NPD" is excluded from the table.

Information withheld due to confidentiality requirements.

7.6 Reliability

IRT offers a direct connection between the test information function, $I(\theta)$, and reliability at a given ability level θ. A commonly used formula in IRT expresses reliability as

$$\alpha(\theta) = \frac{I(\theta)}{I(\theta) + 1}$$

where $I(\theta)$ is the test information function (TIF) at a given level of ability (Nicewander, 2018; Raju, Price, Oshima, & Nering, 2006).

In classical test theory (CTT), reliability is a single coefficient, like Cronbach's α, that assumes uniform measurement precision across all students. In contrast, IRT acknowledges that precision varies depending on the student's ability (θ). Rather than one overall reliability value, IRT uses the TIF to indicate how precise (i.e., reliable) the measurement is at each point on the ability scale. High information at a particular θ implies low error and therefore high reliability; where information is lower, reliability declines. This perspective, often called conditional reliability, means that each ability estimate has its own associated precision.

This formula is essentially a transformation of the standard error (SE) of θ. In IRT, the SE of ability estimate is inversely related to information (Lord, 1980; approximately $SE^2 \approx 1/I(\theta)$ for large item pools). Thus, reliability can be defined as $1 - (SE)^2$ on the θ scale. Substituting $SE^2 \approx 1/I(\theta)$ gives $\rho(\theta) \approx 1 - \frac{1}{I(\theta)}$. If the latent trait is scaled to have variance 1, a common IRT normalization, this expression is equivalent to $\frac{I(\theta)}{I(\theta)+1}$. In other words, test information represents a signal-to-noise ratio, and reliability is the signal proportion: info / (info + 1 unit of error variance).

When defining the range of θ, targeted proficiency levels (e.g., P3–P5) are first mapped to corresponding values on the ability scale. This range is then used to evaluate item information across the relevant θ interval including all domains by clusters.

Table 7.6.1 presents the reliability values for proficiency level (PL) ranges. Due to the limited number of items in each domain, reliability is calculated using all four domains (Listening, Reading, Speaking, and Writing) combined for each cluster. Cluster K2 shows a reliability of 0.9048 within the ability range of [0.05, 2.063], while Cluster 35 exhibits a reliability of 0.9018 within the ability range of [0.679, 2.314]. Cluster 68 has a reliability of 0.8969 for the ability range [0.93, 2.44], and Cluster 912 demonstrates a reliability of 0.8769 within the ability range of [1.056, 2.818].

Table 7.6.1

Reliability values for proficiency level ranges

Cluster	PL range	Reliability
K2	[0.050,2.063]	0.9048
35	[0.679,2.314]	0.9018
68	[0.930,2.440]	0.8969
912	[1.056,2.818]	0.8769

7.7 Test Information Function

With the Rasch measurement model, as with any measurement model following IRT, the relationship between the ability measure (in logits) and the accuracy of test scores can be modeled. It is recognized that tests measure most accurately when the abilities of the examinees and the difficulty of the items are most appropriate for each other. If a test is too difficult for an examinee (i.e., the examinee scores close to zero), or if the test is too easy for an examinee (i.e., the examinee "tops out"), accurate measurement of the examinee's ability cannot be made. The TIF shows graphically how well the test is measuring across the ability measure spectrum in terms of measurement error. High values indicate more accuracy in measurement. Thus, for each test form, Figure 7.7.1.1 through Figure 7.7.4.4 shows the relationship between the ability measure (in logits) on the horizontal axis and measurement accuracy, represented as the Fisher information value (which is the inverse squared of the standard error), on the vertical axis. The TIF then reflects the conditional standard error of measurement.

The TIF is an advanced IRT concept. It is significant, mainly because it provides indices analogous to reliability and standard error of measurement (SEM) in the classical test theory. Without using statistical formulations, we can conceptualize the idea this way: in a well-designed test, every item responded to correctly provides a bit of information about what a student knows and can do, and every item responded to incorrectly indicates what a student does not know and cannot do. When there are a sufficient number of items, information accumulates to provide an accurate estimate of student ability. In this sense, information is directly related to the reliability of test scores: the more information, the higher the reliability and the smaller the SEM.

Test information varies as a function of student ability. The same test can provide a significant amount of information for some students, but little information for other students. Usually, an achievement assessment is designed for students ranging from relatively low ability to relatively high ability. A student in this range is expected to answer some items correctly and some items incorrectly. However, if a student has an extremely high ability that is far beyond the ability level

required by the test, they might answer all items correctly. This is positive from an educational point of view, but challenging from an ability-estimation standpoint, since this test provides little information about the student's true level of ability. The student has high ability, but there is no way to determine how high. Determining true ability would require the administration of several additional items at the top of the difficulty range. From this example, IRT test information is conditioned on ability. Usually, the test information curve is bell shaped—intermediate abilities provide for the greatest test information and high reliability, whereas extreme abilities correspond to less information and low reliability.

Statistically, at every ability point, the TIF is inversely proportional to the square of the conditional standard error of measurement (CSEM). This relationship is used to calculate the CSEM for each obtainable scale score point. The TIF for the Rating Scale Model (RSM) is defined as follows:

$$I(\theta) = \sum_{i=1}^{L} I_i(\theta)$$

where

$I_i(\theta)$ is $\sum_{k=0}^{m} k^2 P_{ik} - (\sum_{k=0}^{m} k P_{ik})^2$; *i* denotes an item, *k* is k item category; P_{ik} is the probability of scoring *k* on item *i* given θ based on the Rating Scale model; $I(\theta)$ is the quantity of test information at an ability level of θ.

Figure 7.7.1.1 through Figure 7.9.4.2 present the Test Information Curves across domains and clusters. Each figure includes four vertical lines representing the four cut scores, which divide the curve into five sections corresponding to the WIDA language proficiency levels (P1-P5) for the domain being tested. It is imperative that each test form measures most accurately in the areas for which it is primarily used to make classification decisions. In other words, optimally, the TIF should be high for the cuts between P1/P2, P2/P3, P3/P4, and P4/P5.

