Washington Language Proficiency Test – II (WLPT-II)

Form A
Technical Report
2008 – 2009 School Year

Randy Dorn
State Superintendent of
Public Instruction

Prepared by
Pearson

for
Office of the Superintendent of Public Instruction
P.O. Box 47220
Olympia, Washington 98504-7220

Draft Submitted: August 31, 2009
Final Submitted: January 29, 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

OVERVIEW OF THE REPORT

The Washington Language Proficiency Test - II (WLPT-II) Technical Report for the 2008 – 2009 school year is divided into nine major sections, which are as follows:

The **Introduction** section presents the background, rationale, purpose, recommended test use, and test accommodations.

The **Test Design and Development** section describes the test development process of WLPT-II. It includes the test specifications, item development, review processes, and test construction.

The **Scoring** section provides a description of the scoring process for open-ended items. It provides information about rater training, intra-rater agreement, inter-rater agreement, and observed rater agreement statistics.

The **Reliability** section explains internal consistency reliability, classical standard error of measurement, and conditional SEM. It also provides the reliability statistics for each of the four modalities: Listening, Reading, Writing, and Speaking.

The **Validity** section describes the validity studies, including evidence of validity based on test content, internal structure, and test unidimensionality.

The **Classical Item-Level and Modality Statistics** section begins with a brief description of Classical Test Theory, followed by item-level summary descriptive statistics. Summary statistics by ethnicity and language groups are also provided.

The **Calibration, Equating, and Scaling** section explains the Rasch and Partial Credit Models, and provides sample item characteristic curves for a one-point item and a two-point item. Then, it summarizes the processes of calibration, equating, and scaling for the 2006 administration of the WLPT-II (Form A) assessment. Results of the 2009 administration of WLPT-II (Form A) calibration, equating, and scaling analyses are also presented. More detailed and comprehensive descriptions of the 2009 WLPT-II equating are available in the separate technical document, *Washington Language Proficiency Test – II Equating Study Report (2008 – 2009 School Year)*.

The **Summary of Operational Test Results** section presents scale score and proficiency level summaries for 2006-2009 for the regular administration. This section also provides summary information for the May 2009 assessment (Wave 2) that used Form C of the WLPT-II.

The **Accuracy and Consistency of Classifications** section presents results on the performance of performance levels, based on methodology from Livingston and Lewis (1995).

1. INTRODUCTION

1.1. Background

Title III of the federal *No Child Left Behind* (NCLB) Act of 2001 requires annual assessment of the English proficiency of Limited English Proficient (LEP) students, or English Language Learners (ELLs). Under the Title III requirements, the English language proficiency standards must be based upon the four modalities of Speaking, Reading, Writing and Listening. Additionally, the assessment must measure English language proficiency in the five domains of Speaking, Reading, Writing, Listening, and Comprehension (*Non-Regulatory Guidance on the Title III State Formula Grant Program. Part II: Standards, Assessments, and Accountability. Elementary and Secondary Education Act, As Amended by the No Child Left Behind Act of 2001, U.S. Department of Education*).

To meet these requirements, the Washington Office of Superintendent of Public Instruction (OSPI) launched an assessment project involving the development, research, and scoring of the WLPT-II. The test was developed for four grade spans (K–2, 3–5, 6–8, 9–12) in four modalities (Listening, Reading, Writing, and Speaking), to assess the English language proficiency of students whose first language is not English. Comprehension was operationally defined as the student's skill to understand spoken and written English language. Thus, Comprehension was measured by assessing the student's overall performance in both Listening and Reading. The test was developed in accordance with the Standards for Educational and Psychological Testing (American Educational Research Association, 1999) and the Washington State English Language Development (ELD) standards (http://www.k12.wa.us/MigrantBilingual/ELD.aspx).

1.2. Rationale and Purpose

In compliance with NCLB, OSPI developed the Washington Language Proficiency Test - II (WLPT-II), which measures student progress toward meeting these standards. In addition to using the Pearson's Stanford English Language Proficiency Test (SELP) items, augmented items were developed to produce custom test forms. Approximately 20% of each test form consisted of augmented items.

In line with the requirements of Title III, WLPT-II measures English language proficiency and determines when a student reaches the transitional level, which results in the student's exiting from English as a Second Language (ESL) or bilingual education programs. After exiting from the program(s), it is expected that ELLs will move into regular academic classes and receive instruction in English.

WLPT-II assesses students at all proficiency levels in Primary (K – 2), Elementary (3 – 5), Middle Grades (6 – 8), and High School (9 – 12). Year-to-year progress in language proficiency is measured longitudinally on the WLPT-II vertical scale. Test results may help schools focus on ways to make instruction more effective so that ELLs become proficient in English. Additionally, the vertical scale, from Pearson's Stanford English Language Proficiency (SELP) test, helps determine whether these students are making adequate progress toward English language proficiency.

1.3. Test Accommodations

The goal of the Washington State Assessment System is to assure every student has the opportunity to participate in the assessment, without providing a special advantage to any one of

them or to any group within the student body. Some assessment procedures, however, may be altered for a student, based on a review of the individual needs. These are available to any student who would benefit by the use of the altered procedures and use them during regular instruction. The decision is made on an individual basis and written in the student's IEP. These alterations in procedures are not used for the first time on state assessments. (Refer to *Washington State's Accommodations Guidelines for Students with Disabilities* for specific accommodations available to students.)

Although accommodations are intended to reduce or even eliminate the effects of a student's disability, they do not reduce learning expectations and should not give a false picture of what the student knows and is able to do. The accommodations provided to a student are the same for classroom instruction as well as district and state assessments, though not all classroom accommodations are appropriate on a standardized assessment. District Assessment Coordinators work with special education providers to ensure that accommodations written into IEPs are available to students at the time of testing. All building testing plans include an assessment accommodations plan that lists accommodations for each student.

Accommodations are practices and procedures in the areas of response, presentation, setting, and timing/scheduling that provide equitable access during assessments for students with disabilities.

- **Response Accommodations** allow students to complete activities, assignments, and assessments in different ways, or to solve or organize problems using some type of assistive technology, device, or organizer.
- **Presentation Accommodations** allow students to access information in ways that do not require them to visually read standard print. These modes of access are auditory, multisensory, tactile, and visual.
- **Scheduling/Setting Accommodations** increase the allowable length of time to complete an assessment or assignment, change the way the time is organized, or change the location in which a test or assignment is given or the conditions of the assessment setting.

For further information, refer to the following link to review the *Accommodations Guidelines*: http://www.k12.wa.us/assessment/WLPTII/pubdocs/2007-2009SPEDAccommodationsManualWLPT-II.pdf**}**

Scheduling/Setting:
All directions are reread verbatim.

- Provides an environment in which the student can read the directions aloud without disrupting other students.
- Directs students to underline or mark assessment directions with a No. 2 pencil.
- Audio records assessment directions for a student.
- Some students may require audio amplification devices to increase clarity. (This is provided in an environment that reduces distraction to others.)
- Provides a student additional breaks during a testing session.
- Allows student to use preferential seating, study carrel, or other school environment.

- Assesses the student individually or in a small group.
- Provides special lighting, auditory, or furniture supports.
- Offers noise buffers, such as ear phones, ear plugs, or headphones that are **not** connected to any audio device.

Presentation:

- Provides assistance in turning pages, handling booklets, etc.
- Provides the student with a No. 2 pencil adapted in size or grip.
- Provides student with a strip of heavy paper to assist in tracking.
- Provides tools to adjust color backgrounds such as overlays. In addition to these procedures, several individualized accommodations may be used for students with disabilities for wider access to assessments that are available to all students.

1.4. Large Type

Pearson has standardized large-type product specifications that serve to ease the test-taking experience for visually impaired students. A large-print version of each form was produced in large type for each of Primary through High School grade spans, with a minimum 18-point font for text and a maximum 24-point font for titles and headers. Pages were printed in black ink on a cream colored, non-glare vellum stock to ease readability of pages. Plastic spiral binding was used to make turning pages easy.

All student responses are written or transcribed verbatim, using a No. 2 pencil into the WLPT-II regular-print response booklets or Primary test booklets that accompany the large-print test materials. The transcribed booklets are processed in the same manner as all other scorable booklets.

2. TEST DESIGN AND DEVELOPMENT

2.1. Overview

The WLPT-II operational test was developed for four grade spans (K–2, 3–5, 6–8, and 9–12) in four modalities (Listening, Reading, Writing, and Speaking) to assess the English language proficiency of ELLs. The test was developed in accordance with the Standards for Educational and Psychological Testing (American Educational Research Association, 1999) and Washington State ELD standards.

WLPT consists of three forms (A, B, and C). Each of the three WLPT-II forms has been previously administered, with Form A administered in 2006 and now 2009, Form B administered in 2007, and Form C administered in 2008. Items needed to augment the SELP were field tested in quasi-operational[1] status during each of these administrations. In 2006, all new items in the Elementary and High School levels were accepted, thus completing Form A for those levels for future administrations. Also in 2006, the Primary level had 3 items rejected by the data review committee, while the Middle level had 1 item rejected. This meant that Form A for two of the four levels would require additional items to be field tested for the 2009 administration.

2.2. Test Specifications by Modality and Grade Span for WLPT-II (Form A)

Listening, Reading, Writing, and Speaking are assessed through several different item types: multiple-choice (MC), constructed-response (CR), short-response (SR), and extended-response (ER) items. The total number of items per grade span varies. Form A was originally administered in 2006. Before the 2006 data review, there were a total of 84 items for Primary, 83 items for Elementary, 92 items for Middle Grades, and 94 items for High School. The data review committee met shortly after the 2006 administration and decided to drop three Reading items from the Primary level and one Reading item from the Middle School level. These items were replaced with new items for the 2009 administration of the WLPT-II for form A.

The test design for the 2009 WLPT-II (FORM A) is shown in Table 1. Speaking has 17 CR items in each grade span. There are 20 MC Listening items for each grade span, while Reading has 24 to 31 MC items across grade spans. Note that Speaking consists of only CR items, while Listening and Reading consist of only MC items.

The Writing modality for each grade span is comprised of the following parts:

- MC section (Writing Conventions) that assesses ELLs' understanding of the conventions of written English at the word and sentence level.
- Pre-writing activity (excluding Primary). Pre-writing items are not scored, and are only intended to help students develop essays.
- Six SR items (for Primary only) in which students must copy printed text – a letter, a word, and a sentence, plus three dictation SR items (Primary only).
- Two ERs, responding to graphics-based prompts.

[1] Quasi-operational status refers to fact that new items are "pilot" tested as part of the operational administration. Once item statistics are available for review, OSPI makes a final determination of whether or not to keep the item on the test form.

For Elementary through High School, the number of Writing Conventions MC items in Form A ranged from 20 to 24, and each of these three grade spans has 2 ER prompts. For Primary, there are 15 Writing Conventions MC items, 6 SR items, and 2 ER prompts.

Comprehension consists of Listening and Reading subtests. Thus, the percentage of total items comprised from Comprehension ranged from 52 percent to 54 percent across grade spans.

Table 1: Test Specifications – Number of Items by Modality and Grade Span

Grade Span	Speaking	Listening	Reading		Writing			Total Number of Items
					Writing Conventions	Short Writing	Writing Prompt	
	CR	MC	MC	Passages	MC	SR	ER	
Primary: K-2	17	20	24	5	15	6	2	84
Elementary: 3-5	17	20	24	5	20	0	2	83
Middle Grades: 6-8	17	20	29	5	24	0	2	92
High School: 9-12	17	20	31	5	24	0	2	94

Table 2 provides the maximum number of points by modality and grade span. The percentage of total points for Comprehension ranged from 38 percent to 42 percent.

Table 2: Maximum Number of Points by Modality and Grade Span

Grade Span	Speaking	Listening	Reading		Writing			Total Number of Points
					Writing Conventions	Short Writing	Writing Prompt	
	CR	MC	MC	Passages	MC	SR	ER	
Primary: K-2	38	20	24	5	15	10	8	115
Elementary: 3-5	38	20	24	5	20	0	8	110
Middle Grades: 6-8	38	20	29	5	24	0	8	119
High School: 9-12	38	20	31	5	24	0	8	121

2.3. Item Mapping to Washington ELD Standards by Grade Span

Harcourt (now Pearson) conducted an alignment study comparing SELP Form A to the Washington State ELD standards as part of the company's proposal for the Washington project. Additionally, a committee of Washington state educators performed a second alignment study using the state's English Language Proficiency Descriptors, which are broader than the state's ELD standards, to confirm the general gaps in the SELP forms. This committee recommended that SELP forms be augmented in the Reading, Writing, and Speaking subtests, aimed at advanced proficiency learners at each grade span, i.e., advanced proficiency second graders for the K-2 (Primary) test, advanced proficiency fifth graders for the 3-5 (Elementary) test, and so on for the 6-8 (Middle Grades) and 9-12 (High School) tests. Because the item types are parallel across all three SELP forms, alignment of an item type from Form A implies a match for the same item type on Form B and/or Form C. The full results of the two alignment studies can be found in the *Washington Language Proficiency Test – II Technical Report (2005 – 2006 School Year)*.

2.4. Item Development

To create a new and fully aligned assessment for ELLs, and also to meet the reporting requirements for NCLB, Pearson made use of a bank of field-tested English language proficiency (ELP) items, in addition to developing new items. The Pearson ELP item bank includes items developed for the Stanford English Language Proficiency Test (SELP) Forms A, B, and C. The WLPT-II (Form A) was developed from SELP Form A. The 2009 WLPT-II (Form A) was identical to the 2006 WLPT-II (Form A) for the Elementary and High School forms. The 2009 WLPT-II (Form A) Primary form had three new Reading items compared to the 2006 WLPT-II (Form A) and the 2009 WLPT-II (Form A) Middle School form had one new Reading item compared to the 2006 WLPT-II (Form A).

Items in the bank (for all three SELP forms) were originally submitted by educators of English language learners. Assessment specialists reviewed the items to ensure the following:

- Item soundness
- Freedom of item language, cultural, or gender bias
- Appropriateness of topic, vocabulary, and language structure for each grade span
- Match to the Teachers of English to Speakers of Other Languages (TESOL) standards and individual state ESL standards

Only test items judged to be of acceptable quality and fairness to students were approved to be included on the WLPT-II. Questions were also sampled in ELL classrooms to ensure that the directions are clear and easy-to-follow, and that they are reliable indicators of student achievement.

To develop augmented items for WLPT-II, OSPI convened committees of Washington state educators for an item writing meeting in October 2005. At the meeting, facilitators first provided intensive item-writing training. Next, facilitators worked closely with the writers during the development of augmented Reading items for passages. Lastly, writers were asked to work in small groups, led by the facilitators, to develop the augmented Writing and Speaking items. After the item-writing conference, the newly-developed, augmented items were reviewed by Harcourt (now Pearson) content and editorial staff and were then compiled into review booklets.

2.5. Content and Item Bias & Sensitivity Reviews

In August 2005, a committee composed of twelve Washington State ESL professionals, including classroom teachers, school administrators, and university faculty, reviewed SELP Forms A, B, and C for bias and sensitivity. The committee recommended various revisions to items in the three forms.

In the week following the October 2005 item writing meeting, additional Washington State educators reviewed the newly created augmented items for alignment of content to ELD standards and for bias and sensitivity.

2.6. Test Construction

SELP and augmented items represent a broad range of difficulty at all grade levels. Items range from very easy for students with little or no ability in English to very difficult for students with advanced ability in English. The original proposed final version of Form A that was administered in 2006 was submitted to OSPI for bias and sensitivity review, as well as alignment

to the Washington ELD Standards. OSPI provided final approval on the form to be printed. For the 2009 WLPT-II (Form A), the Primary form was updated with three Reading items and the Middle School form was updated with one Reading item. These items were approved by OSPI.

2.7. Data Review

In April 2006, a data review committee consisting of Washington ESL professionals reviewed each augmented item on Form A and the associated item statistics. The committee decided not to use three Reading items on the Primary form and one item on the High School form. These items were excluded from the 2006 equating study, reported results, and all subsequent statistical analyses.

The item statistics used at the 2006 Data Review were based on 50% of the total testing population. The statistics provided included response-option distributions, item means, item-total correlations, differential item function (DIF) statistics, and response-total correlations for MC items.

For MC items, the item mean is the proportion of students that answer an item correctly (i.e., *p*-value). For the CR, SR, and ER items, the item mean is the average number of points earned.

The item-total correlation is an index of association between item score and the total test score. It shows the ability of the item to discriminate between low- and high-ability students. An item with a large item-total correlation discriminates more effectively between the low- and the high-ability students than an item with a small item-total correlation. In the case of a dichotomous item, the index is also referred to as a point-biserial correlation. In the case of a polytomous item, the index is also referred to as a point-polyserial correlation.

The response-total correlation is an index of association between a particular item response option and the total-test score. It shows the relationship between a response option and the total score. The response-total correlation for the correct response is equivalent to the item-total correlation. The response-total correlations for the incorrect response-options tend to be negative in value for well-written items.

The statistics for the three new Reading items on the Primary form and one new item on the Middle School form for the 2009 WLPT-II (Form A) were presented to OSPI for review in April 2009. OSPI decided to keep each of the four items, so no items were dropped from the 2009 WLPT-II assessment.

2.8. Differential Item Functioning

This section provides information about Differential Item Functioning (DIF) analyses for the WLPT-II assessment. For the WLPT-II DIF analyses, the reference group was male students, and the focal group was female students. Since WLPT-II was a mixed-format examination, composed of Multiple Choice (MC) and Open-Ended (OE) items, the DIF procedure used consisted of Mantel's (1963) extension of the Mantel-Haenszel procedure for the OE items and the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) for the MC items. For OE items, the DIF procedure used the Mantel statistic in conjunction with the Standardized Mean Difference (SMD) while for the MC items, the Mantel-Haenszel procedure was used in conjunction with the Delta Scale.

2.8.1. Mantel χ^2

The Mantel χ^2 is a conditional mean comparison of the ordered response categories for reference and focal groups combined over values of the matching variable score. By "ordered" we mean that a response of "1" on an item is better than "0," "2" is better than "1," and so on. "Conditional," on the other hand, refers to the comparison of members from the two groups who received the same score on the matching variable, i.e., the total test score in the analysis for the WLPT-II.

Table 3 shows a $2 \times T \times K$ contingency table, where T is the number of response categories and K is the number of levels of the matching variable. The values, y_1, y_2, …, y_T are the T scores that can be gained on the item. The values, n_{Ftk} and n_{Rtk}, represent the numbers of focal and reference groups who are at the k^{th} level of the matching variable and gain an item score of y_t. The "+" indicates total number over a particular index (Zwick, Donoghue, & Grima, 1993).

Table 3: $2 \times T$ Contingency Table at the k^{th} Level[1]

Group	**Item Score**				**Total**
	y_1	y_2		y_T	
Reference	n_{R1k}	n_{R2k}	…	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	…	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	…	n_{+Tk}	n_{++k}

[1] Zwick, et al. (1993)

The Mantel statistics is defined as the following formula:

$$Mantel\ \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k Var(F_k)}$$

where $F_k = \sum_t y_t \cdot n_{Ftk}$ is the sum of scores for the focal group at the k^{th} level of the matching variable,

$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t \cdot n_{+tk}$ is the expectation of F_k under the null hypothesis, and

$Var(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \left[\left(n_{++k} \sum_t y_t^2 n_{+tk} \right) - \left(\sum_t y_t n_{+tk} \right)^2 \right]$ is the variance of F_k under the null hypothesis.

Under H_0, the Mantel statistic has a chi-square distribution with one degree of freedom. In DIF applications, rejecting H_0 suggests that the students of the reference and focal groups who are similar in overall test performance tend to differ in their mean performance on an item. In the case of dichotomous items, on the other hand, the statistic is identical to the Mantel-Haenszel (1959) statistic without the continuity correction (Zwick, et al., 1993).

2.8.2. Standardized Mean Difference (SMD)

A summary statistic to accompany the Mantel approach is the Standardized Mean Difference (SMD) between the reference and focal groups proposed by Dorans and Schmitt (1991). This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the values of the matching variable. SMD has the following form (adapted from Dorans & Schmitt, 1991):

$$\text{SMD} = \sum_k p_{Fk} m_{Rk} - \sum_k p_{Fk} m_{Fk}$$

where $p_{Fk} = \frac{n_{F+k}}{n_{F++}}$ is the proportion of the focal group members who are at the k^{th} level of the matching variable,

$m_{Fk} = \frac{1}{n_{F+k}(\sum_t y_t \cdot n_{Ftk})}$ is the mean item score of the focal group members at the k^{th} level, and

m_{Rk} is the analogous value for the reference group.

As can be seen from the equation above, the SMD is the difference between the weighted-item mean of the reference group and the unweighted-item mean of the focal group. The weights for the reference group are applied to make the weighted number of the reference-group students the same as in the focal group within the same ability. A negative SMD value (or "<" in this report) implies that the focal group has a higher mean item score than the reference group, conditional on the matching variable.

2.8.3. DIF classification for OE items

The SMD is divided by the total group item standard deviation to obtain an effect-size value for the SMD. This effect-size SMD is then examined in conjunction with the Mantel χ^2 to obtain DIF classifications that are depicted in Table 4 below.

Table 4: DIF Classification for OE Items

Category	Description	Criterion[1]
AA	No DIF	Non-significant Mantel χ^2 or Significant Mantel χ^2 and $\lvert SMD/SD \rvert \leq .17$
BB	Weak DIF	Significant Mantel χ^2 and $.17 < \lvert SMD/SD \rvert \leq .25$
CC	Strong DIF	Significant Mantel χ^2 and $.25 < \lvert SMD/SD \rvert$

[1] SD is the total group standard deviation of the item score in its original metric

For the MC items, the Mantel-Haenszel Chi-square (*M-H* χ^2) is used in conjunction with the *M-H* odds ratio that is transferred to the delta scale (*D*). The odds of a correct response (proportion passing divided by proportion failing) are *P/Q* or *P/(1-P)*. The odds ratio, on the other hand, is simply the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. For a given item, the odds ratio is defined as follows:

$$\alpha_{M-H} = \frac{P_r / Q_r}{P_f / Q_f}$$

And the corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups. Thus, the odds ratio is equal to 1:

$$H_0 : \alpha_{M-H} = \frac{P_r / Q_r}{P_f / Q_f} = 1$$

2.8.4. The Delta Scale

In order to make the odds ratio symmetrical around zero with its range being in the interval $-\infty$ to $+\infty$, the odds ratio is transformed into a log odds ratio as per the following: $\beta_{M-H} = \ln(\alpha_{M\text{-}H})$. The simple natural logarithm transformation of this odds ratio is symmetrical around zero, in which zero has the interpretation of equal odds. This DIF measure is a signed index, where a positive value signifies DIF in favor of the reference group, while a negative value indicates DIF in favor of the focal group. β_{M-H} also has the advantage of being transformed linearly to other interval scale metrics (Camilli & Shepard, 1994). This fact is utilized in creating the delta scale (*D*), which is defined as $D = -2.35 \cdot \beta_{M-H}$.

2.8.5. DIF classification for MC items

The *M-H* χ^2 is examined in conjunction with the delta scale (D) to obtain DIF classifications depicted in Table 5 below. The four new items that were field tested in on WLPT-II (Form A) were MC items. None of these items in 2009 had a DIF flag.

Table 5: DIF Classification for MC Items

Category	Description	Criterion
A	No DIF	Non-significant *M-H* χ^2 or $\|D\| < 1.0$
B	Weak DIF	Significant *M-H* χ^2 and $\|D\| < 1.5$ or
		Non-significant *M-H* χ^2 and $\|D\| \geq 1.0$
C	Strong DIF	Significant *M-H* χ^2 and $\|D\| \geq 1.5$

3. SCORING

All multiple-choice items are scored as correct or incorrect and are machine scored. The Directions for Administering (DFA) contain administration and scoring instructions, along with scoring rubrics for the Speaking items. The Speaking subtest is an individually administered, free-response assessment, and each item was scored by the test proctor, who was provided additional scoring information in the DFA. The multiple choice items were scored by Scoring Operations (SCOPS) while the Writing short-answer (SA) and extended-responses (ER) items were scored by Performance Scoring Center (PSC). At least 10% of the Writing items received a second reading for reliability and accuracy purposes. Anchor papers, training sets, and rubrics were used as scoring guides. If questions arose during scoring, the problem was discussed by the group to maintain consistency in scoring.