7.7.1 Listening

Figure 7.7.1.1

Test Information Curve: List K–2

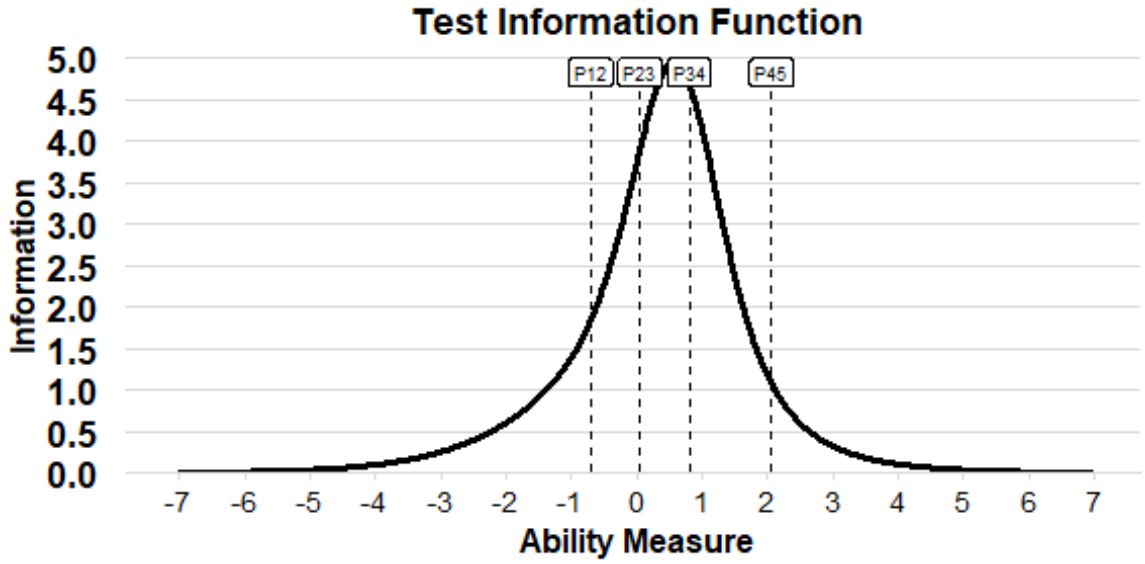


Figure 7.7.1.2

Test Information Curve: List 3–5

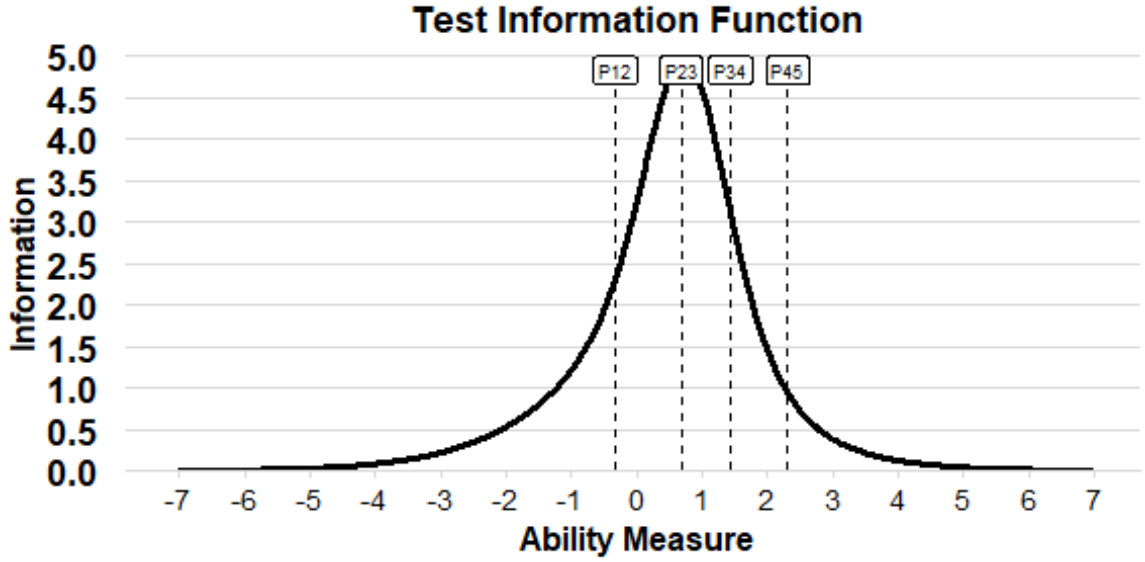


Figure 7.7.1.3

Test Information Curve: List 6–8

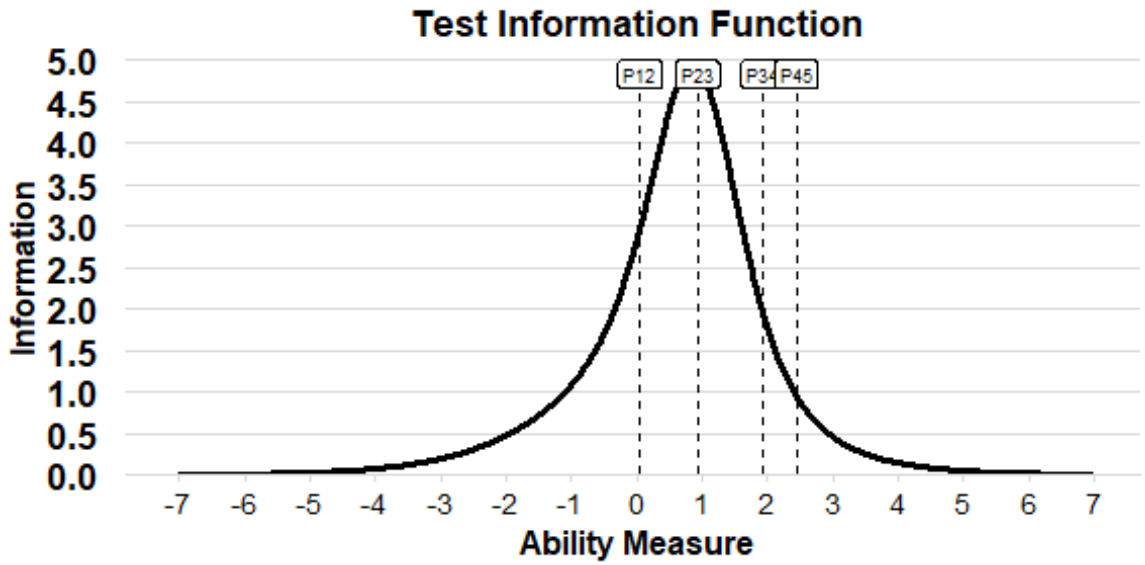
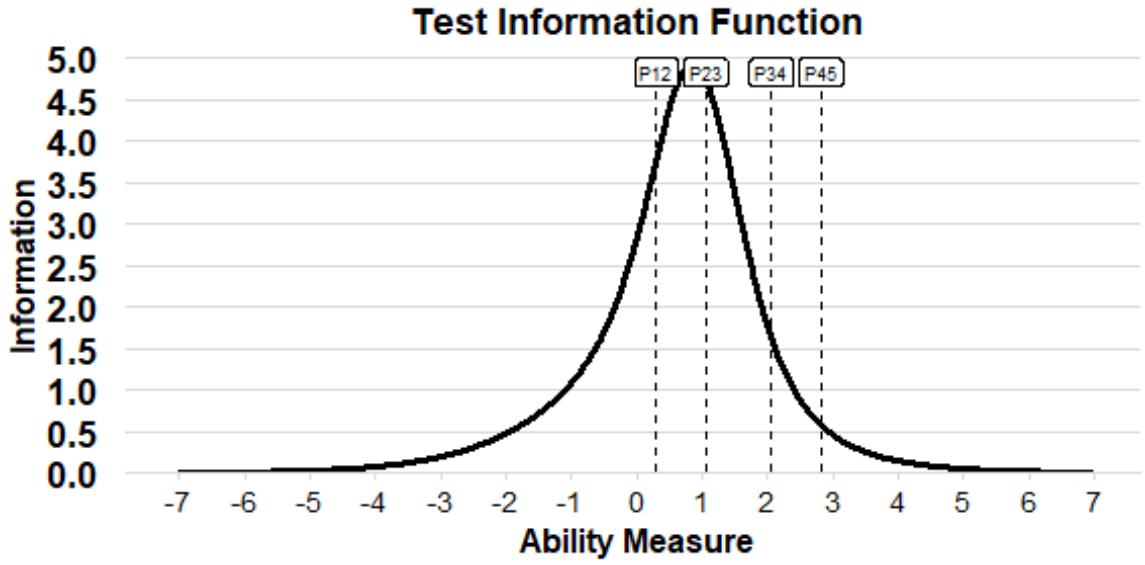


Figure 7.7.1.4

Test Information Curve: List 9–12



7.7.2 Reading

Figure 7.7.2.1

Test Information Curve: Read K–2

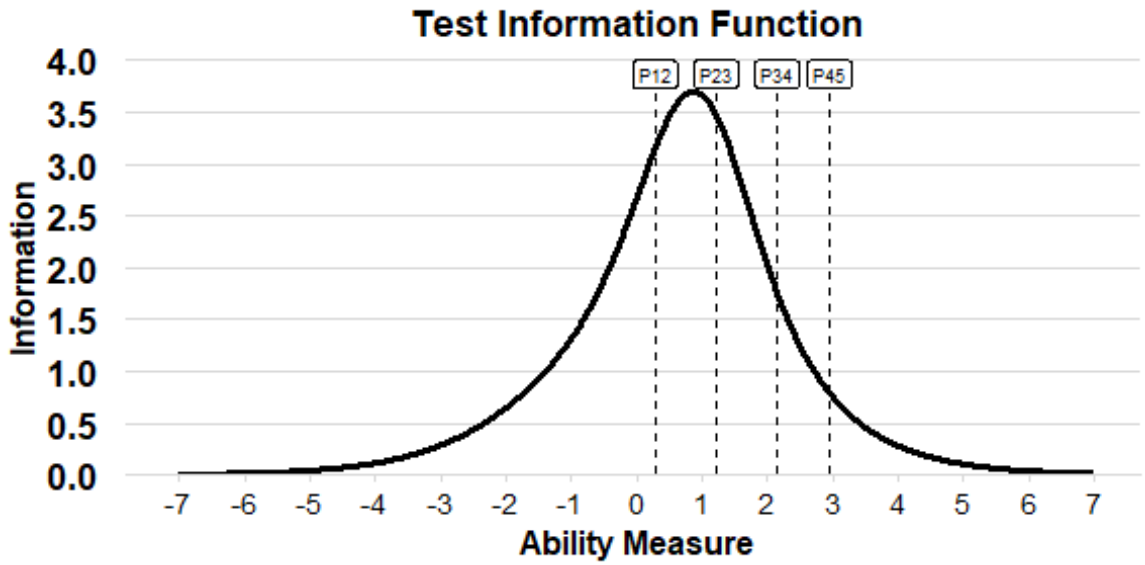


Figure 7.7.2.2

Test Information Curve: Read 3–5

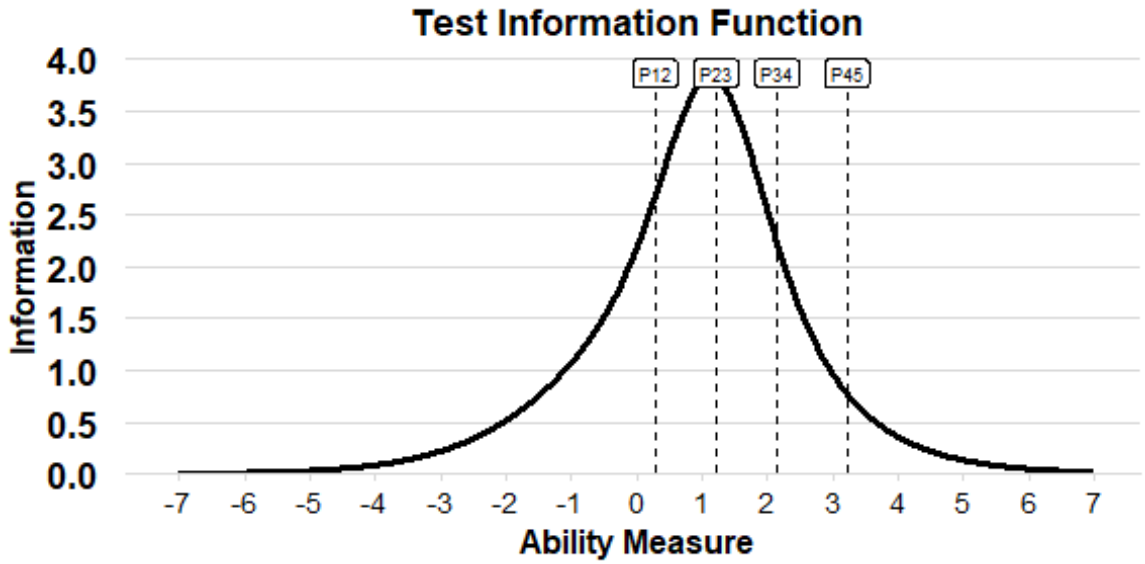


Figure 7.7.2.3

Test Information Curve: Read 6–8

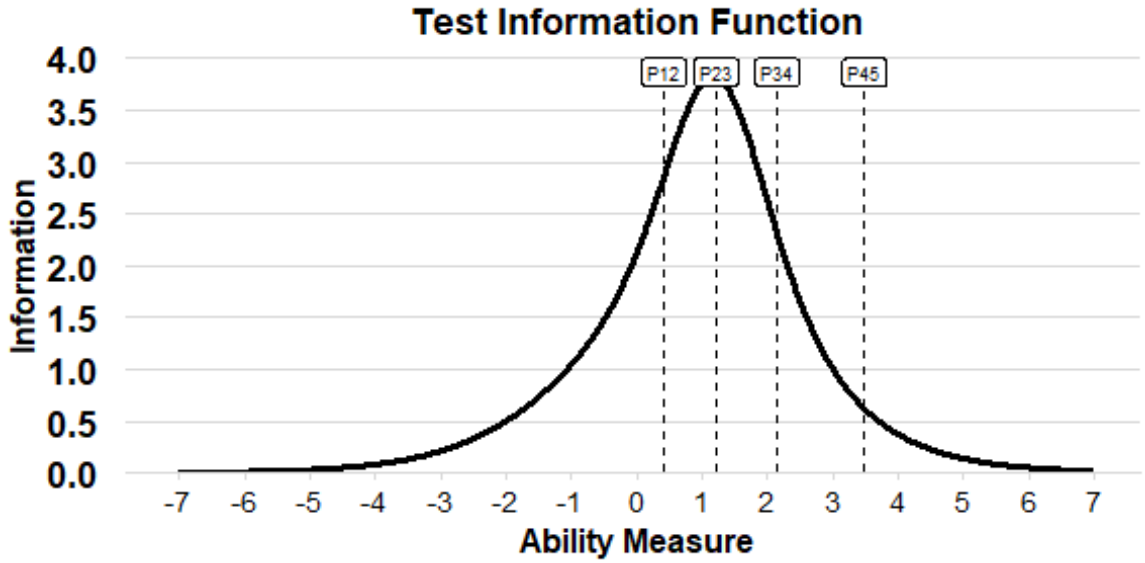
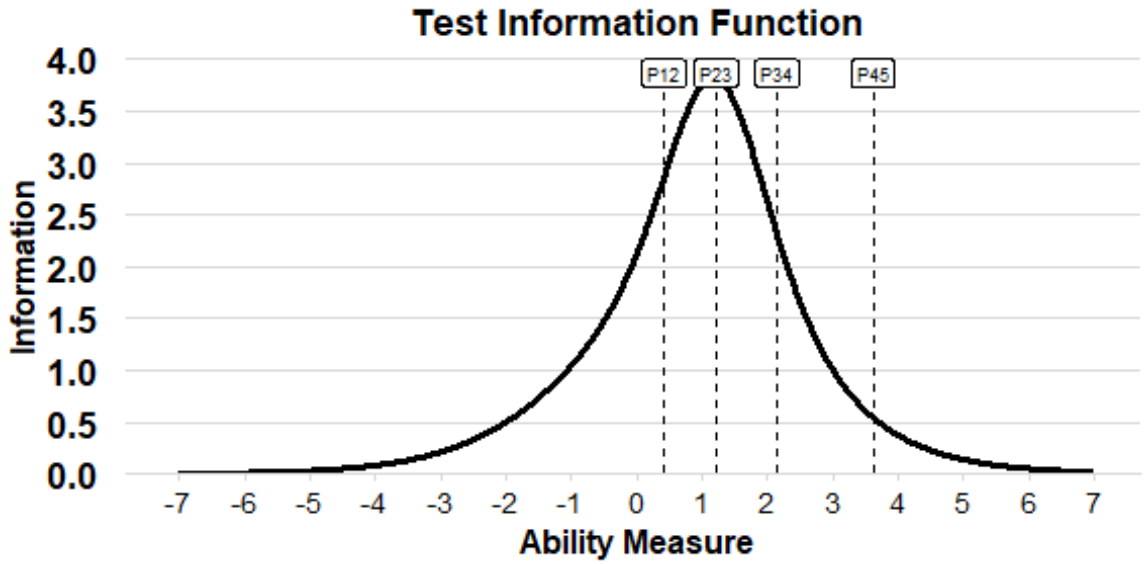