3.1. Rater Training and Intra-Rater Agreement

All PSC scorers had a minimum of a Bachelor's degree and successfully completed generalized training in performance assessment scoring. In addition to the general scorer training, all scorers assigned to score the WLPT-II test were required to qualify on project-specific training with rubrics, anchor papers, and practice papers.

The accuracy of scoring was monitored by Scoring Directors and Scoring Supervisors who are seasoned PSC scorers and who had extensive experience in all facets of scoring.

The Scoring Directors and Scoring Supervisors monitored scoring through the PSC backreading system. In this case, unlike blind second scoring, the Scoring Supervisors review the scores entered by their scorers. This feature allowed the Scoring Supervisor to monitor the scores being assigned by a scorer and to intervene as needed to ensure the accuracy of scoring. (Scoring Directors also backread their Scoring Supervisors in the same manner to ensure their scoring accuracy.) The targeted agreement rate for scoring student responses was 70% perfect agreement, with no more than 5% non-adjacent agreement. Scorers failing to achieve this agreement rate were retrained. Scorers who failed to maintain the minimum agreement rate for scoring following retraining were removed from the project.

In addition to regular student responses, scorers scored validity responses each day to measure their intra-rater reliability. Validity responses are student papers that have been pre-scored by scoring experts. Each scorer completed a blind scoring of numerous validity papers throughout the day. A daily validity report was prepared indicating the number and percent in perfect agreement, *within* ± 1 score point agreement, and *beyond* ± 1 score point agreement. The targeted agreement for validity responses was 80 percent perfect agreement, plus 20 percent adjacent agreement. Scorers failing to achieve this validity agreement rate were given a "must pass" targeted calibration set. Scorers who failed the calibration set (or who passed the set but thereafter failed to maintain the minimum validity agreement rate) were removed from the project. The table below summarizes the overall results of the readers' daily intra-rater agreement for WLPT-II scoring. The summary in Table 6 indicates that the agreement rates met the target.

Table 6: Mean Intra-Rater Agreement Statistics Across Daily Validity Sets by Grade Span

	Intra-Rater Agreement	
Grade Span	Mean % Perfect	Mean ± 1 Adjacent
Primary: Grades K–2	98	2
Elementary: Grades 3–5	83	17
Middle Grades: Grades 6–8	83	17
High School: Grades 9–12	91	8

3.2. Inter-Rater Agreement

During the scoring process, a second score (also called a blind read) monitoring process was followed to measure the scorers' inter-rater reliability. Ten percent of the student papers were read by two scorers. Two definitions were followed to check the accuracy and reliability of the scores. The first definition, *% Perfect*, addressed the percent perfect agreement between the first and second ratings. Under this definition, agreement is present as long as the score arising from the second rating matches exactly the score from the first rating. The second definition, *± 1 Adjacent*, addresses the percent of agreement between adjacent score categories. For this definition, agreement is present when discrepancies between the first and second ratings are within ± 1 score point. There was no third reading for non-adjacent scores. The first reader's score was final unless overridden by a supervisor's backreading score.

Data from the second score procedure were analyzed under the two previously stated definitions of inter-rater agreement. The targeted agreement rate for responses was 70% perfect agreement with no more than 5% greater than ±1 score point discrepancy. Table 7 provides the rater agreement statistics for the Writing items on the 2009 WLPT-II. The statistics indicate that the degree of the inter-rater agreement was on target.

Table 7: Inter-Rater Agreement Statistics by Grade Span

	Inter-Rater Agreement		
Grade Span	% Perfect	± 1 Adjacent	Total (Perfect +Adjacent)
Primary: Grades K–2	93.1	6.6	99.7
Elementary: Grades 3–5	84.8	14.6	99.4
Middle Grades: Grades 6–8	82.8	16.7	99.5
High School: Grades 9–12	84.4	14.9	99.3

3.3. Research File

After 100% of PSC scoring was completed, the Operations department merged all scoring files to create a Scored File. This file was verified by Pearson's Assessment and Information Quality group (AIQ) based on the description values in the file layout. Once verified, a Research File for the 2009 WLPT-II test was created and verified by AIQ again. After the verification and approval by AIQ, the Research File was forwarded to Psychometric and Research Service (PRS). PRS used this file for item analysis and equating. Once all analyses were completed, PRS provided Measurement Services (MS) with raw score to scale score conversion tables and scaled cut score tables. These tables were then used to update the student data to create a Student Data File.

4. RELIABILITY

4.1. Classical Test Theory

There are useful indices available within the framework of Classical Test Theory (CTT), for estimating the precision of the raw test scores and the reliability of assessments. Within CTT, an observed test score is defined as the sum of a student's true score and error (X = T + E, where X = the observed score, T = the true score, and E = error). A true score is considered the student's true standing on the measure, while the error score reflects a random error component. Thus, error is the discrepancy between a student's observed and true score.

The reliability coefficient of a measure is the proportion of variance in observed scores accounted for by the variance in true scores. The coefficient can be interpreted as the degree to which scores remain consistent over parallel forms of an assessment (Ferguson & Takane, 1989; Crocker & Algina, 1986). There are several methods for estimating reliability:

- In the **Test-Retest Method**, the same test is administered on two occasions to determine whether examinees respond consistently over a brief period of time.
- In the **Parallel Forms Method**, equivalent forms of a test are administered to the same group of subjects to determine whether examinees respond consistently on two parallel test forms.
- In the **Internal Consistency Method**, a single form is administered to the same group of subjects to determine whether examinees respond consistently across the items within a test.

Because the WLPT-II is a secure test that should not be administered twice, internal consistency was utilized.

4.2. Internal Consistency Reliability

The Internal Consistency Method investigates the stability of scores from one sample of content to another by estimating how consistently individuals respond to items. A basic estimate of internal consistency reliability is the split-half method, in which the test is split into two parallel halves and scores on each half-test are correlated. Which items contribute to which half-test's score can have an impact on the resulting correlation.

To counter this concern, *Cronbach's Coefficient Alpha* statistic (Cronbach, 1951) was used. Coefficient alpha is equivalent to the average split-half correlation based on all possible divisions of a test into two halves. Coefficient alpha can be used on any combinations of both dichotomous (two score values) and polytomous (two or more score values) test items and is computed using the following formula:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} S_j^2}{S_X^2}\right),$$

where n is the number of items,

S_j^2 is the variance of students' scores on item j, and

S_X^2 is the variance of the total-test scores.

Cronbach's alpha ranges in value from 0.0 and 1.0, where higher values indicate greater proportion of observed score variance is true score variance. Two factors affect estimates of internal consistency: test length and homogeneity of items. The longer the test, the more observed score variance is likely to be true score variance. The more similar the items, the more likely examinees will respond consistently across items within the test.

4.3. Classical Standard Error of Measurement

The purpose of a reliability coefficient is to estimate the proportion of observed score variance that is true score variance. With this statistic, one can infer the proportion of observed score variance that is error variance. The Standard Error of Measurement (SEM) is another way of understanding reliability. The SEM is the square root of the error variance. This statistic indicates the amount of measurement error in a set of observed test scores. The SEM is inversely related to the reliability of a test; therefore, the greater the reliability, the lower the SEM. With a lower SEM, there is more confidence in the accuracy, or precision, of the observed test scores. The SEM is calculated using the following equation:

$$SEM = \sigma_X \sqrt{1 - \rho_{XX}} ,$$

where σ_X is the population standard deviation of observed scores and

ρ_{XX} is the population reliability coefficient.

For a sample of examinees, an estimate of the SEM, when the reliability coefficient is estimated via Coefficient Alpha, is

$$Est(SEM) = S_X \sqrt{1 - \alpha} ,$$

where S_X is the sample standard deviation of observed scores.

4.4. Item Response Theory Conditional SEM

Unlike the classical SEM, the conditional SEM based on Item Response Theory (IRT) is not the same value across test scores. For example, if a person gets either a few or a large number of items correct (i.e., scores at the extremes of the score distribution), the conditional standard error will be greater in value than it will be if the person gets a moderate number of items correct. The conditional SEM (on the scale score metric) at each score point for the 2009 WLPT-II (FORM A) is presented in the raw score to scaled score conversion tables in Tables A1 to A20 in Appendix A.

4.5. Inter-Rater Reliability

Another source of measurement error occurs during the evaluation of student work. Inter-rater reliability investigates the extent to which examinees would obtain the same score if the assessment task is scored by different scorers. One way to estimate this type of reliability is to have two raters score each student's paper and then obtain the correlation between scores. In this case, reliability is defined as similarity of students' rank orderings by two raters. Another way to obtain evidence of inter-rater reliability is to calculate the percent agreement between raters. If raters always agree in their assignment of scores, there is 100% agreement. If raters never agree in their assignment of scores, there is 0% agreement. The choice between using a correlation

coefficient or percent agreement depends on whether students' absolute (actual) or relative (rank order) score level is important for a particular interpretation and use. If the actual score is more important, interjudge agreement is the appropriate statistic. If rank order is all that matters, correlations between scores provided by different raters is the appropriate statistic. The Scoring section (Section 3.2) of this report provides the results on inter-rater agreement for WLPT-II.

4.6. Reliability of the Modalities

Table 8 provides raw score descriptive statistics and alpha coefficients by grade for the four main modalities, for the composite (total) test score, and for the Comprehension score (the combination of Listening and Reading). Table 8 includes the following information for each grade level tested:

- Number of items (N Items)
- Maximum raw score observed
- Maximum raw score possible
- Number of students included in the analysis (*N*)
- Average raw score (Mean)
- Standard deviation of raw scores (SD)
- Cronbach's Alpha estimate of internal consistency reliability (reliability estimate)
- CTT Standard error of measurement (SEM)
- Spearman-Brown Predicted Reliability

For the Listening modality of WLPT-II, the Cronbach alpha reliability ranged from 0.50 to 0.77 across grades with a median of 0.68, whereas for the Reading modality it ranged between 0.77 and 0.86 with a median of 0.81. For the Speaking modality the Cronbach alpha reliability ranged from 0.88 to 0.94 with a median of .90, and for the Writing modality, it ranged from 0.78 to 0.84 with a median of 0.80. Generally speaking, the Speaking modality showed higher Cronbach alpha reliability estimates than the other modalities for all grades. The Cronbach alpha reliability of the Comprehension score ranged from .81 to .88 with a median of 0.85. The Cronbach alpha reliability of the overall test was consistently high over all grades, ranging from 0.92 to 0.94, with a median of 0.92.

As mentioned above, test length can affect estimates of score reliability. The Listening test had the fewest number of points, which contributed to its lower reliability estimates. In general, the median reliability estimates for the Reading, Listening, and Writing scores were below that which is preferred. The reliability estimates for the Speaking, Comprehension, and total test scores were in an appropriate range. Because of the relatively lower reliability estimates, caution should be used when making any score based inferences from the listening test scores at all grade levels. Caution should also be used when making score based inferences about the Reading and Writing test scores.

In order to interpret the reliabilities of subtests with different test length based on a common test length, the Spearman-Brown prophecy formula was used to estimate what the reliability would be if the number of items were increased by factor k. In Table 8, the factor k was determined by the multiplier associated with the increased test length that allowed the number of test items be equivalent to the number of items in the Composite score.

$$r_{kk} = \frac{kr_{11}}{1+(k-1)r_{11}}$$

where,
k is the multiplier associated with the increased test length and
kr_{11} is the known reliability of the given test length

Table 8: Descriptive Statistics and Reliability by Grade and Modality

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Cronbach Reliability	SEM	Spearman-Brown Predicted Reliability[h]
K	Composite[c]	84	115	112	13,153	53.97	15.46	0.92	4.46	0.92
	Listening	20	20	20	13,153	15.01	3.09	0.73	1.59	0.92
	Reading	24	24	24	13,153	4.35	3.71	0.80	1.66	0.93
	Speaking	17	38	38	13,153	24.99	9.09	0.92	2.52	0.98
	Writing	23	33	32	13,153	9.63	4.66	0.78	2.19	0.93
	Comprehension[d]	44	44	43	13,153	19.35	5.36	0.81	2.37	0.89
	Social[e]	37	58	58	13,153	40.00	11.02	0.91	3.28	0.96
	Academic[f]	47	57	56	13,153	13.97	7.52	0.86	2.78	0.92
	Productive[g]	25	56	54	13,153	30.86	10.63	0.91	3.13	0.97
1	Composite[c]	84	115	113	13,183	77.16	15.30	0.92	4.25	0.92
	Listening	20	20	20	13,183	16.88	1.91	0.55	1.28	0.84
	Reading	24	24	24	13,183	10.48	5.12	0.84	2.04	0.95
	Speaking	17	38	38	13,183	30.78	6.78	0.90	2.17	0.98
	Writing	23	33	33	13,183	19.02	5.77	0.84	2.34	0.95
	Comprehension[d]	44	44	44	13,183	27.36	5.97	0.83	2.48	0.90
	Social[e]	37	58	58	13,183	47.66	7.80	0.88	2.73	0.94
	Academic[f]	47	57	57	13,183	29.50	10.02	0.90	3.13	0.94
	Productive[g]	25	56	56	13,183	41.69	8.68	0.89	2.82	0.97
2	Composite[c]	84	115	115	11,320	91.28	13.68	0.92	3.78	0.92
	Listening	20	20	20	11,320	17.59	1.60	0.50	1.13	0.81
	Reading	24	24	24	11,320	16.26	5.19	0.86	1.93	0.96
	Speaking	17	38	38	11,320	33.27	5.40	0.88	1.85	0.97
	Writing	23	33	33	11,320	24.16	4.94	0.82	2.12	0.94
	Comprehension[d]	44	44	44	11,320	33.85	5.96	0.85	2.31	0.92
	Social[e]	37	58	58	11,320	50.86	6.25	0.86	2.32	0.93
	Academic[f]	47	57	57	11,320	40.42	9.39	0.91	2.89	0.94
	Productive[g]	25	56	56	11,320	46.57	7.02	0.87	2.49	0.96

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Cronbach Reliability	SEM	Spearman-Brown Predicted Reliability[h]
	Composite[c]	83	110	108	7,605	76.57	13.77	0.92	3.96	0.92
	Listening	20	20	20	7,605	13.82	3.53	0.73	1.84	0.92
	Reading	24	24	24	7,605	12.83	4.14	0.77	2.00	0.92
	Speaking	17	38	38	7,605	33.44	5.17	0.88	1.78	0.97
3	Writing	22	28	28	7,605	16.48	4.53	0.80	2.02	0.94
	Comprehension[d]	44	44	44	7,605	26.65	6.78	0.84	2.74	0.91
	Social[e]	37	58	58	7,605	47.26	7.53	0.88	2.65	0.94
	Academic[f]	46	52	51	7,605	29.31	7.89	0.87	2.86	0.92
	Productive[g]	19	46	46	7,605	37.09	5.75	0.88	2.00	0.97
	Composite[c]	83	110	109	6,732	83.21	13.14	0.92	3.74	0.92
	Listening	20	20	20	6,732	15.14	3.24	0.71	1.73	0.91
	Reading	24	24	24	6,732	15.01	4.35	0.80	1.94	0.93
	Speaking	17	38	38	6,732	34.36	4.61	0.88	1.62	0.97
4	Writing	22	28	28	6,732	18.69	4.26	0.79	1.94	0.94
	Comprehension[d]	44	44	43	6,732	30.16	6.78	0.85	2.62	0.91
	Social[e]	37	58	58	6,732	49.51	6.80	0.87	2.45	0.94
	Academic[f]	46	52	51	6,732	33.71	7.83	0.88	2.76	0.93
	Productive[g]	19	46	46	6,732	38.60	5.21	0.87	1.89	0.97
	Composite[c]	83	110	109	5,611	87.30	13.38	0.93	3.58	0.93
	Listening	20	20	20	5,611	15.89	3.16	0.73	1.63	0.92
	Reading	24	24	24	5,611	16.65	4.36	0.82	1.85	0.94
	Speaking	17	38	38	5,611	34.66	4.57	0.88	1.56	0.97
5	Writing	22	28	28	5,611	20.09	4.20	0.80	1.88	0.94
	Comprehension[d]	44	44	44	5,611	32.55	6.80	0.87	2.49	0.92
	Social[e]	37	58	58	5,611	50.56	6.85	0.88	2.33	0.94
	Academic[f]	46	52	52	5,611	36.74	7.85	0.89	2.66	0.93
	Productive[g]	19	46	46	5,611	39.30	5.24	0.87	1.86	0.97

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Cronbach Reliability	SEM	Spearman-Brown Predicted Reliability[h]
	Composite[c]	92	119	116	5,010	89.61	13.63	0.92	3.86	0.92
	Listening	20	20	20	5,010	16.02	2.76	0.68	1.55	0.91
	Reading	29	29	29	5,010	17.40	4.73	0.79	2.18	0.92
	Speaking	17	38	38	5,010	34.60	4.85	0.90	1.55	0.98
6	Writing	26	32	31	5,010	21.58	4.58	0.79	2.10	0.93
	Comprehension[d]	49	49	48	5,010	33.42	6.66	0.84	2.70	0.91
	Social[e]	37	58	58	5,010	50.62	6.69	0.88	2.29	0.95
	Academic[f]	55	61	60	5,010	38.98	8.56	0.87	3.04	0.92
	Productive[g]	19	46	46	5,010	39.86	5.56	0.89	1.82	0.98
	Composite[c]	92	119	115	3,842	90.35	15.16	0.93	3.89	0.93
	Listening	20	20	20	3,842	15.96	2.95	0.72	1.56	0.92
	Reading	29	29	29	3,842	18.07	4.95	0.81	2.17	0.93
	Speaking	17	38	38	3,842	34.19	5.63	0.92	1.60	0.98
7	Writing	26	32	32	3,842	22.13	4.69	0.80	2.09	0.93
	Comprehension[d]	49	49	49	3,842	34.03	7.12	0.86	2.70	0.92
	Social[e]	37	58	58	3,842	50.15	7.73	0.91	2.36	0.96
	Academic[f]	55	61	59	3,842	40.19	8.91	0.88	3.03	0.93
	Productive[g]	19	46	46	3,842	39.65	6.47	0.92	1.84	0.98
	Composite[c]	92	119	117	3,319	91.09	16.60	0.94	3.91	0.94
	Listening	20	20	20	3,319	16.04	3.19	0.77	1.54	0.94
	Reading	29	29	29	3,319	18.76	5.25	0.83	2.14	0.94
	Speaking	17	38	38	3,319	33.84	6.09	0.93	1.67	0.99
8	Writing	26	32	32	3,319	22.45	4.95	0.82	2.11	0.94
	Comprehension[d]	49	49	49	3,319	34.80	7.69	0.88	2.66	0.93
	Social[e]	37	58	58	3,319	49.87	8.50	0.92	2.40	0.97
	Academic[f]	55	61	61	3,319	41.21	9.49	0.90	3.02	0.94
	Productive[g]	19	46	46	3,319	39.46	7.03	0.93	1.91	0.98

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Cronbach Reliability	SEM	Spearman-Brown Predicted Reliability[h]
	Composite[c]	94	121	118	3,749	85.83	17.62	0.94	4.26	0.94
	Listening	20	20	20	3,749	13.02	2.87	0.64	1.73	0.89
	Reading	31	31	31	3,749	18.77	5.30	0.81	2.29	0.93
	Speaking	17	38	38	3,749	32.59	7.28	0.94	1.78	0.99
9	Writing	26	32	32	3,749	21.45	5.26	0.82	2.25	0.94
	Comprehension[d]	51	51	49	3,749	31.79	7.36	0.85	2.89	0.91
	Social[e]	37	58	58	3,749	45.61	9.19	0.91	2.69	0.96
	Academic[f]	57	63	61	3,749	40.22	9.86	0.89	3.22	0.93
	Productive[g]	19	46	46	3,749	38.35	8.34	0.94	2.05	0.99
	Composite[c]	94	121	119	2,923	87.38	16.96	0.94	4.20	0.94
	Listening	20	20	20	2,923	13.14	2.89	0.65	1.71	0.90
	Reading	31	31	31	2,923	19.54	5.37	0.82	2.27	0.93
	Speaking	17	38	38	2,923	32.71	6.65	0.93	1.79	0.99
10	Writing	26	32	32	2,923	21.99	5.19	0.82	2.23	0.94
	Comprehension[d]	51	51	50	2,923	32.68	7.44	0.85	2.86	0.91
	Social[e]	37	58	58	2,923	45.85	8.56	0.90	2.65	0.96
	Academic[f]	57	63	63	2,923	41.53	9.84	0.89	3.19	0.93
	Productive[g]	19	46	46	2,923	38.65	7.67	0.93	2.07	0.98
	Composite[c]	94	121	118	2,287	90.83	15.50	0.93	4.02	0.93
	Listening	20	20	20	2,287	13.57	2.80	0.64	1.69	0.89
	Reading	31	31	31	2,287	20.62	5.14	0.82	2.21	0.93
	Speaking	17	38	38	2,287	33.54	5.64	0.91	1.66	0.98
11	Writing	26	32	32	2,287	23.09	4.98	0.81	2.16	0.94
	Comprehension[d]	51	51	51	2,287	34.20	7.14	0.85	2.80	0.91
	Social[e]	37	58	58	2,287	47.11	7.50	0.89	2.49	0.95
	Academic[f]	57	63	63	2,287	43.71	9.41	0.89	3.10	0.93
	Productive[g]	19	46	46	2,287	39.83	6.50	0.91	1.94	0.98

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Cronbach Reliability	SEM	Spearman-Brown Predicted Reliability[h]
	Composite[c]	94	121	117	1,864	91.94	13.91	0.92	3.97	0.92
	Listening	20	20	20	1,864	13.63	2.62	0.59	1.68	0.87
	Reading	31	31	31	1,864	21.13	4.85	0.80	2.19	0.92
	Speaking	17	38	38	1,864	33.68	5.11	0.90	1.65	0.98
12	Writing	26	32	32	1,864	23.50	4.56	0.78	2.12	0.93
	Comprehension[d]	51	51	49	1,864	34.76	6.61	0.82	2.78	0.90
	Social[e]	37	58	57	1,864	47.31	6.72	0.87	2.46	0.94
	Academic[f]	57	63	63	1,864	44.63	8.63	0.87	3.06	0.92
	Productive[g]	19	46	46	1,864	40.17	5.83	0.89	1.90	0.98

2

a Maximum points possible

b Maximum points observed

c Composite score is based on Listening, Reading, Speaking, and Writing subtest items

d Comprehension score is based on Listening and Reading subtest items

e Social score is based on Listening and Speaking subtest items

f Academic score is based on Writing and Reading subtest items

g Productive score is based on Writing CR and Speaking subtest items

h In order to interpret the reliabilities of subtests with different test length based on a common test length, the Spearman-Brown prophecy formula was used to estimate what the reliability would be if the number of items were increased by factor *k*: $r_{kk} = \frac{kr_{11}}{1+(k-1)r_{11}}$.

5. VALIDITY OF INFERENCES MADE FROM TEST SCORES

Any assessments constructed using the Pearson ELP item bank adhere to the validity-related standards set forth in the Standards for Educational and Psychological Testing. The judgments about the validity of scores for these assessments are based on the following sources of evidence of validity from the *Stanford English Language Proficiency Test Technical Manual, 2005, Harcourt Assessment, Inc.* (now Pearson):

- Test content—"…a critical part of the item review process included the appropriateness of the match of the item to the instructional standard being assessed." (p. 23)
- Internal structure—"Harcourt Assessment (now Pearson) examined the fit between the way the construct (theoretical attribute) was assessed and the way students were able to respond." (p. 24)
- Relationships to other variables—"…analyses of the relationship of test scores to variables external to the test." (p. 24)

5.1. Test Content Validity

Evidence for the validity of scores, based on test content, is demonstrated by the extent to which the material on the test represents the skills, knowledge, and understanding of the domain tested. As part of the development of the Pearson ELP item bank, writers were trained to write items aligning with the instructional standards set forth in the test blueprint. In addition, a critical part of the item review process included examining how well the item matched the instructional standard being assessed. Only those items relating specifically to an instructional standard were included in the test forms.