Figure 7.7.2.4

Test Information Curve: Read 9–12



7.7.3 Speaking

Figure 7.7.3.1

Test Information Curve: Spek K–2

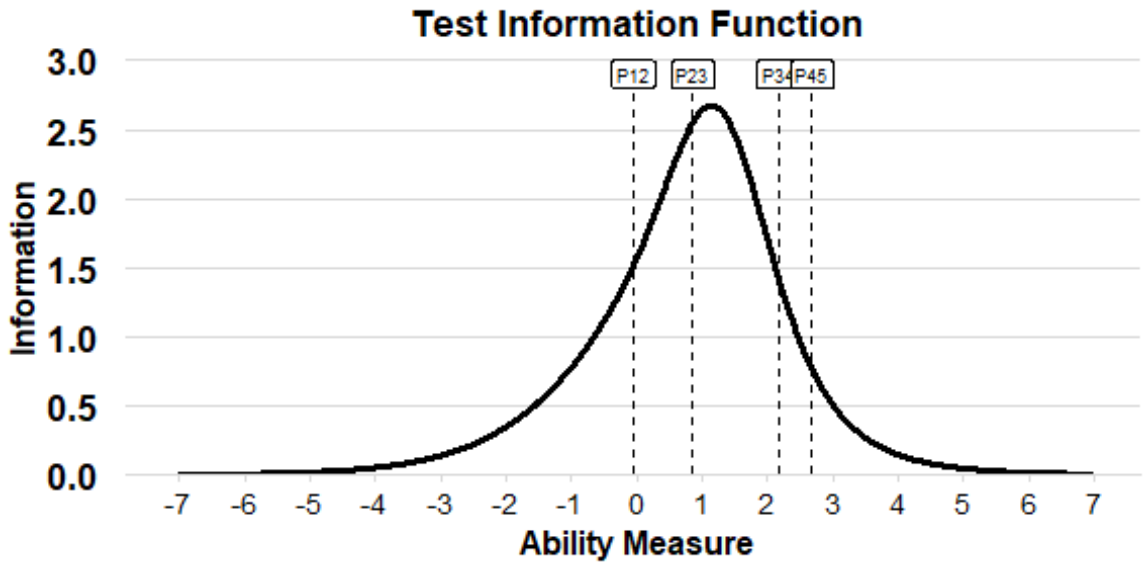


Figure 7.7.3.2

Test Information Curve: Spek 3–5

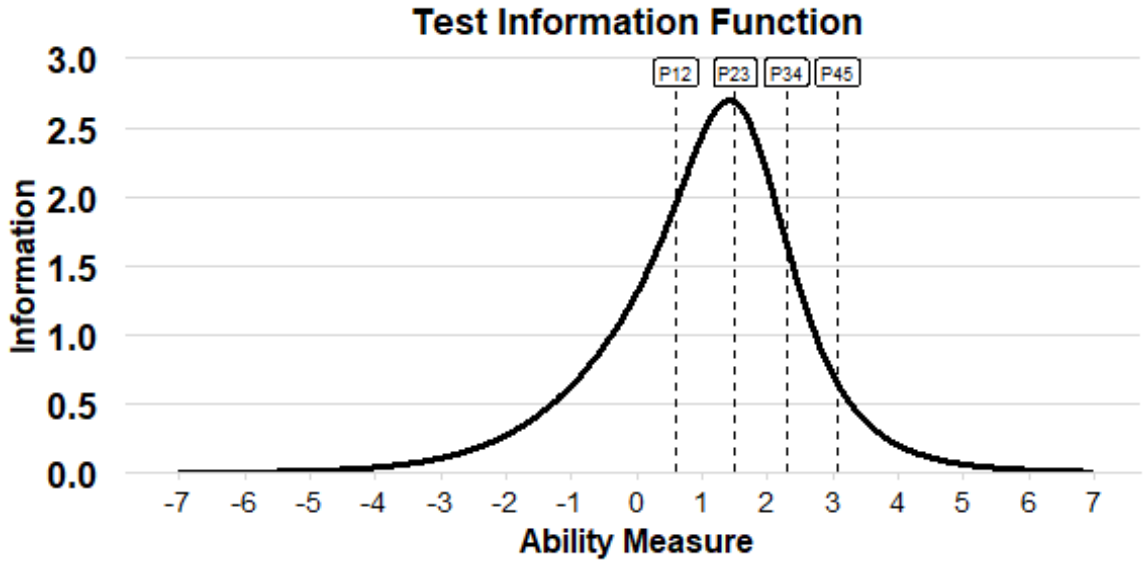


Figure 7.7.3.3

Test Information Curve: Spek 6–8

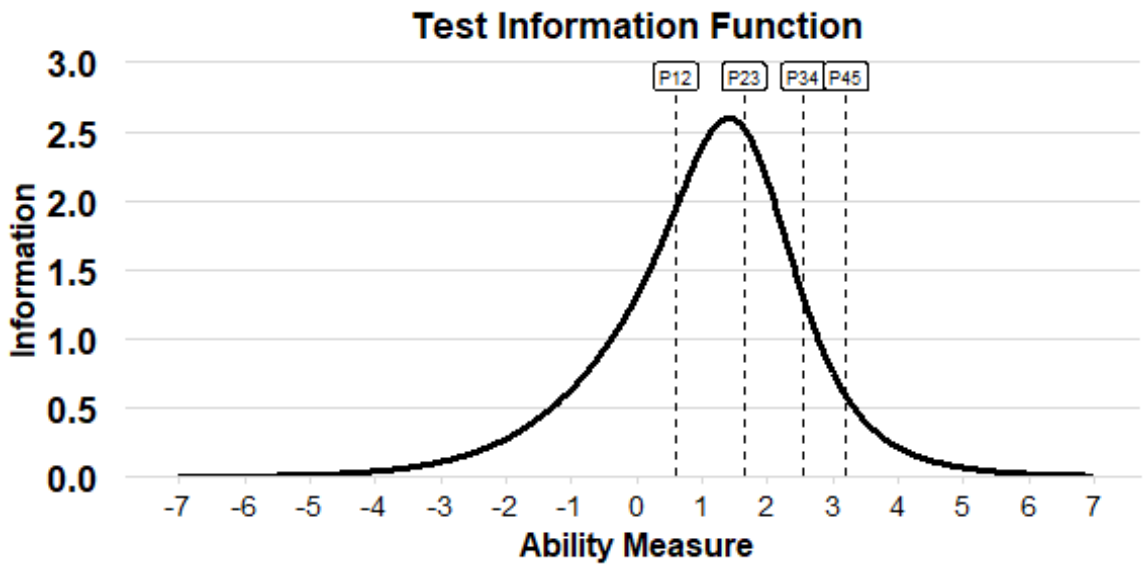
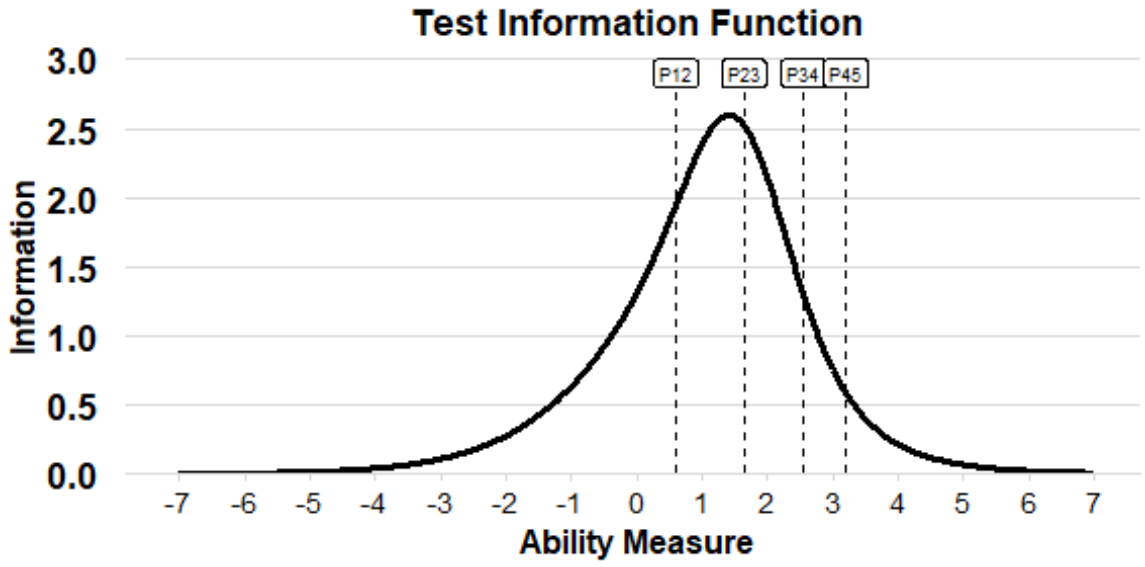