The 2009 WLPT-II (Form A) items (original Form A SELP and augmented items) were reviewed by Pearson ESL experts, OSPI ESL staff, and Washington State ESL professionals through bias and sensitivity reviews, an alignment study, and item writing meetings. Only those items meeting the specific intent of the Washington State ELD standards were selected. Several SELP items were slightly revised to incorporate the committees' recommendations. All augmented items on the test met the requests of the committees, including the state alignment committee, and were approved as appropriate by OSPI.

For the 2009 WLPT-II (Form A) test to appropriately align with the Washington State ELD standards, the items in the Pearson ELP item bank were reviewed to match the instructional standards for each grade span. The item mapping functioned as item maps for creating a majority of the test items and offered concrete evidence for the alignment to the Washington State ELD standards. Details of the item alignment study can be found in the *Washington Language Proficiency Test – II Form A Technical Report (2005 – 2006 School Year)*.

5.2. Internal Structure of WLPT-II

An English language proficiency test should detect performance and proficiency differences among students. In developing the structure of the test forms, assessment specialists examined the construct being assessed in terms of how it was assessed and how students were able to respond. Content experts examined the test blueprints and items to be sure the test would logically relate to the most current empirical and theoretical understanding of the constructs being assessed. To examine how consistently each item functions with the overall intent of the test, point-biserial and point-polyserial correlation coefficients were calculated, revealing how well an item discriminates between low- and high-achieving students. The evidence for the validity of the internal structure of the 2009 WLPT-II test is also depicted by the point-biserial correlation and point-polyserial correlation coefficients (item-total correlations), which are contained in Tables B1 – B13 in Appendix B.

In addition to discriminating between low- and high-achieving students, it is important that test modalities perform well together. An assessment procedure should not be a random collection of assessment tasks or test questions. Each task in the assessment should contribute positively to the total result. The interrelationship among the tasks on an assessment is known as the internal structure of the assessment. Typical questions that investigate the relationships among assessment parts include (Nitko, 2004):

- Do all of the assessment tasks "work together" so that each task contributes positively toward assessing the quality of interest?
- If different parts of the assessment procedure are to provide unique information, do the results support this uniqueness?
- If different parts of the assessment procedure are to provide the same or similar information, do the results support this?

To investigate the answers to these questions, correlations were obtained among the four modalities. Table 9 presents the intercorrelations among the four modalities by grade.

Students in grades K – 2 showed low correlations between spoken English (Listening/Speaking) and written English (Reading/Writing). Such outcomes were not surprising considering that students in this age group do not usually read or write well yet, but can have listening and speaking skills. Generally speaking, the correlations between modalities were relatively higher for grades 3 – 12 than grades K – 2. This indicates that the construct validity of the test became stronger for higher grades than Primary grades.

Table 9: Intercorrelations Among Modalities by Grade

Grade	Modality	Listening	Reading	Speaking	Writing
K	Listening	1.00			
	Reading	0.23	1.00		
	Speaking	0.52	0.18	1.00	
	Writing	0.43	0.61	0.37	1.00
1	Listening	1.00			
	Reading	0.29	1.00		
	Speaking	0.43	0.30	1.00	
	Writing	0.43	0.69	0.47	1.00
2	Listening	1.00			
	Reading	0.36	1.00		
	Speaking	0.42	0.39	1.00	
	Writing	0.42	0.72	0.48	1.00
3	Listening	1.00			
	Reading	0.56	1.00		
	Speaking	0.48	0.35	1.00	
	Writing	0.58	0.66	0.44	1.00
4	Listening	1.00			
	Reading	0.59	1.00		
	Speaking	0.49	0.38	1.00	
	Writing	0.57	0.65	0.45	1.00
5	Listening	1.00			
	Reading	0.63	1.00		
	Speaking	0.55	0.43	1.00	
	Writing	0.62	0.68	0.51	1.00
6	Listening	1.00			
	Reading	0.55	1.00		
	Speaking	0.51	0.37	1.00	
	Writing	0.58	0.69	0.49	1.00
7	Listening	1.00			
	Reading	0.59	1.00		
	Speaking	0.58	0.44	1.00	
	Writing	0.64	0.71	0.59	1.00
8	Listening	1.00			
	Reading	0.64	1.00		
	Speaking	0.64	0.49	1.00	
	Writing	0.69	0.73	0.64	1.00

Grade	Modality	Listening	Reading	Speaking	Writing
9	Listening	1.00			
	Reading	0.58	1.00		
	Speaking	0.55	0.53	1.00	
	Writing	0.62	0.74	0.67	1.00
10	Listening	1.00			
	Reading	0.59	1.00		
	Speaking	0.54	0.53	1.00	
	Writing	0.62	0.74	0.63	1.00
11	Listening	1.00			
	Reading	0.58	1.00		
	Speaking	0.52	0.49	1.00	
	Writing	0.62	0.73	0.59	1.00
12	Listening	1.00			
	Reading	0.52	1.00		
	Speaking	0.45	0.43	1.00	
	Writing	0.55	0.68	0.55	1.00

Note: The restriction of the range of scores on the modalities could have resulted in the attenuation of the correlation coefficients between any two modalities.

5.3. Evidence of Unidimensionality of WLPT-II

The unidimensionality of a test can also be examined to provide evidence for the valid internal structure or construct validity. Pearson has adopted the Rasch model (Rasch, 1980) for dichotomous items and the partial credit model (Masters, 1982) for polytomous items as the underlying Item Response Theory (IRT) models for establishing the WLPT-II scale. As with other IRT models, these models assume unidimensionality, in that a single latent trait underlies test performance. In the case of the WLPT-II, the latent trait is English language skills.

To check the unidimensionality assumption for the WLPT-II, a principal component analysis (Stevens, 1996) was conducted for each of the four grade spans. For the purposes of testing unidimensionality, the datasets from the calibration and scaling were used. These calibration datasets comprised the entire WA state population who were administered the 2009 WLPT-II. After eliminating anomalies and other exclusion criteria used in the equating process, approximately 96 percent of the total testing population from 2009 was represented.

Polychoric correlation coefficients were utilized because the items were scored either dichotomously or polytomously. To interpret the results with regard to test unidimensionality, the first and second principal component eigenvalues were compared without rotation. Table 10 summarizes this comparison for each grade span.

Table 10: Principal Component Eigenvalues by Grade Span

Grade Span	Component Number	Eigenvalue	Eigenvalue Ratio
Primary: Grades K-2	1	20.42	3.51
	2	5.82	
Elementary: Grades 3-5	1	13.77	3.31
	2	4.16	
Middle Grades: Grades 6-8	1	15.93	3.71
	2	4.30	
High School: Grades 9-12	1	16.82	4.49
	2	3.75	

The generally accepted standard for determining the unidimensionality of a test requires the eigenvalue of the first component or factor to be at least three times larger than the second component or factor (Hattie, 1985). The observed eigenvalue ratios ranged from 3.31 to 4.49. Thus, this criterion was satisfied at each grade span.

6. CLASSICAL ITEM-LEVEL AND MODALITY-LEVEL STATISTICS

6.1. Item-Level Statistics

The item-level statistics for the 2009 WLPT-II (FORM A) are presented by grade level in Tables B1 – B13 in Appendix B by grade. The following item information and statistics are presented for each item by grade[3]:

- Modality
- Item Sequence
- Item Mean
- Item-Total correlation

6.2. Composite-Level Statistics by Ethnicity and Home Language

Table 11 and Table 12 contain summary statistics on the total test (composite) score by ethnicity and by native language for each grade. For presentation purposes, ethnicity was recoded to have six categories, including the four most populous Washington State ethnic groups: Asian, Black/African, Hispanic, and Caucasian. Students reporting an ethnicity but not belonging to any of these four groups were categorized into Other. Students who had missing values on ethnicity were grouped as Unidentified.

Home language was also recoded to have eight categories, including the six most populous languages among non-English speakers in Washington State: Spanish, Russian, Vietnamese, Ukrainian, Korean, and Tagalog. Similar to ethnicity, Other represents those marking a language as any other than one of these six languages, while Unidentified represents missing values on language.

The statistics shown in each table are as follows:

- Total number of items (N Items**)**
- Maximum score observed
- Maximum score possible
- Minimum score observed
- Number of students (*N*)
- Average raw score (Mean)
- Standard deviation of raw scores (SD)

Table 11 presents descriptive statistics by grade and ethnicity. As Table 11 shows, looking across all grades and ethnicities, the raw score means of grades K and 1 (both in the Primary level), and grade 3 (the first grade in the Elementary level) are comparatively lower than the other grades. In addition, it can be seen that performance increases dramatically between grades K and 1 and again between grades 1 and 2. These increases are expected due to the large gains in cognitive ability for students progressing through grades K through 2. Grade 3 is the first grade in the next level, so the raw score mean presented in the table can not be directly compared with the means presented for grades K-2. However, similar to the Primary level, the raw score means generally increase for higher grade-levels within the Elementary level. This pattern is not always found in the Middle and High School levels. This may be an artifact of the population of students that are

[3] The item difficulty, infit, and outfit are also presented in Appendix B. These values are described later in this report.

included in the higher levels of the assessment. Students have the opportunity to transition out of the program each year, so students that remain in the program are those that did not demonstrate enough language ability, as measured by the WLPT II, to transition out.

With regards to ethnic representation, Hispanic is the largest ethnicity group across all grades. Asian and Caucasian students generally achieved better than the other ethnicities across all grades.

Table 12 presents the descriptive statistics by grade and language. As can be seen from the table, Spanish is the largest language group. There is also a large group of students with an unidentified language. Korean and Tagalog generally performed slightly higher than the other language groups when differences were observed.

Table 11: Descriptive Statistics by Grade and Ethnicity

Grade	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
K	Black/African	84	115	104	0	448	57.61	14.41
	Asian	84	115	106	0	2,112	61.32	16.19
	Caucasian	84	115	103	0	1,380	55.19	15.78
	Hispanic	84	115	104	0	7,361	52.49	13.97
	Other	84	115	106	15	329	54.57	14.31
	Unidentified	84	115	112	0	1,523	48.66	17.58
1	Black/African	84	115	110	21	469	79.37	14.56
	Asian	84	115	113	21	1,987	83.87	15.39
	Caucasian	84	115	112	0	1,486	79.06	15.08
	Hispanic	84	115	112	0	7,990	75.02	14.55
	Other	84	115	108	24	237	77.91	14.55
	Unidentified	84	115	111	10	1,014	76.92	17.33
2	Black/African	84	115	112	27	337	89.39	14.41
	Asian	84	115	115	19	1,450	95.58	12.63
	Caucasian	84	115	114	11	1,232	94.09	12.50
	Hispanic	84	115	114	0	7,133	90.16	13.39
	Other	84	115	111	39	272	92.02	12.56
	Unidentified	84	115	115	13	896	89.88	16.85
3	Black/African	83	110	100	20	264	73.69	14.46
	Asian	83	110	104	8	847	78.62	15.06
	Caucasian	83	110	108	20	739	79.64	12.92
	Hispanic	83	110	105	0	4,955	76.23	13.05
	Other	83	110	100	19	204	76.19	13.23
	Unidentified	83	110	105	0	596	74.09	17.13
4	Black/African	83	110	104	28	237	80.53	13.07
	Asian	83	110	106	5	773	84.77	13.99
	Caucasian	83	110	109	3	682	85.30	13.05
	Hispanic	83	110	105	0	4,453	83.07	12.64
	Other	83	110	103	46	123	81.24	12.56
	Unidentified	83	110	105	0	464	80.84	15.75
5	Black/African	83	110	107	39	220	82.77	15.27
	Asian	83	110	108	16	707	88.20	14.60
	Caucasian	83	110	107	35	529	88.65	12.38
	Hispanic	83	110	109	0	3,655	87.56	12.60
	Other	83	110	107	36	111	86.37	13.83
	Unidentified	83	110	106	0	389	84.19	16.85
6	Black/African	92	119	110	20	190	84.66	16.34
	Asian	92	119	115	0	612	90.80	14.66
	Caucasian	92	119	116	30	444	91.21	13.40
	Hispanic	92	119	115	0	3,233	89.59	12.82
	Other	92	119	107	8	81	85.22	16.68
	Unidentified	92	119	112	25	450	89.40	15.43
7	Black/African	92	119	110	48	182	86.26	14.46
	Asian	92	119	115	0	438	89.03	18.61
	Caucasian	92	119	114	38	355	91.72	14.22
	Hispanic	92	119	115	22	2,394	90.64	14.42
	Other	92	119	110	53	68	88.75	13.71
	Unidentified	92	119	115	0	405	90.91	16.27
8	Black/African	92	119	110	0	156	82.96	18.73
	Asian	92	119	117	0	517	93.28	16.13
	Caucasian	92	119	114	33	272	92.68	14.88
	Hispanic	92	119	116	0	2,037	91.21	16.06
	Other	92	119	115	35	51	89.45	15.24
	Unidentified	92	119	116	0	286	89.42	20.00

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Grade	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
9	Black/African	94	121	109	6	194	81.26	18.27
	Asian	94	121	118	18	626	87.76	16.74
	Caucasian	94	121	116	19	296	89.60	14.88
	Hispanic	94	121	114	0	2,166	85.50	17.68
	Other	94	121	112	46	63	86.62	15.03
	Unidentified	94	121	113	0	404	83.91	19.65
10	Black/African	94	121	112	20	213	79.12	19.39
	Asian	94	121	116	25	494	91.47	14.41
	Caucasian	94	121	115	25	259	90.38	14.53
	Hispanic	94	121	114	0	1,629	87.08	16.73
	Other	94	121	114	57	52	89.33	12.76
	Unidentified	94	121	119	0	276	85.02	20.27
11	Black/African	94	121	114	42	167	86.31	15.23
	Asian	94	121	116	0	425	93.79	14.11
	Caucasian	94	121	118	28	260	90.74	15.48
	Hispanic	94	121	118	0	1,204	90.46	15.60
	Other	94	121	107	52	33	90.94	14.04
	Unidentified	94	121	116	37	198	90.57	17.10
12	Black/African	94	121	115	0	168	87.80	15.40
	Asian	94	121	115	0	362	93.39	14.02
	Caucasian	94	121	116	36	220	90.75	14.55
	Hispanic	94	121	116	0	953	92.64	12.71
	Other	94	121	108	58	35	88.37	13.93
	Unidentified	94	121	117	31	126	91.08	17.51

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Table 12: Descriptive Statistics by Grade and Language

Grade	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
K	Spanish	84	115	104	0	7,496	52.61	13.93
	Russian	84	115	103	0	609	52.28	15.77
	Vietnamese	84	115	106	0	692	57.88	15.63
	Ukrainian	84	115	94	4	324	52.45	14.28
	Korean	84	115	103	0	202	66.31	14.86
	Tagalog	84	115	87	23	141	60.35	12.89
	Other	84	115	112	0	1,415	48.35	17.74
	Unidentified	84	115	106	0	2,274	59.96	16.28
1	Spanish	84	115	112	0	8,066	75.12	14.52
	Russian	84	115	112	0	695	78.33	15.21
	Vietnamese	84	115	113	23	684	82.82	14.57
	Ukrainian	84	115	107	16	375	77.34	14.64
	Korean	84	115	111	50	193	89.62	13.11
	Tagalog	84	115	108	30	159	81.52	14.80
	Other	84	115	111	10	928	76.75	17.58
	Unidentified	84	115	112	3	2,083	81.49	15.79
2	Spanish	84	115	114	0	7,248	90.23	13.34
	Russian	84	115	114	34	576	94.32	12.06
	Vietnamese	84	115	114	19	461	95.82	12.09
	Ukrainian	84	115	113	46	311	93.74	11.87
	Korean	84	115	112	56	131	100.31	10.62
	Tagalog	84	115	113	32	150	95.79	12.01
	Other	84	115	115	13	822	89.57	17.05
	Unidentified	84	115	115	11	1,621	92.86	13.90
3	Spanish	83	110	105	0	4,998	76.33	12.98
	Russian	83	110	108	20	367	80.03	13.27
	Vietnamese	83	110	101	18	253	76.89	15.10
	Ukrainian	83	110	103	27	198	79.32	12.99
	Korean	83	110	102	44	80	85.38	12.29
	Tagalog	83	110	99	42	74	80.19	12.34
	Other	83	110	105	0	565	74.05	17.18
	Unidentified	83	110	104	8	1,070	76.36	14.91
4	Spanish	83	110	105	0	4,472	83.12	12.56
	Russian	83	110	103	25	297	86.01	12.44
	Vietnamese	83	110	105	26	174	83.91	15.39
	Ukrainian	83	110	106	11	181	85.83	12.28
	Korean	83	110	106	5	96	86.24	15.32
	Tagalog	83	110	101	50	103	86.32	9.67
	Other	83	110	105	0	446	80.69	15.82
	Unidentified	83	110	109	3	963	82.70	14.07
5	Spanish	83	110	109	0	3,656	87.61	12.51
	Russian	83	110	106	45	238	89.45	10.71
	Vietnamese	83	110	105	28	161	89.53	13.55
	Ukrainian	83	110	106	35	146	88.49	13.05
	Korean	83	110	108	17	90	88.83	16.89
	Tagalog	83	110	107	49	79	90.23	10.27
	Other	83	110	106	0	379	84.27	16.75
	Unidentified	83	110	107	16	862	85.64	15.31
6	Spanish	92	119	115	0	3,220	89.74	12.49
	Russian	92	119	112	30	194	89.81	14.96
	Vietnamese	92	119	113	0	135	89.48	16.79
	Ukrainian	92	119	113	53	122	91.08	12.23
	Korean	92	119	115	62	74	96.30	12.15
	Tagalog	92	119	114	0	79	92.28	15.02
	Other	92	119	112	25	427	90.13	14.89
	Unidentified	92	119	116	8	759	87.56	16.18

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Grade	Language	N Items	Max Pointsa	Max Pointsb	Min Pointsc	N	Mean	SD
7	Spanish	92	119	115	22	2,378	90.77	14.36
	Russian	92	119	114	38	158	91.35	14.92
	Vietnamese	92	119	112	0	106	85.33	20.54
	Ukrainian	92	119	112	42	92	89.79	15.42
	Korean	92	119	115	48	66	94.64	14.39
	Tagalog	92	119	114	59	48	95.23	11.92
	Other	92	119	115	0	387	90.79	16.51
	Unidentified	92	119	115	18	607	88.23	16.15
8	Spanish	92	119	116	0	2,022	91.37	15.85
	Russian	92	119	113	33	137	90.87	16.31
	Vietnamese	92	119	114	0	91	90.58	18.38
	Ukrainian	92	119	114	44	68	95.62	13.25
	Korean	92	119	117	48	88	96.53	15.15
	Tagalog	92	119	116	11	65	95.45	15.66
	Other	92	119	116	0	278	89.38	20.05
	Unidentified	92	119	117	0	570	89.14	17.41
9	Spanish	94	121	114	0	2,150	85.65	17.51
	Russian	94	121	113	19	147	89.63	15.15
	Vietnamese	94	121	110	32	124	84.58	18.33
	Ukrainian	94	121	113	32	82	90.79	14.10
	Korean	94	121	115	18	108	91.78	15.88
	Tagalog	94	121	118	67	70	93.21	11.49
	Other	94	121	112	0	391	83.82	19.69
	Unidentified	94	121	116	6	677	84.65	17.68
10	Spanish	94	121	114	0	1,616	87.30	16.54
	Russian	94	121	115	44	130	90.52	14.47
	Vietnamese	94	121	116	41	111	88.78	13.97
	Ukrainian	94	121	108	59	65	91.66	11.69
	Korean	94	121	115	61	79	96.57	11.36
	Tagalog	94	121	115	61	62	95.97	10.72
	Other	94	121	119	0	261	85.07	20.55
	Unidentified	94	121	116	20	599	85.10	18.12
11	Spanish	94	121	118	0	1,205	90.59	15.57
	Russian	94	121	118	53	117	92.26	14.04
	Vietnamese	94	121	115	56	88	92.44	12.70
	Ukrainian	94	121	115	30	77	88.73	15.68
	Korean	94	121	116	69	57	99.40	10.04
	Tagalog	94	121	114	72	51	98.31	9.36
	Other	94	121	116	37	185	90.44	17.23
	Unidentified	94	121	115	0	507	89.53	15.90
12	Spanish	94	121	116	0	956	92.70	12.67
	Russian	94	121	112	54	98	89.62	14.23
	Vietnamese	94	121	113	53	80	91.81	13.09
	Ukrainian	94	121	109	48	55	93.49	13.29
	Korean	94	121	114	72	42	100.55	9.23
	Tagalog	94	121	114	78	38	98.58	7.75
	Other	94	121	117	31	121	90.92	17.76
	Unidentified	94	121	116	0	474	89.69	15.42

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

6.3. Modality-Level Descriptive Statistics

Table 8 showed the classical statistics of central tendency, variability, and score precision for the four modality scores, as well as the overall, composite score.

Table 13 and Table 14 present the following summary statistics by grade span and ethnicity, and by grade span and language for the four modalities (as well as Comprehension), respectively:

- Number of items (N Items)
- Maximum points possible
- Maximum points observed
- Minimum points observed
- Number of students (*N*)
- Average raw score (Mean)
- Standard deviation of raw scores (SD)

Table 13 presents the descriptive statistics by grade span and ethnicity for each modality. As can be seen from the table, performance across ethnic groups was relatively similar. When differences were found, Asian and/or Caucasian groups achieved slightly higher than the other ethnicities.

Table 14 presents the descriptive statistics by grade span and language. Again, Spanish is the largest language group for all grade spans. Korean and Tagalog generally performed better than other language groups.

Table 13: Descriptive Statistics by Grade Span and Ethnicity for Modalities

Grade Span	Modality	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Primary (Grades K-2)	Composite[d]	Black/African	84	115	112	0	1,254	74.29	19.48
		Asian	84	115	115	0	5,549	78.35	20.62
		Caucasian	84	115	114	0	4,098	75.54	21.46
		Hispanic	84	115	114	0	22,484	72.45	20.70
		Other	84	115	111	15	838	73.33	21.18
		Unidentified	84	115	115	0	3,433	67.76	24.78
	Listening	Black/African	20	20	20	0	1,254	16.25	2.58
		Asian	20	20	20	0	5,549	16.97	2.17
		Caucasian	20	20	20	0	4,098	16.66	2.58
		Hispanic	20	20	20	0	22,484	16.38	2.47
		Other	20	20	20	0	838	16.56	2.39
		Unidentified	20	20	20	0	3,433	15.72	3.43
	Reading	Black/African	24	24	24	0	1,254	9.79	6.59
		Asian	24	24	24	0	5,549	11.50	7.15
		Caucasian	24	24	24	0	4,098	10.18	6.89
		Hispanic	24	24	24	0	22,484	9.85	6.49
		Other	24	24	24	0	838	9.66	7.11
		Unidentified	24	24	24	0	3,433	9.30	6.84
	Speaking	Black/African	17	38	38	0	1,254	31.02	6.73
		Asian	17	38	38	0	5,549	30.48	7.22
		Caucasian	17	38	38	0	4,098	30.43	7.73
		Hispanic	17	38	38	0	22,484	29.37	8.01
		Other	17	38	38	0	838	29.97	7.34
		Unidentified	17	38	38	0	3,433	27.07	10.28
	Writing	Black/African	23	33	33	0	1,254	17.23	7.53
		Asian	23	33	33	0	5,549	19.41	7.87
		Caucasian	23	33	33	0	4,098	18.27	8.04
		Hispanic	23	33	33	0	22,484	16.84	7.69
		Other	23	33	33	0	838	17.14	8.15
		Unidentified	23	33	33	0	3,433	15.66	8.39
	Comprehension[e]	Black/African	44	44	44	0	1,254	26.04	8.11
		Asian	44	44	44	0	5,549	28.46	8.35
		Caucasian	44	44	44	0	4,098	26.84	8.39
		Hispanic	44	44	44	0	22,484	26.24	7.93
		Other	44	44	44	0	838	26.22	8.51
		Unidentified	44	44	44	0	3,433	25.03	9.05