Figure 7.7.3.4

Test Information Curve: Spek 9–12



7.7.4 Writing

Figure 7.7.4.1

Test Information Curve: Writ K–2

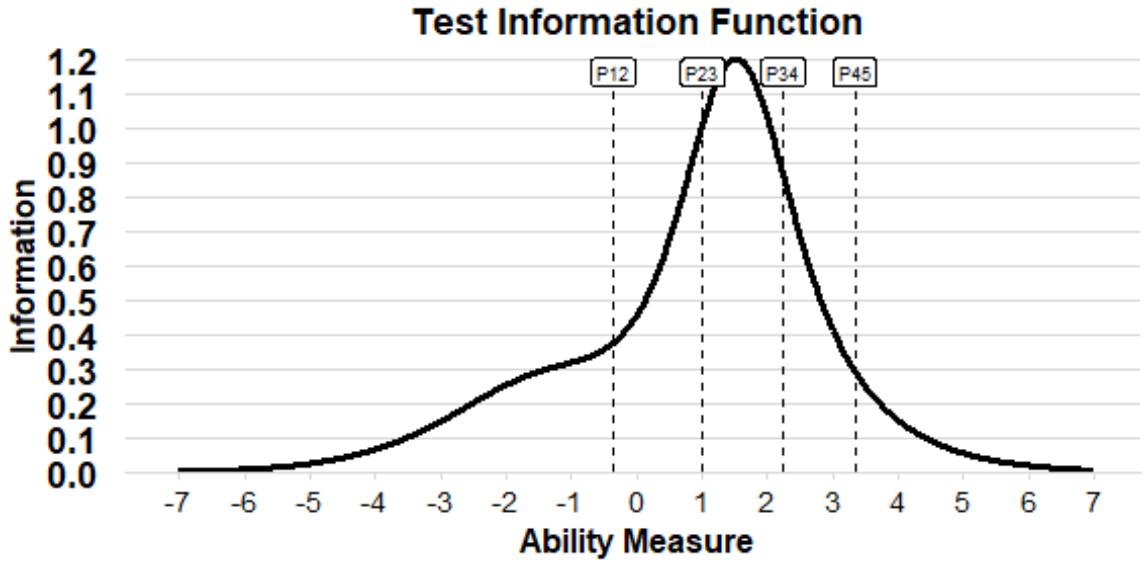


Figure 7.7.4.2

Test Information Curve: Writ 3–5

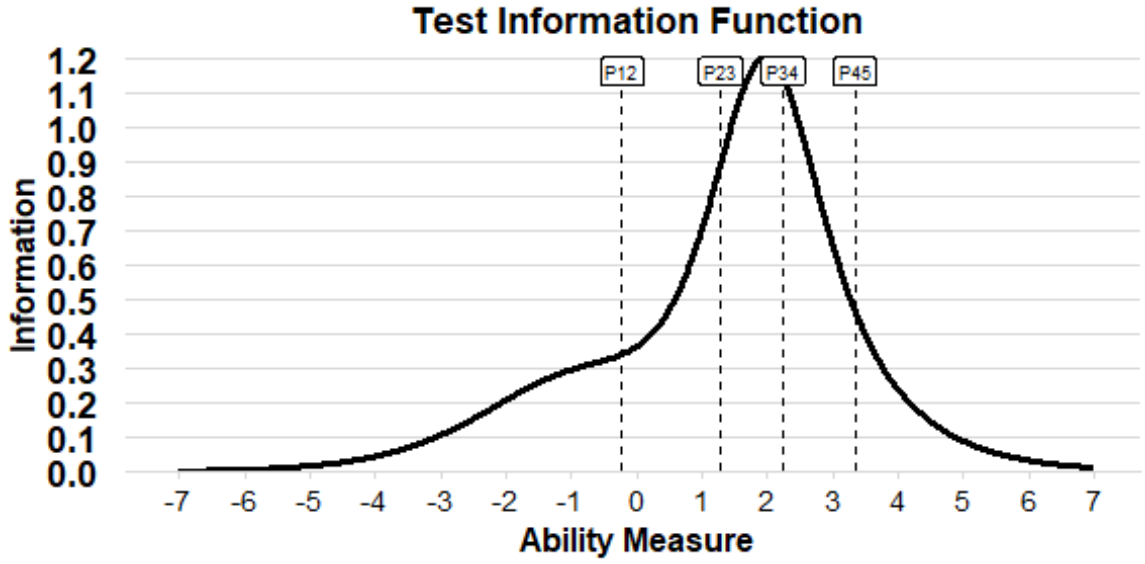


Figure 7.7.4.3

Test Information Curve: Writ 6–8

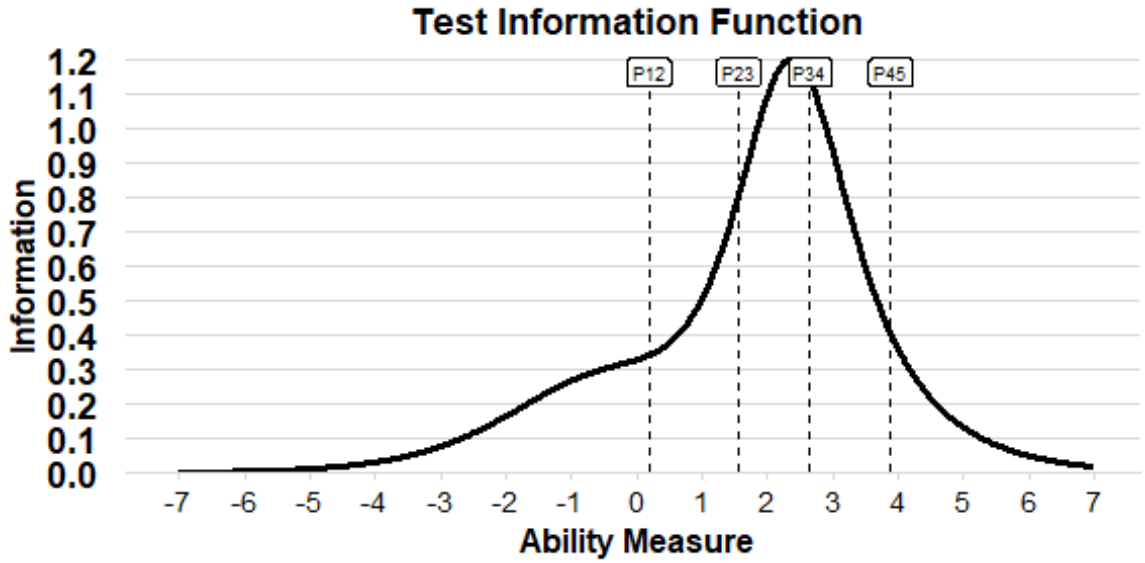
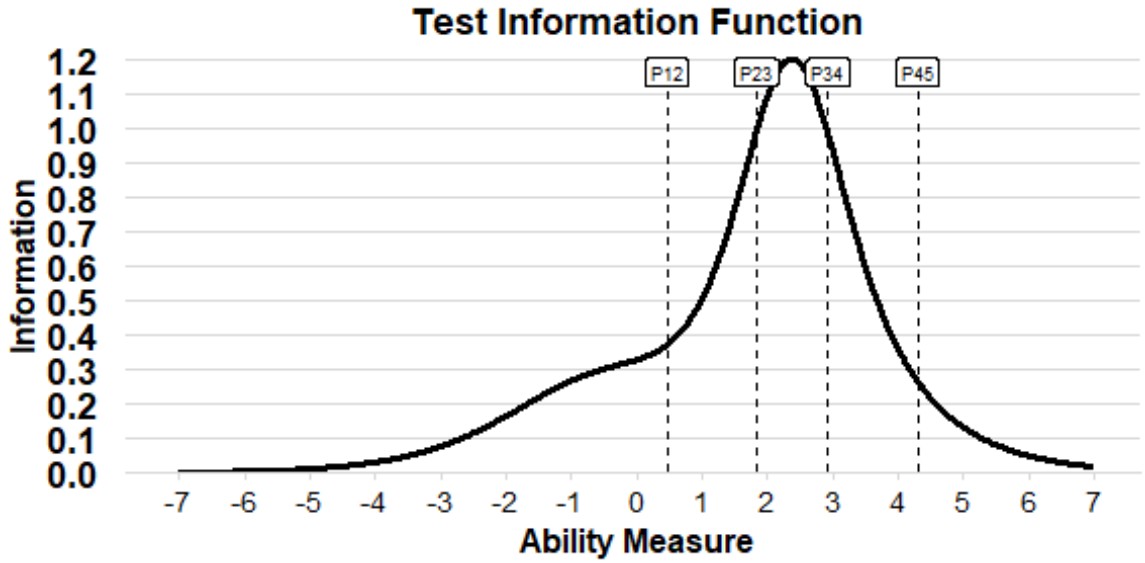


Figure 7.7.4.4

Test Information Curve: Writ 9–12



7.8 Test Characteristic Curve

For each test form, the test characteristic curve graphically shows the relationship between the ability measure (in logits) on the horizontal axis and the expected raw score on the vertical axis. Four vertical lines indicate the four cut scores, dividing the figure into five sections for each of the WIDA proficiency levels (P1-P5) for the domain being tested. As would be expected, higher raw scores are expected by higher ability levels to be placed into higher language proficiency levels. The relative width of each section between the cut score lines, however, gives an indication of how many points must be earned to be placed into a WIDA language proficiency level.

In item response theory, the definition of an expected score according to Andrich (1978) is used. The formula for a true score is given in the equation below:

$$Expected\ Score(\theta_n) = \sum_{i=1}^{I} \left[\sum_{k=0}^{K} [k \times P_{nik}] \right]$$

where

n is an examinee, *i* denotes an item, and *k* is k item category; P_{nik} is the probability of person *n* scoring *k* on item *i* based on the Rating Scale model; ES_n is the expected score for an examinee with ability level θ_n.

Figures 7.8.1.A through 7.8.4.D present the Test Characteristic Curves across domains and clusters. As with the Test Information Curves, four vertical lines represent the WIDA proficiency level cut scores, dividing each curve into five categories.

7.8.1 Listening

Figure 7.8.1.1

Test Characteristic Curve: List K–2

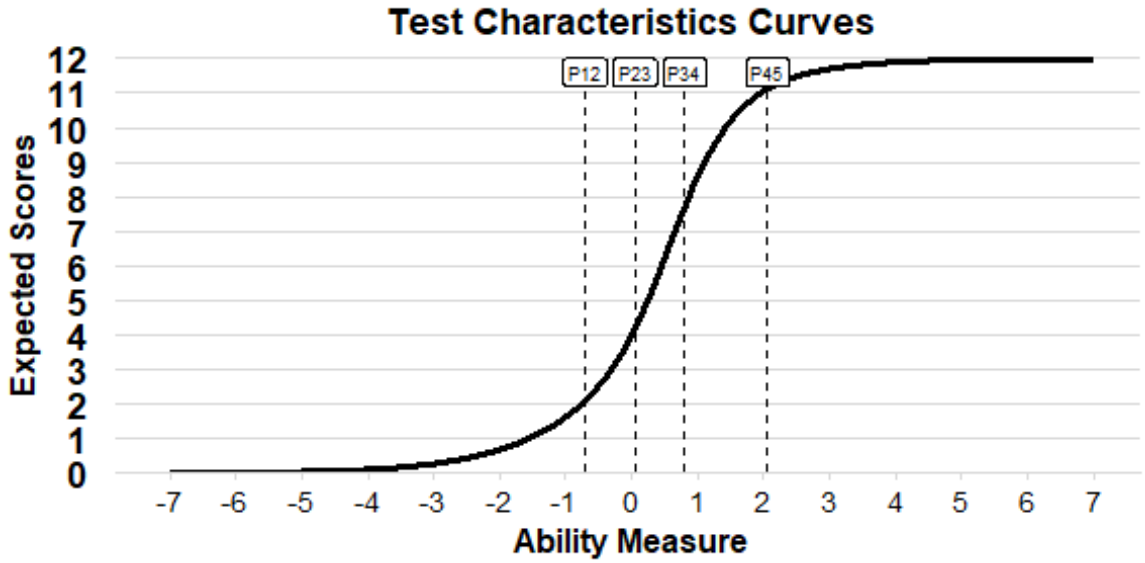


Figure 7.8.1.2

Test Characteristic Curve: List 3–5

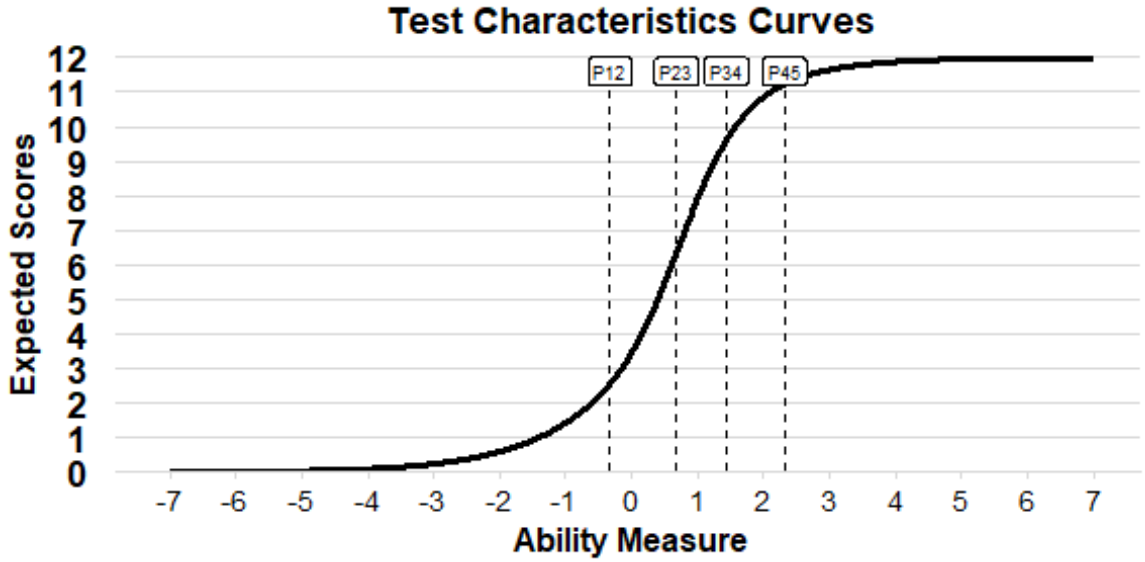


Figure 7.8.1.3

Test Characteristic Curve: List 6–8

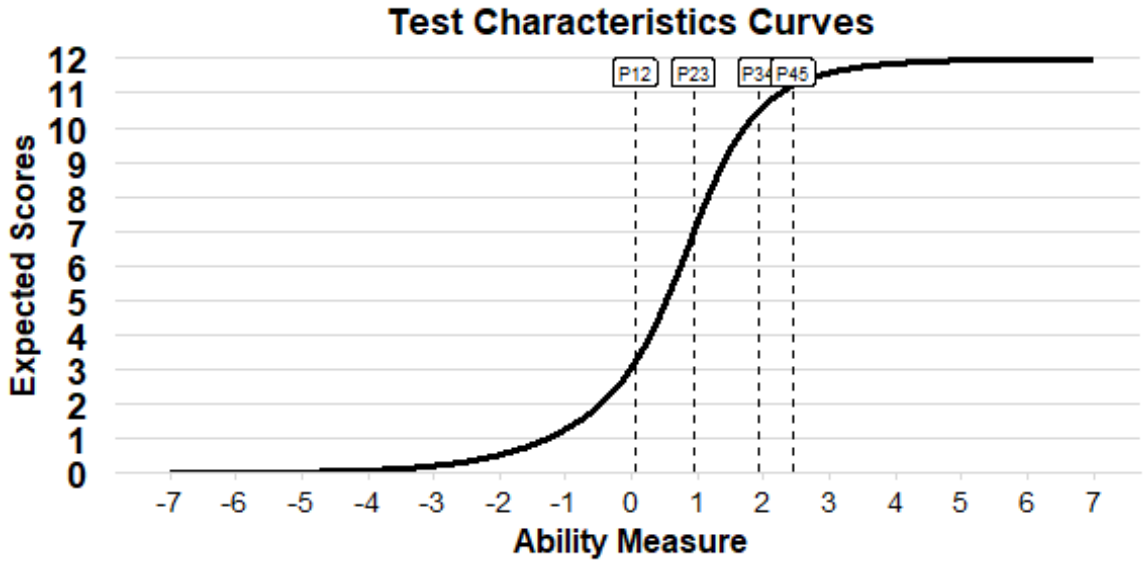
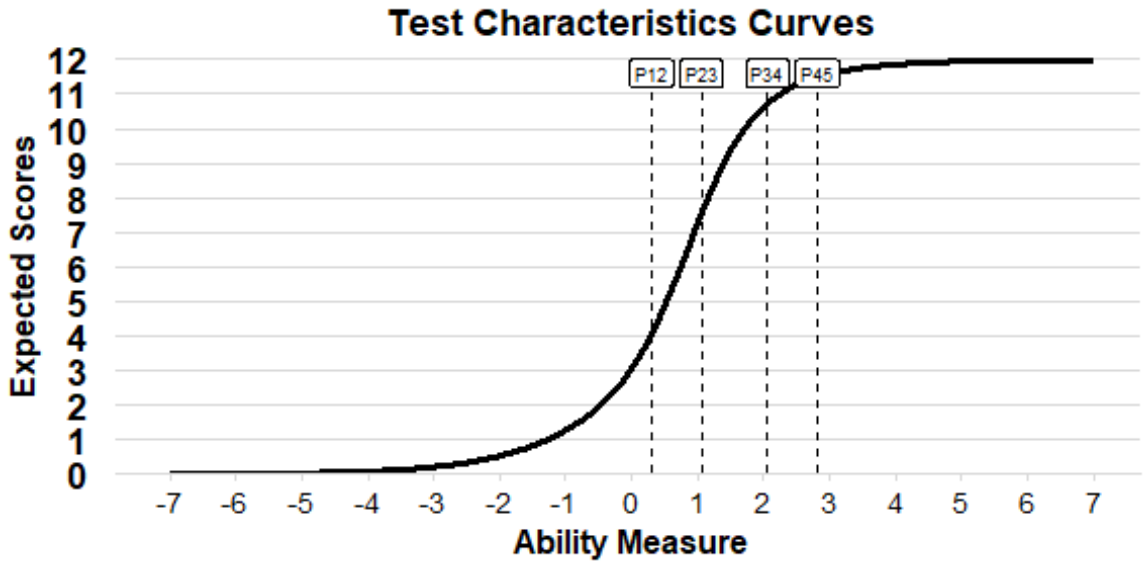