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Elementary (Grades 3-5)	Composite[d]	Black/African	83	110	107	20	721	78.71	14.79
		Asian	83	110	108	5	2,327	83.57	15.10
		Caucasian	83	110	109	3	1,950	84.06	13.34
		Hispanic	83	110	109	0	13,063	81.73	13.60
		Other	83	110	107	19	438	80.19	13.82
		Unidentified	83	110	106	0	1,449	78.97	17.15
	Listening	Black/African	20	20	20	2	721	13.91	3.73
		Asian	20	20	20	0	2,327	15.12	3.62
		Caucasian	20	20	20	0	1,950	15.44	3.30
		Hispanic	20	20	20	0	13,063	14.84	3.33
		Other	20	20	20	1	438	14.43	3.67
		Unidentified	20	20	20	0	1,449	14.34	3.92
	Reading	Black/African	24	24	24	0	721	13.75	4.84
		Asian	24	24	24	0	2,327	15.54	4.61
		Caucasian	24	24	24	0	1,950	14.93	4.53
		Hispanic	24	24	24	0	13,063	14.58	4.45
		Other	24	24	24	1	438	13.89	4.61
		Unidentified	24	24	24	0	1,449	14.09	4.94
	Speaking	Black/African	17	38	38	12	721	33.82	4.37
		Asian	17	38	38	2	2,327	33.56	5.13
		Caucasian	17	38	38	0	1,950	34.68	4.40
		Hispanic	17	38	38	0	13,063	34.24	4.68
		Other	17	38	38	12	438	33.95	4.35
		Unidentified	17	38	38	0	1,449	33.06	6.40
	Writing	Black/African	22	28	28	0	721	17.23	4.93
		Asian	22	28	28	0	2,327	19.35	4.63
		Caucasian	22	28	28	0	1,950	19.01	4.34
		Hispanic	22	28	28	0	13,063	18.08	4.48
		Other	22	28	28	3	438	17.93	4.75
		Unidentified	22	28	27	0	1,449	17.48	5.16
	Comprehension[e]	Black/African	44	44	44	2	721	27.66	7.74
		Asian	44	44	44	0	2,327	30.67	7.51
		Caucasian	44	44	44	3	1,950	30.37	7.02
		Hispanic	44	44	44	0	13,063	29.41	6.98
		Other	44	44	43	3	438	28.32	7.42
		Unidentified	44	44	44	0	1,449	28.43	8.08

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Middle Grades (Grades 6-8)	Composite[d]	Black/African	92	119	110	0	528	84.71	16.51
		Asian	92	119	117	0	1,567	91.12	16.40
		Caucasian	92	119	116	30	1,071	91.75	14.06
		Hispanic	92	119	116	0	7,664	90.35	14.26
		Other	92	119	115	8	200	87.50	15.40
		Unidentified	92	119	116	0	1,141	89.94	16.97
	Listening	Black/African	20	20	20	0	528	14.87	3.60
		Asian	20	20	20	0	1,567	16.23	3.10
		Caucasian	20	20	20	4	1,071	16.29	2.86
		Hispanic	20	20	20	0	7,664	16.03	2.78
		Other	20	20	20	0	200	15.53	3.02
		Unidentified	20	20	20	0	1,141	15.91	3.32
	Reading	Black/African	29	29	28	0	528	16.10	5.34
		Asian	29	29	29	0	1,567	19.34	5.31
		Caucasian	29	29	29	4	1,071	18.50	4.94
		Hispanic	29	29	29	0	7,664	17.76	4.76
		Other	29	29	29	1	200	16.82	5.26
		Unidentified	29	29	29	0	1,141	18.18	5.22
	Speaking	Black/African	17	38	38	0	528	33.28	5.16
		Asian	17	38	38	0	1,567	32.98	5.80
		Caucasian	17	38	38	5	1,071	34.62	4.56
		Hispanic	17	38	38	0	7,664	34.61	5.36
		Other	17	38	38	6	200	33.73	5.05
		Unidentified	17	38	38	0	1,141	33.90	6.30
	Writing	Black/African	26	32	30	0	528	20.46	5.49
		Asian	26	32	32	0	1,567	22.57	5.06
		Caucasian	26	32	31	5	1,071	22.33	4.60
		Hispanic	26	32	31	0	7,664	21.95	4.53
		Other	26	32	30	1	200	21.43	5.10
		Unidentified	26	32	31	0	1,141	21.96	5.10
	Comprehension[e]	Black/African	49	49	46	0	528	30.97	8.13
		Asian	49	49	49	0	1,567	35.57	7.79
		Caucasian	49	49	48	10	1,071	34.80	7.03
		Hispanic	49	49	48	0	7,664	33.79	6.70
		Other	49	49	47	1	200	32.35	7.53
		Unidentified	49	49	49	0	1,141	34.09	7.72

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
High School (Grades 9-12)	Composite[d]	Black/African	94	121	115	0	742	83.26	17.68
		Asian	94	121	118	0	1,907	91.13	15.28
		Caucasian	94	121	118	19	1,035	90.32	14.87
		Hispanic	94	121	118	0	5,952	88.08	16.51
		Other	94	121	114	46	183	88.50	14.00
		Unidentified	94	121	119	0	1,004	86.43	19.30
	Listening	Black/African	20	20	20	0	742	12.11	3.15
		Asian	20	20	20	0	1,907	13.58	2.71
		Caucasian	20	20	20	3	1,035	13.68	2.59
		Hispanic	20	20	20	0	5,952	13.28	2.78
		Other	20	20	20	5	183	13.42	2.82
		Unidentified	20	20	20	0	1,004	13.08	3.13
	Reading	Black/African	31	31	31	0	742	17.91	5.93
		Asian	31	31	31	0	1,907	21.46	5.27
		Caucasian	31	31	30	2	1,035	20.23	5.12
		Hispanic	31	31	31	0	5,952	19.48	5.05
		Other	31	31	30	8	183	19.30	5.01
		Unidentified	31	31	31	0	1,004	19.34	5.62
	Speaking	Black/African	17	38	38	0	742	32.42	5.79
		Asian	17	38	38	0	1,907	32.49	5.55
		Caucasian	17	38	38	3	1,035	33.56	5.43
		Hispanic	17	38	38	0	5,952	33.25	6.72
		Other	17	38	38	15	183	34.01	4.42
		Unidentified	17	38	38	0	1,004	32.26	7.95
	Writing	Black/African	26	32	32	0	742	20.82	5.78
		Asian	26	32	32	0	1,907	23.59	4.87
		Caucasian	26	32	32	2	1,035	22.86	4.78
		Hispanic	26	32	32	0	5,952	22.07	5.00
		Other	26	32	31	8	183	21.79	4.84
		Unidentified	26	32	32	0	1,004	21.75	5.70
	Comprehension[e]	Black/African	51	51	47	0	742	30.02	8.26
		Asian	51	51	51	0	1,907	35.05	7.19
		Caucasian	51	51	49	6	1,035	33.91	6.96
		Hispanic	51	51	49	0	5,952	32.76	6.98
		Other	51	51	49	14	183	32.71	6.95
		Unidentified	51	51	50	0	1,004	32.42	7.96

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Table 14: Descriptive Statistics by Grade Span and Language

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Primary (Grades K-2)	Composite[d]	Spanish	84	115	114	0	22,810	72.53	20.67
		Russian	84	115	114	0	1,880	74.79	22.28
		Vietnamese	84	115	114	0	1,837	76.68	21.14
		Ukrainian	84	115	113	4	1,010	74.41	21.47
		Korean	84	115	112	0	526	83.33	19.32
		Tagalog	84	115	113	23	450	79.65	19.55
		Other	84	115	115	0	3,165	67.38	24.94
		Unidentified	84	115	115	0	5,978	76.38	20.62
	Listening	Spanish	20	20	20	0	22,810	16.40	2.46
		Russian	20	20	20	0	1,880	16.56	2.69
		Vietnamese	20	20	20	0	1,837	16.89	2.34
		Ukrainian	20	20	20	2	1,010	16.44	2.66
		Korean	20	20	20	0	526	17.35	1.91
		Tagalog	20	20	20	7	450	17.07	1.77
		Other	20	20	20	0	3,165	15.67	3.48
		Unidentified	20	20	20	0	5,978	16.71	2.37
	Reading	Spanish	24	24	24	0	22,810	9.86	6.50
		Russian	24	24	24	0	1,880	10.10	6.94
		Vietnamese	24	24	24	0	1,837	10.99	7.05
		Ukrainian	24	24	24	0	1,010	9.93	6.63
		Korean	24	24	24	0	526	13.16	7.34
		Tagalog	24	24	24	0	450	12.01	7.17
		Other	24	24	24	0	3,165	9.29	6.81
		Unidentified	24	24	24	0	5,978	10.65	7.04
	Speaking	Spanish	17	38	38	0	22,810	29.42	7.98
		Russian	17	38	38	0	1,880	29.90	8.20
		Vietnamese	17	38	38	0	1,837	30.07	7.52
		Ukrainian	17	38	38	0	1,010	29.80	7.79
		Korean	17	38	38	0	526	31.30	6.81
		Tagalog	17	38	38	4	450	30.54	6.81
		Other	17	38	38	0	3,165	26.85	10.40
		Unidentified	17	38	38	0	5,978	30.65	7.18
	Writing	Spanish	23	33	33	0	22,810	16.85	7.69
		Russian	23	33	33	0	1,880	18.24	8.15
		Vietnamese	23	33	33	0	1,837	18.73	8.06
		Ukrainian	23	33	33	0	1,010	18.23	7.96
		Korean	23	33	33	0	526	21.53	7.32
		Tagalog	23	33	33	3	450	20.02	7.61
		Other	23	33	33	0	3,165	15.57	8.40
		Unidentified	23	33	33	0	5,978	18.38	7.89
	Comprehension[e]	Spanish	44	44	44	0	22,810	26.25	7.93
		Russian	44	44	44	0	1,880	26.66	8.53
		Vietnamese	44	44	44	0	1,837	27.89	8.32
		Ukrainian	44	44	43	2	1,010	26.37	8.22
		Korean	44	44	44	0	526	30.51	8.29
		Tagalog	44	44	44	9	450	29.09	8.05
		Other	44	44	44	0	3,165	24.96	9.08
		Unidentified	44	44	44	0	5,978	27.35	8.41

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Elementary (Grades 3-5)	Composite[d]	Spanish	83	110	109	0	13,126	81.79	13.52
		Russian	83	110	108	20	902	84.49	12.96
		Vietnamese	83	110	105	18	588	82.43	15.67
		Ukrainian	83	110	106	11	525	84.11	13.32
		Korean	83	110	108	5	266	86.86	15.07
		Tagalog	83	110	107	42	256	85.75	11.34
		Other	83	110	106	0	1,390	78.96	17.17
		Unidentified	83	110	109	3	2,895	81.23	15.26
	Listening	Spanish	20	20	20	0	13,126	14.85	3.31
		Russian	20	20	20	3	902	15.58	3.15
		Vietnamese	20	20	20	2	588	15.05	3.58
		Ukrainian	20	20	20	0	525	15.57	3.35
		Korean	20	20	20	0	266	15.72	3.53
		Tagalog	20	20	20	7	256	15.71	2.87
		Other	20	20	20	0	1,390	14.34	3.92
		Unidentified	20	20	20	0	2,895	14.55	3.78
	Reading	Spanish	24	24	24	0	13,126	14.58	4.45
		Russian	24	24	24	1	902	15.13	4.56
		Vietnamese	24	24	24	1	588	15.15	4.54
		Ukrainian	24	24	24	1	525	14.85	4.29
		Korean	24	24	24	0	266	17.20	4.64
		Tagalog	24	24	23	3	256	15.82	4.06
		Other	24	24	24	0	1,390	14.11	4.95
		Unidentified	24	24	24	0	2,895	14.56	4.77
	Speaking	Spanish	17	38	38	0	13,126	34.27	4.64
		Russian	17	38	38	0	902	34.71	4.44
		Vietnamese	17	38	38	2	588	33.38	5.74
		Ukrainian	17	38	38	4	525	34.51	4.58
		Korean	17	38	38	2	266	32.98	4.98
		Tagalog	17	38	38	18	256	34.18	3.94
		Other	17	38	38	0	1,390	33.05	6.45
		Unidentified	17	38	38	0	2,895	33.78	4.82
	Writing	Spanish	22	28	28	0	13,126	18.09	4.47
		Russian	22	28	28	0	902	19.07	4.22
		Vietnamese	22	28	28	3	588	18.84	4.69
		Ukrainian	22	28	27	1	525	19.18	4.17
		Korean	22	28	28	0	266	20.96	4.59
		Tagalog	22	28	28	8	256	20.04	3.67
		Other	22	28	27	0	1,390	17.47	5.16
		Unidentified	22	28	28	0	2,895	18.34	4.89
	Comprehension[e]	Spanish	44	44	44	0	13,126	29.43	6.96
		Russian	44	44	44	6	902	30.71	6.83
		Vietnamese	44	44	44	3	588	30.20	7.41
		Ukrainian	44	44	44	3	525	30.42	6.85
		Korean	44	44	44	0	266	32.92	7.53
		Tagalog	44	44	42	10	256	31.54	6.27
		Other	44	44	44	0	1,390	28.44	8.10
		Unidentified	44	44	44	1	2,895	29.11	7.78