Figure 7.8.1.4

Test Characteristic Curve: List 9–12



7.8.2 Reading

Figure 7.8.2.1

Test Characteristic Curve: Read K–2

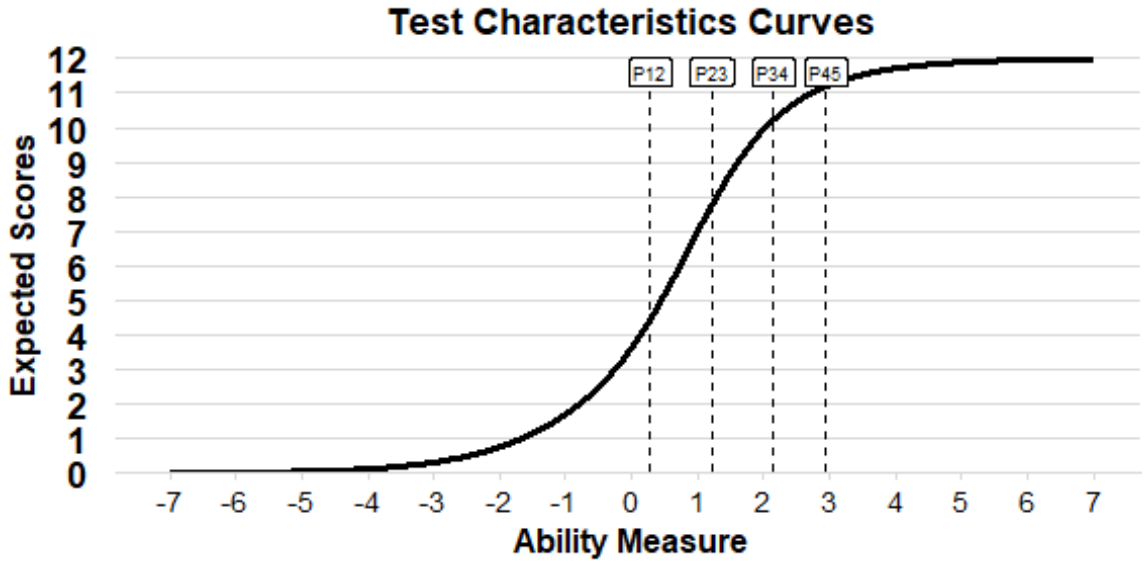


Figure 7.8.2.2

Test Characteristic Curve: Read 3–5

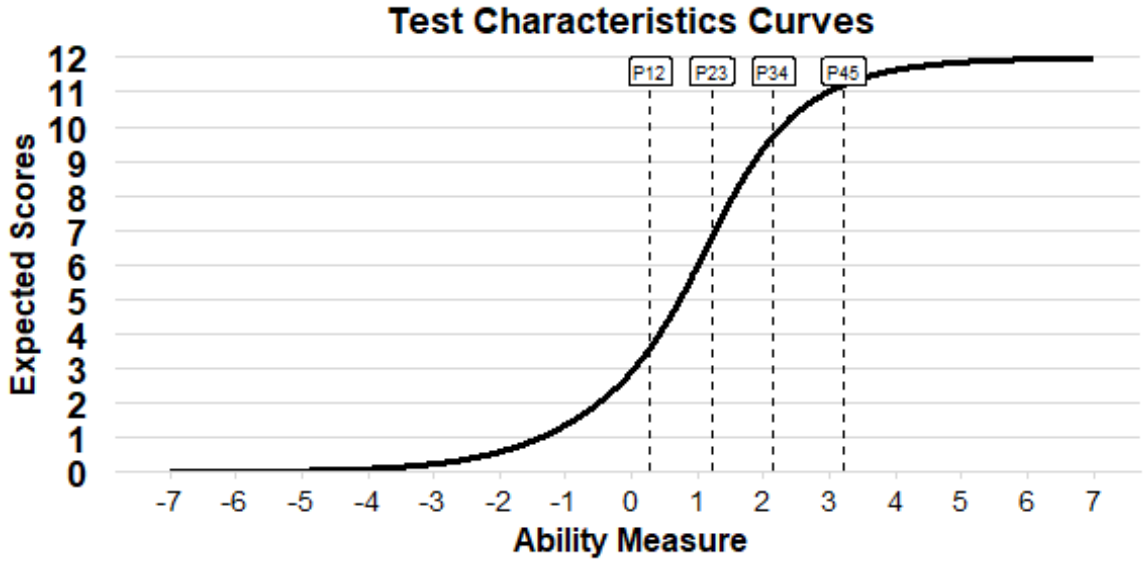


Figure 7.8.2.3

Test Characteristic Curve: Read 6–8

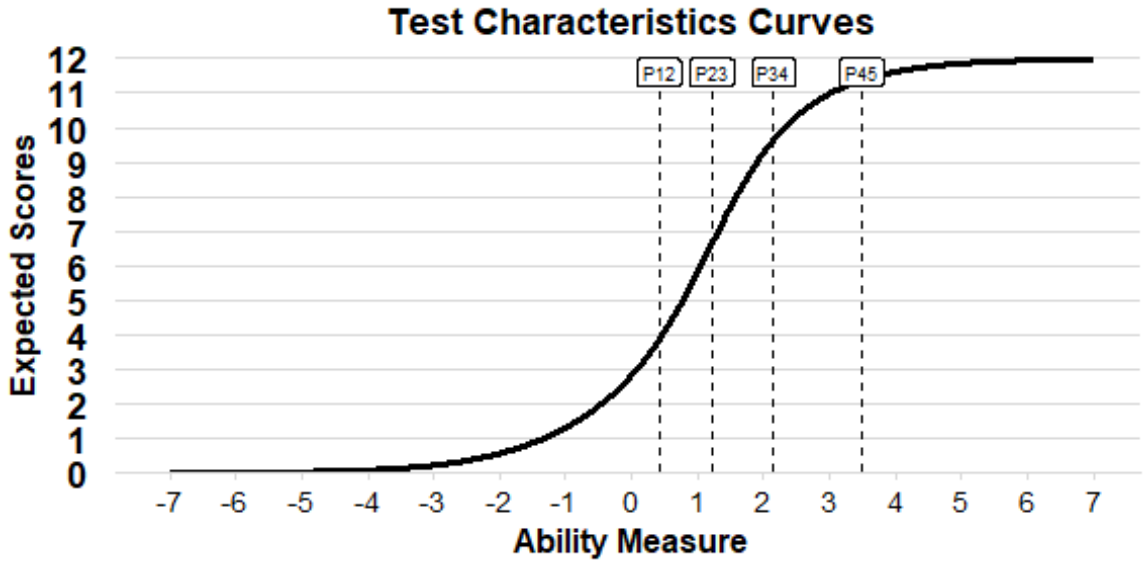
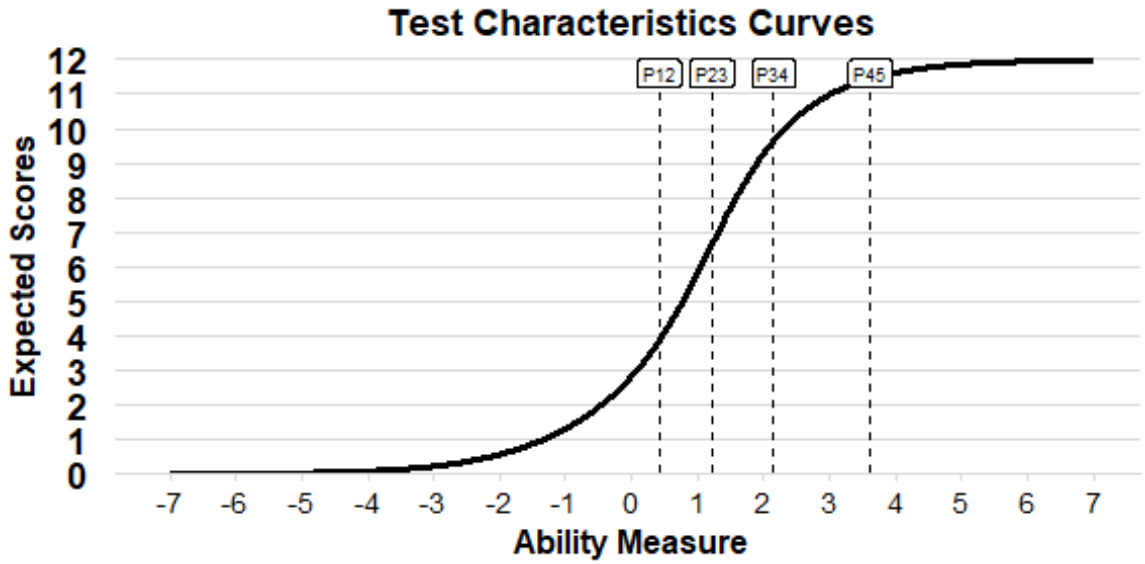


Figure 7.8.2.4

Test Characteristic Curve: Read 9–12



7.8.3 Speaking

Figure 7.8.3.1

Test Characteristic Curve: Spek K–2

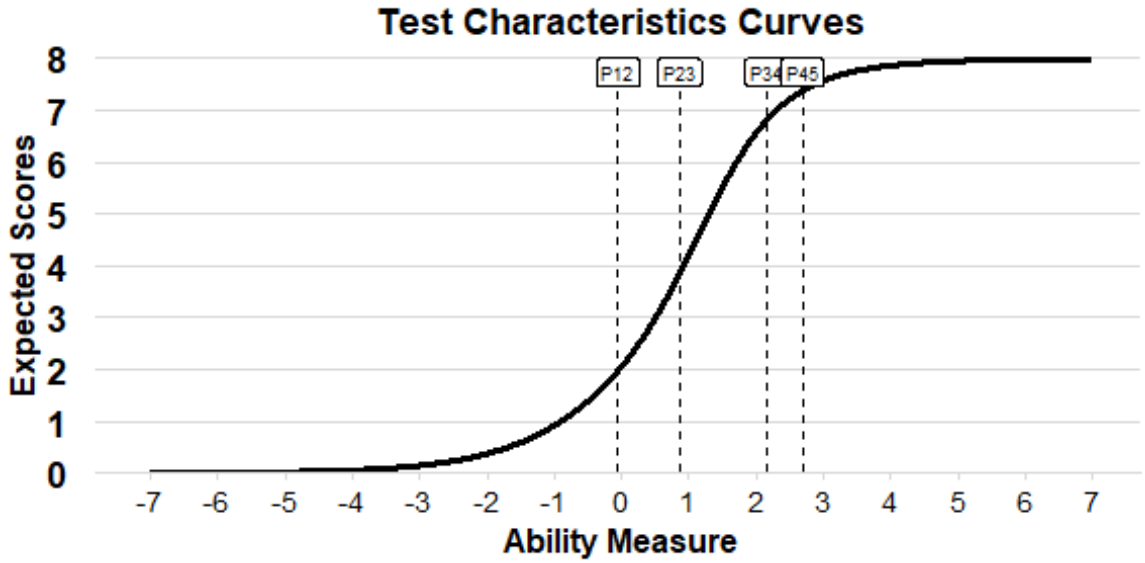


Figure 7.8.3.2

Test Characteristic Curve: Spek 3–5

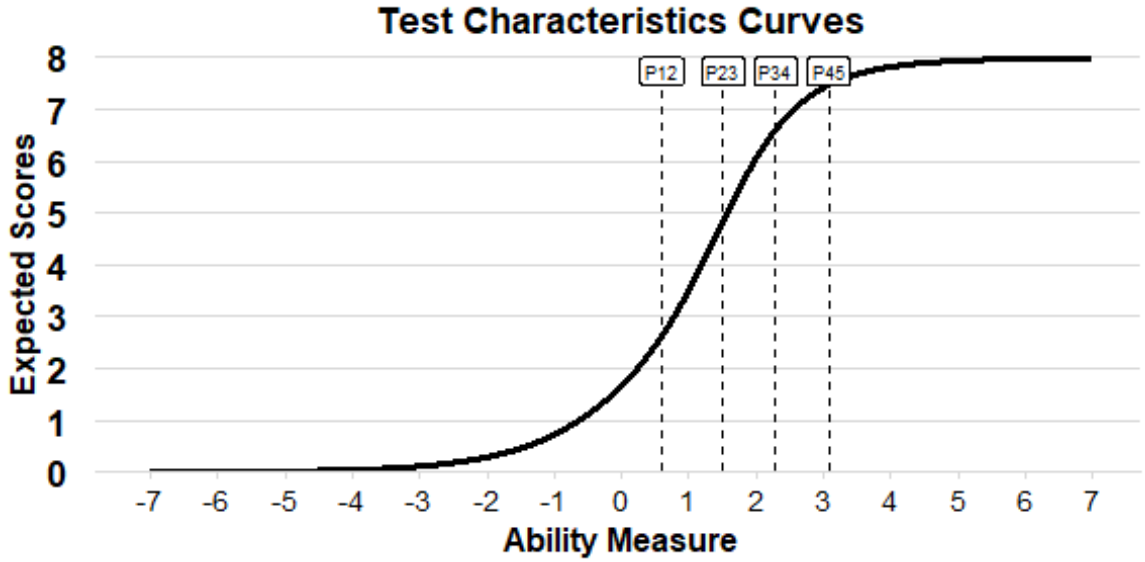


Figure 7.8.3.3

Test Characteristic Curve: Spek 6–8

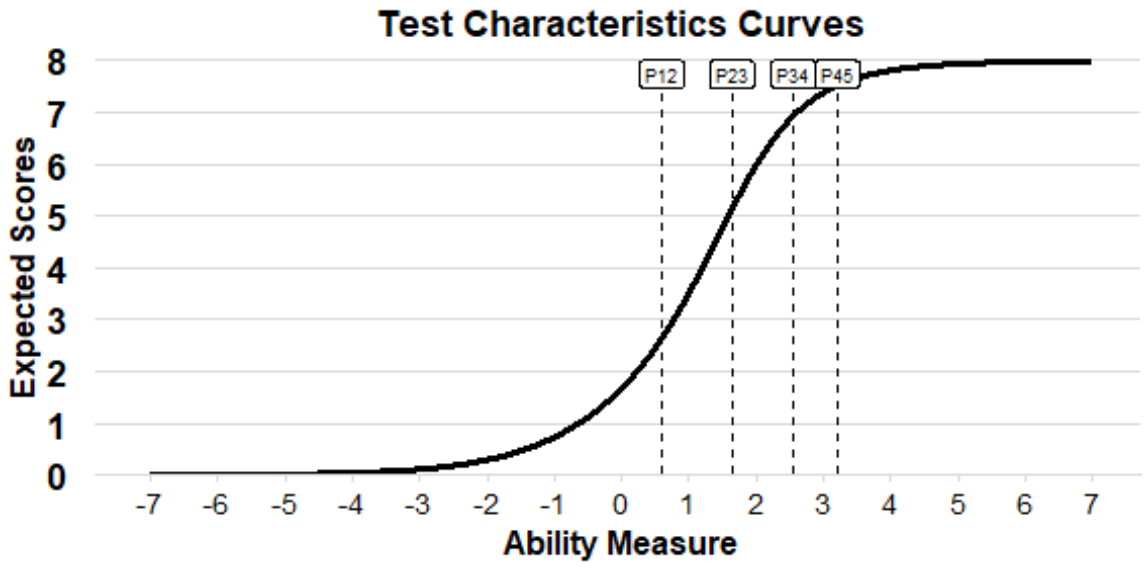
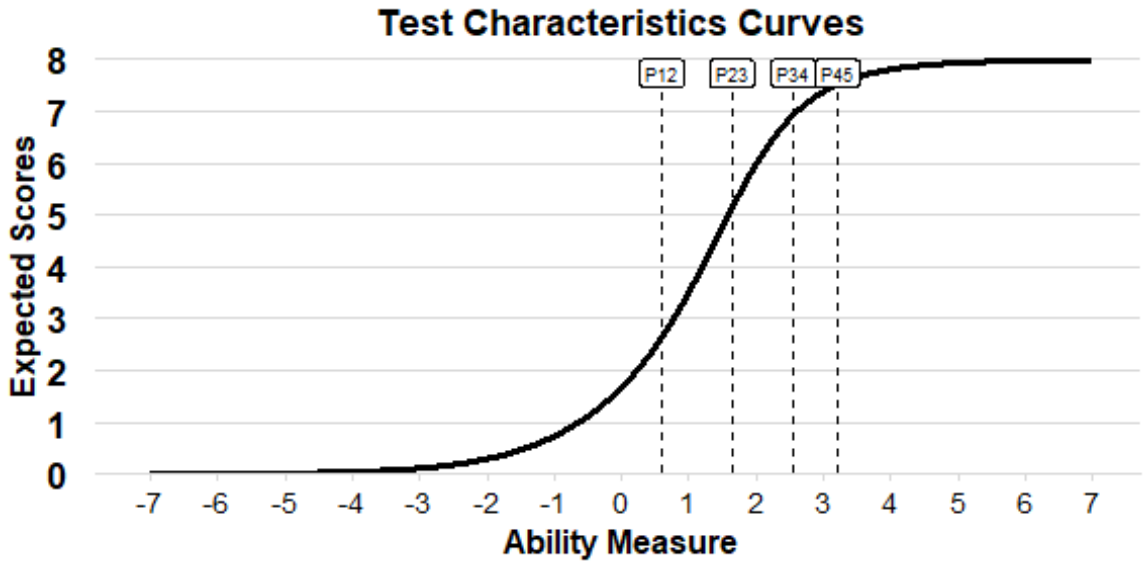


Figure 7.8.3.4

Test Characteristic Curve: Spek 9–12



7.8.4 Writing

Figure 7.8.4.1

Test Characteristic Curve: Writ K–2

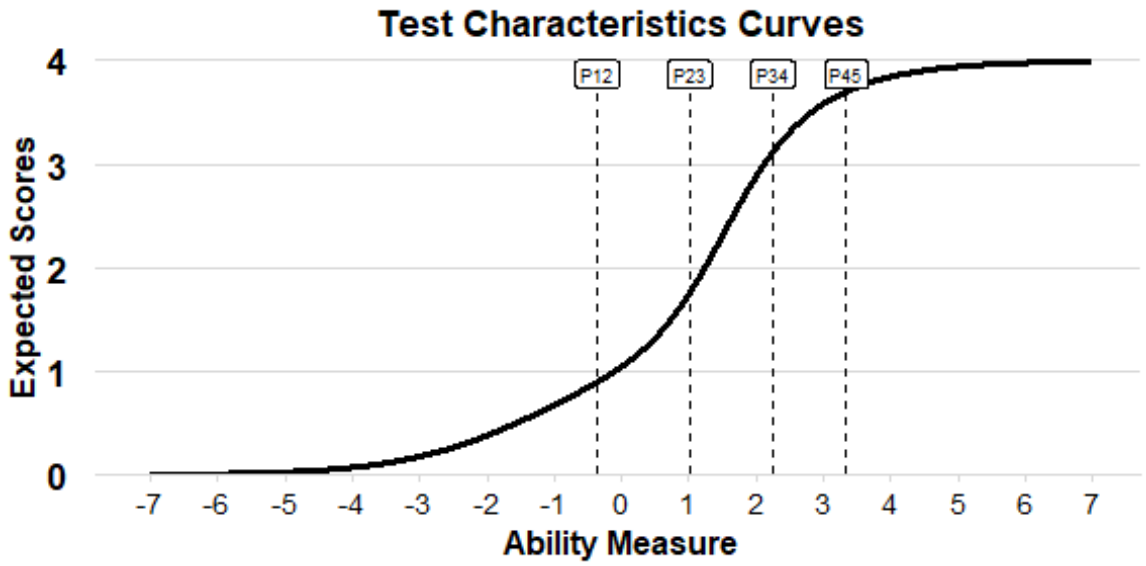


Figure 7.8.4.2

Test Characteristic Curve: Writ 3–5

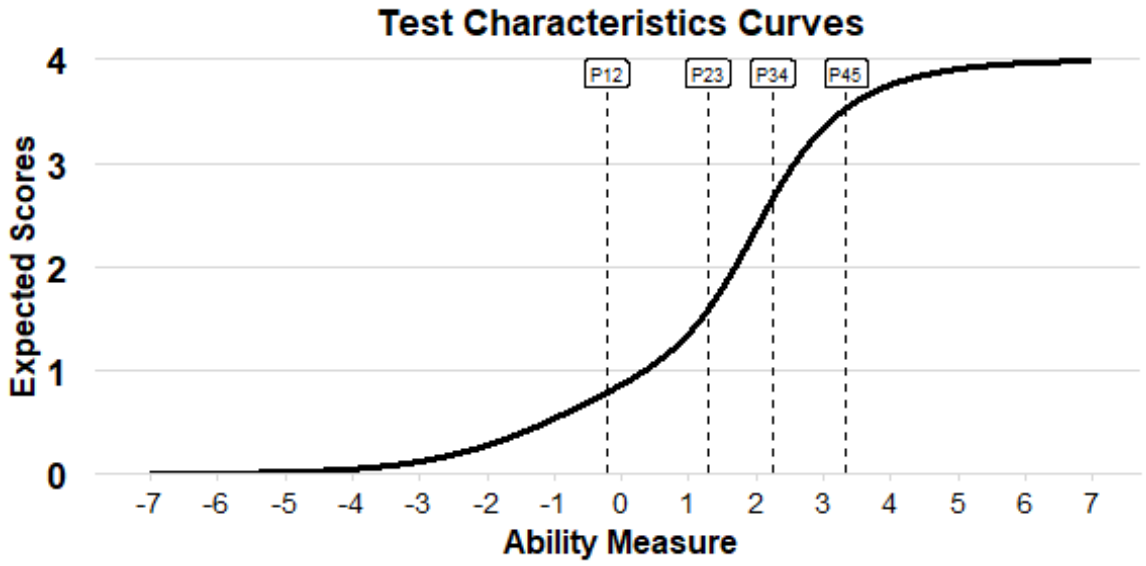


Figure 7.8.4.3

Test Characteristic Curve: Writ 6–8

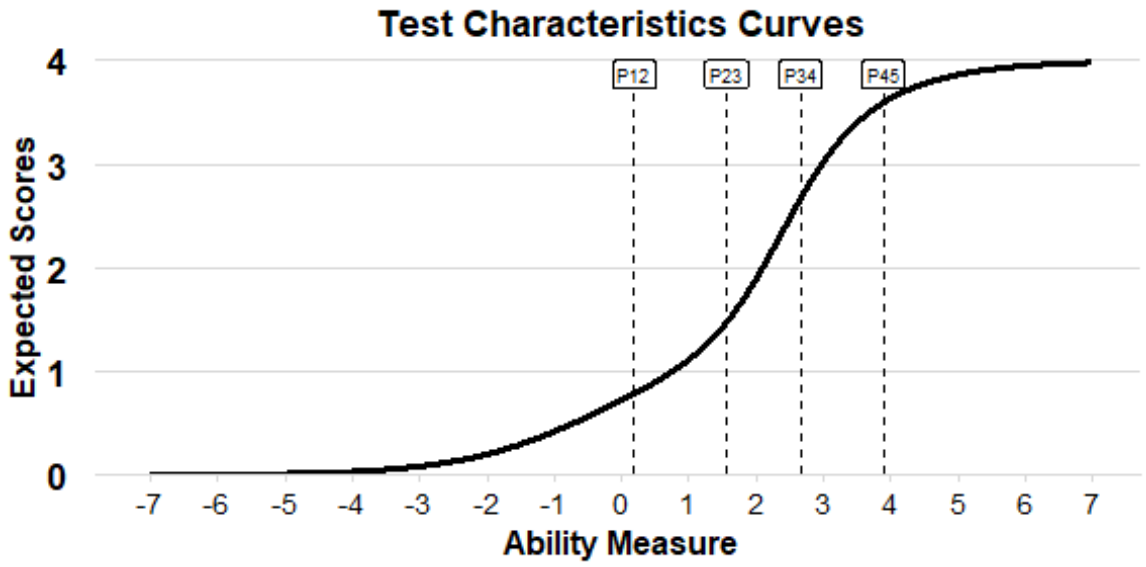
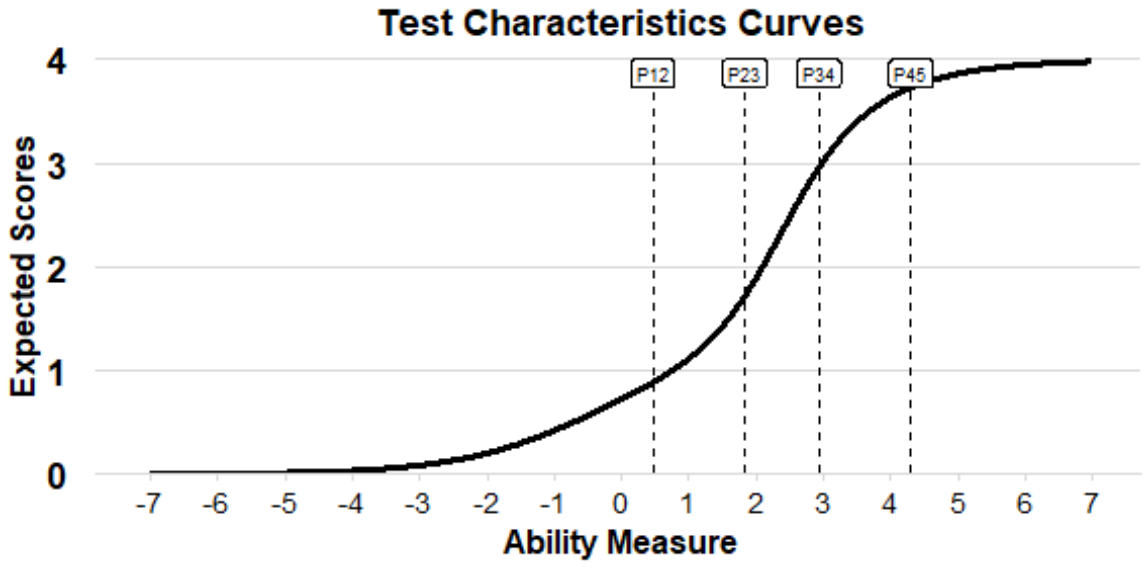


Figure 7.8.4.4

Test Characteristic Curve: Writ 9–12



7.9 TIF and TCC across Clusters

Figure 7.9.1.1.A through Figure 7.9.4.2.B present the Test Information Functions (TIFs) and Test Characteristic Curves (TCCs) for each domain, with curves for all clusters plotted together to allow for easier comparison. The suffix "A" denotes the Test Information Function, while "B" indicates the Test Characteristic Curve.

7.9.1 Listening

Figure 7.9.1.1.A

Test Information Functions: List K2–612

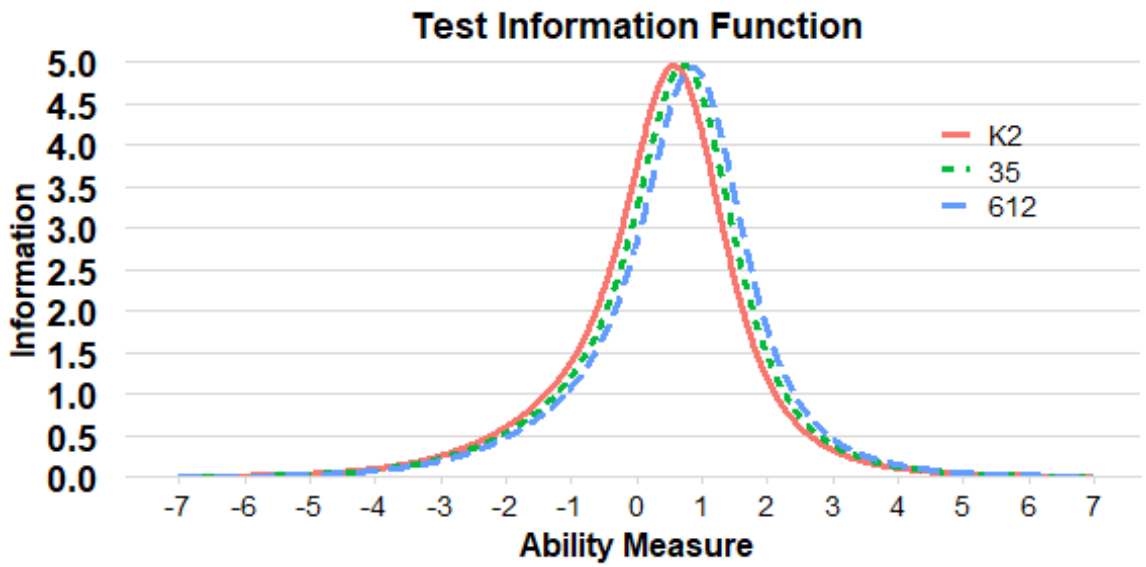
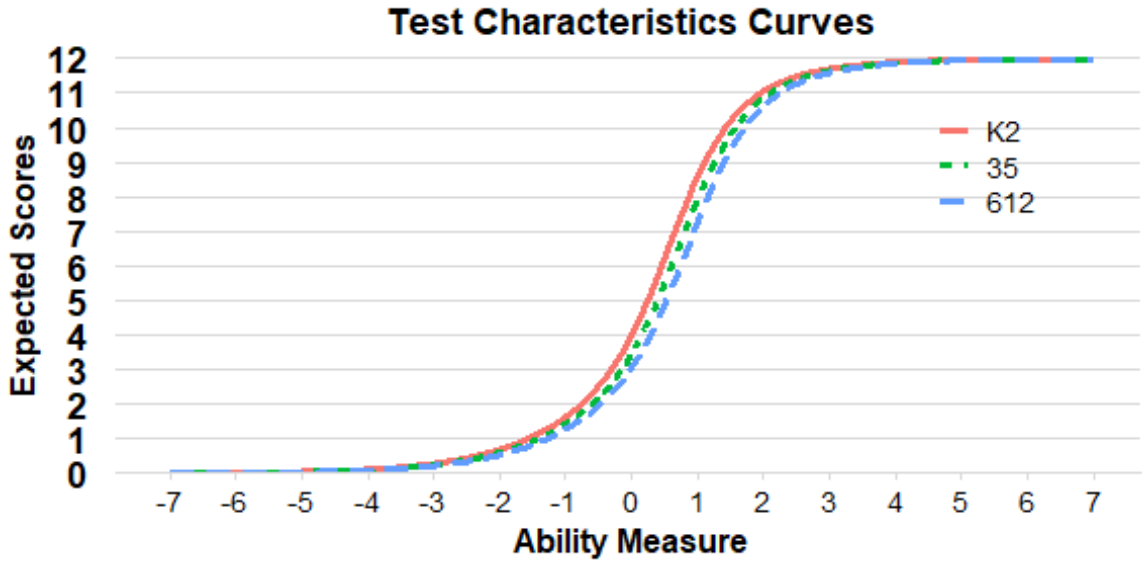


Figure 7.9.1.2.B

Test Characteristic Curves: List K2–612



7.9.2 Reading

Figure 7.9.2.1.A

Test Information Functions: Read K2–612

Figure 7.9.2.2.B

Test Characteristic Curves: Read K2–612

7.9.3 Speaking

Figure 7.9.3.1.A

Test Information Functions: Spek K2–612

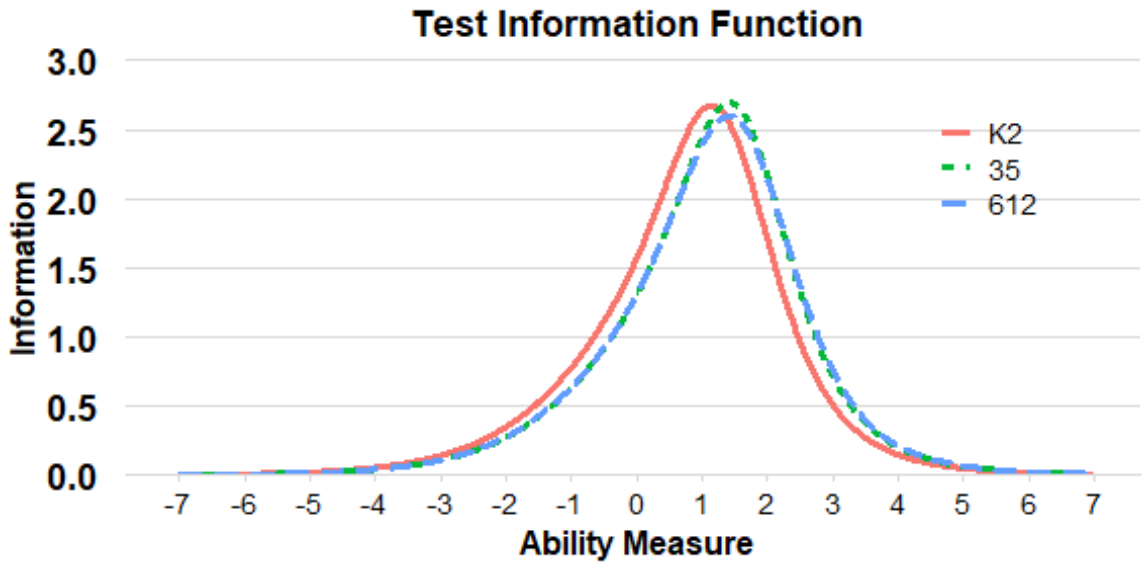
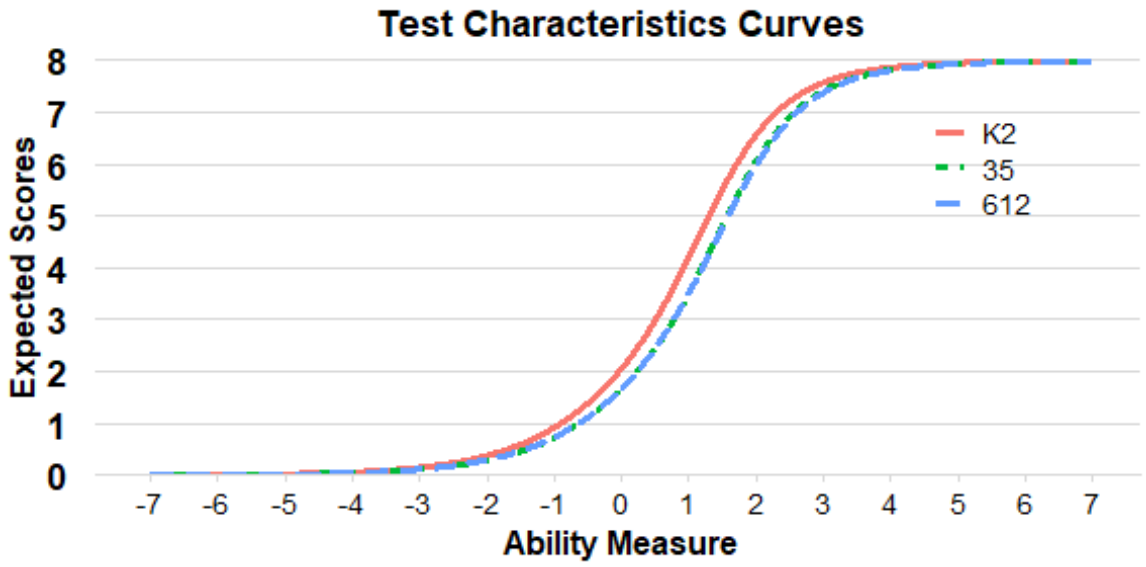


Figure 7.9.3.2.B

Test Characteristic Curves: Spek K2–612



7.9.4 Writing

Figure 7.9.4.1.A

Test Information Functions: Writ K2–612

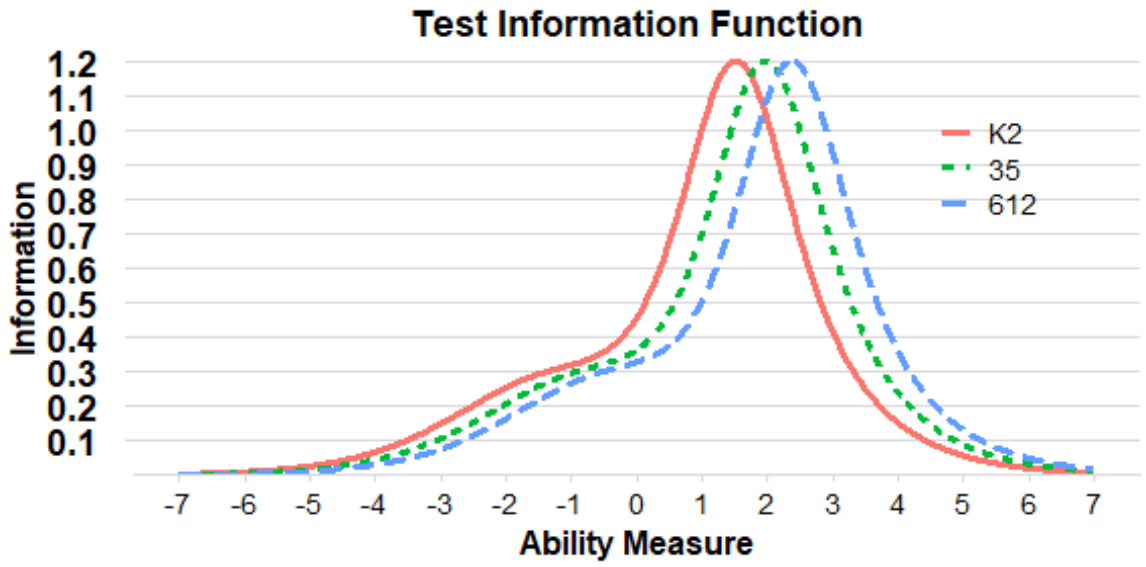


Figure 7.9.4.2.B

Test Characteristic Curves: Writ K2–612

8. Validation Study

The Alternate Screener Validation Study was conducted to evaluate the effectiveness of WIDA Alternate Screener, a new assessment tool designed to identify English learners (ELs) with the most significant cognitive disabilities. The study aimed to determine whether Alternate Screener scores correlate with those from the established Alternate ACCESS test, and whether test administrators (TAs) believe the scores reflect students' English language proficiency. Alternate Screener, a shortened version of Alternate ACCESS, includes nine items across four language domains and is intended for students who meet specific eligibility criteria. In the study, 18 TAs from 11 states administered the screener to 141 students, with 103 students' results matching Alternate ACCESS scores. The findings showed strong positive correlations between Alternate Screener and Alternate ACCESS scores, particularly in the Overall Composite score ($r = 0.833$), and moderate to high correlations across domains and grade clusters. Classification accuracy was acceptable, with the highest exact matches being in Speaking and Writing. While most TAs agreed that Alternate Screener scores aligned with their perceptions of students' proficiency, about one-third expressed concerns, often citing student disengagement. Additionally, a large majority of TAs supported the use of the Alternate Screener for EL identification. Despite these promising results, the study acknowledged limitations, such as a smaller-than-expected sample size, limited representation of higher-proficiency students, and the use of a pre-publication version of Alternate Screener. The report concludes that Alternate Screener is a valid tool for its intended purpose but recommends further validation following its official release.

The report outlines several future directions to improve the WIDA Alternate Screener and its validation process. While the current study provides encouraging evidence supporting Alternate Screener use, it also acknowledges the need for further refinement and validation. One of the key recommendations is to conduct a follow-up validation study after Alternate Screener is officially published. This future study would aim to address the limitations identified in the current research, such as the small sample size and the underrepresentation of students scoring at or above proficiency level 3 on Alternate ACCESS. Additionally, the follow-up would seek to gather more robust data on classification accuracy and test administrator perceptions, particularly to understand and mitigate concerns about student engagement and score validity. WIDA is also expected to use feedback from test administrators, SEAs, TAC, and subcommittees regarding training, ease of administration, and item engagement to refine the screener's design and implementation. Based on the feedback, WIDA plans to update Alternate Screener items to better align with the targeted difficulties in the near future.

These improvements and the subsequent validation efforts will help ensure c the reliability and effectiveness of Alternate Screener as a tool for identifying English learners with the most significant cognitive disabilities.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The 1996 NAEP Technical Report*. National Center for Education Statistics. Washington, DC.

American Educational Research Association, the American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561- 573.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2006). *A User's Guide to WINSTEPS: Rasch Model Computer Programs.* https://www.researchgate.net/publication/238169941_A_User''s_Guide_to_Winsteps_Rasch-Model_Computer_Program.

Lord, F. M. (2012). *Applications of Item Response Theory to Practical Testing Problems.* Routledge.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, *23*(2), 351.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*(3), 169-180.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago, IL: Mesa Press.

1025 W. Johnson St. | Madison, WI 53706-1706

(866) 276-7735 | help@wida.us | wida.wisc.edu