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
Middle Grades (Grades 6-8)	Composite[d]	Spanish	92	119	116	0	7,620	90.49	14.05
		Russian	92	119	114	30	489	90.61	15.32
		Vietnamese	92	119	114	0	332	88.46	18.57
		Ukrainian	92	119	114	42	282	91.76	13.72
		Korean	92	119	117	48	228	95.91	13.98
		Tagalog	92	119	116	0	192	94.09	14.55
		Other	92	119	116	0	1,092	90.17	16.89
		Unidentified	92	119	117	0	1,936	88.24	16.55
	Listening	Spanish	20	20	20	0	7,620	16.05	2.76
		Russian	20	20	20	4	489	16.22	3.08
		Vietnamese	20	20	20	0	332	15.81	3.53
		Ukrainian	20	20	20	4	282	16.34	2.78
		Korean	20	20	20	5	228	17.43	2.44
		Tagalog	20	20	20	0	192	16.35	2.81
		Other	20	20	20	0	1,092	15.93	3.34
		Unidentified	20	20	20	0	1,936	15.61	3.27
	Reading	Spanish	29	29	29	0	7,620	17.78	4.75
		Russian	29	29	29	4	489	18.27	5.12
		Vietnamese	29	29	29	0	332	18.83	5.49
		Ukrainian	29	29	28	5	282	18.55	4.70
		Korean	29	29	29	5	228	21.94	4.77
		Tagalog	29	29	29	0	192	19.54	4.88
		Other	29	29	29	0	1,092	18.27	5.20
		Unidentified	29	29	29	0	1,936	17.67	5.40
	Speaking	Spanish	17	38	38	0	7,620	34.67	5.25
		Russian	17	38	38	5	489	34.26	5.17
		Vietnamese	17	38	38	0	332	31.70	7.18
		Ukrainian	17	38	38	6	282	34.53	4.52
		Korean	17	38	38	9	228	33.25	4.72
		Tagalog	17	38	38	0	192	34.51	4.80
		Other	17	38	38	0	1,092	33.95	6.27
		Unidentified	17	38	38	0	1,936	33.33	5.56
	Writing	Spanish	26	32	31	0	7,620	21.99	4.49
		Russian	26	32	31	5	489	21.86	4.77
		Vietnamese	26	32	32	0	332	22.13	5.10
		Ukrainian	26	32	30	8	282	22.33	4.51
		Korean	26	32	31	10	228	23.29	4.73
		Tagalog	26	32	32	0	192	23.69	4.66
		Other	26	32	31	0	1,092	22.02	5.05
		Unidentified	26	32	32	0	1,936	21.62	5.35
	Comprehension[e]	Spanish	49	49	48	0	7,620	33.84	6.65
		Russian	49	49	48	10	489	34.49	7.50
		Vietnamese	49	49	47	0	332	34.64	8.45
		Ukrainian	49	49	48	15	282	34.89	6.66
		Korean	49	49	49	15	228	39.36	6.64
		Tagalog	49	49	48	0	192	35.89	6.98
		Other	49	49	49	0	1,092	34.20	7.72
		Unidentified	49	49	48	0	1,936	33.28	7.95

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
High School (Grades 9-12)	Composite[d]	Spanish	94	121	118	0	5,927	88.24	16.37
		Russian	94	121	118	19	492	90.49	14.53
		Vietnamese	94	121	116	32	403	88.89	15.33
		Ukrainian	94	121	115	30	279	90.96	13.91
		Korean	94	121	116	18	286	95.91	13.20
		Tagalog	94	121	118	61	221	96.09	10.40
		Other	94	121	119	0	958	86.34	19.45
		Unidentified	94	121	116	0	2,257	86.92	17.11
	Listening	Spanish	20	20	20	0	5,927	13.30	2.77
		Russian	20	20	20	3	492	13.67	2.54
		Vietnamese	20	20	19	6	403	13.12	2.73
		Ukrainian	20	20	19	4	279	13.86	2.39
		Korean	20	20	19	5	286	14.15	2.18
		Tagalog	20	20	20	6	221	14.17	2.36
		Other	20	20	20	0	958	13.07	3.16
		Unidentified	20	20	20	0	2,257	12.96	3.03
	Reading	Spanish	31	31	31	0	5,927	19.51	5.05
		Russian	31	31	30	3	492	20.48	5.10
		Vietnamese	31	31	30	6	403	21.28	4.69
		Ukrainian	31	31	29	2	279	20.47	4.56
		Korean	31	31	31	4	286	23.69	4.45
		Tagalog	31	31	31	10	221	22.58	4.59
		Other	31	31	31	0	958	19.29	5.64
		Unidentified	31	31	31	0	2,257	19.40	5.79
	Speaking	Spanish	17	38	38	0	5,927	33.32	6.63
		Russian	17	38	38	3	492	33.41	5.35
		Vietnamese	17	38	38	0	403	30.96	6.22
		Ukrainian	17	38	38	6	279	33.63	5.55
		Korean	17	38	38	4	286	33.11	4.56
		Tagalog	17	38	38	22	221	34.57	3.33
		Other	17	38	38	0	958	32.23	8.01
		Unidentified	17	38	38	0	2,257	32.58	5.92
	Writing	Spanish	26	32	32	0	5,927	22.11	4.97
		Russian	26	32	31	8	492	22.93	4.61
		Vietnamese	26	32	32	6	403	23.53	4.76
		Ukrainian	26	32	32	7	279	22.99	4.56
		Korean	26	32	32	5	286	24.95	4.68
		Tagalog	26	32	32	12	221	24.76	3.62
		Other	26	32	32	0	958	21.74	5.72
		Unidentified	26	32	32	0	2,257	21.99	5.48
	Comprehension[e]	Spanish	51	51	49	0	5,927	32.82	6.96
		Russian	51	51	49	6	492	34.15	6.92
		Vietnamese	51	51	48	12	403	34.40	6.61
		Ukrainian	51	51	46	10	279	34.33	6.11
		Korean	51	51	48	9	286	37.84	5.92
		Tagalog	51	51	49	19	221	36.75	5.87
		Other	51	51	50	0	958	32.36	8.02
		Unidentified	51	51	51	0	2,257	32.36	8.06

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing modality items.

[e] Comprehension score is based on Listening and Reading modality items.

7. CALIBRATION, EQUATING, AND SCALING

7.1. Background

The WLPT-II scale scores were derived within the framework of Item Response Theory (IRT). IRT is widely used because it promotes equity of results from year to year, through what has been referred to as test-free measurement. Simply stated, test-free measurement means that, given a student's responses to two exams scaled using IRT, the student will achieve the same scaled score on both exams except for measurement error. This holds true regardless of differences in the overall difficulties of the exams. In other words, measurement is test-free in the sense that the results are dependent only upon the ability of the student and are independent of item difficulties.

The Rasch model (Rasch, 1980) for dichotomous items and the Partial Credit Model (PCM; Masters, 1982) for polytomous items were used to develop, calibrate, equate, and scale WLPT-II. These measurement models are regularly used to construct test forms, for scaling and equating, and to develop and maintain large item banks. All item and test analyses, including item-fit analysis, scaling, equating, diagnosis, and performance prediction were accomplished within this framework. The statistical software used to calibrate and scale WLPT-II was WINSTEPS, Version 3.63 (Linacre, 2006).

7.2. The Rasch and Partial Credit Models

The most basic expression of the Rasch model is the item response function (IRF), which expresses the probability of a correct response to an item as a function of ability level. The probability of a correct response is bounded by 0 (certainty of an incorrect response) and 1 (certainty of a correct response). The ability scale is, in theory, unbounded. In practice, the ability scale tends to range from -5 to +5 logits for heterogeneous ability groups.

As an example, consider Figure 1, which depicts a dichotomously scored item that falls at approximately 0.75 on an ability scale that ranges from -5 to +5 (horizontal axis). The curve (j = 1) shows the probability of obtaining a correct response (a score of 1). When a person answers an item at the same level as his or her ability, that person has a probability of .50 of answering the item correctly. Simply stated, in a group of 100 people, all of whom have an ability of 0.75, we would expect approximately 50% to answer the item correctly. A person whose ability was above 0.75 would have a higher probability of answering the item correctly, while a person whose ability is below 0.75 would have a lower probability of answering the item correctly. This makes intuitive sense and is the basic formulation of Rasch measurement for test items having only 2 possible categories (i.e., correct or incorrect).

Figure 2 extends this formulation to show the probabilities of obtaining an incorrect (score of 0) or correct (score of 1) response. The thick dotted curve (j = 0) shows the probability of getting a score of "0," while the solid curve (j = 1) shows the probability of getting a score of "1." The point at which the two curves cross indicates the transition point on the ability scale where the most likely response changes from a "0" to a "1." Here, the probability of answering the item correctly or incorrectly is .50. The thick dotted curve shows that, of a group of 100 examinees whose ability was greater than .75, less than a 50% would be likely to answer the item incorrectly and, of a group of 100 examinees whose ability was less than .75, more than 50% would be likely to answer the item incorrectly.

Figure 1: Sample Item Characteristic Curve

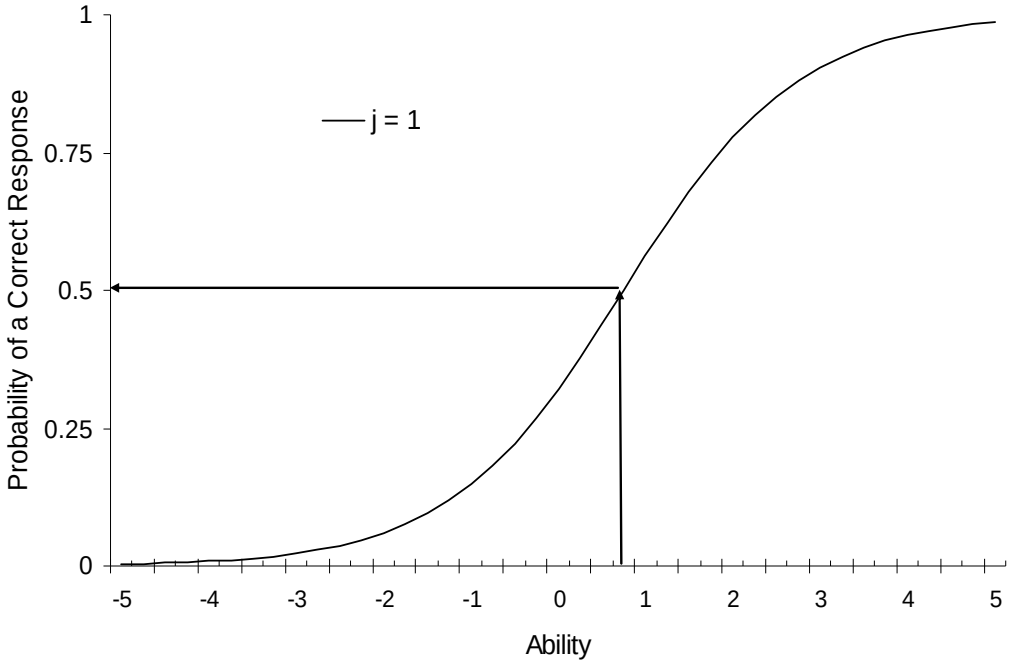
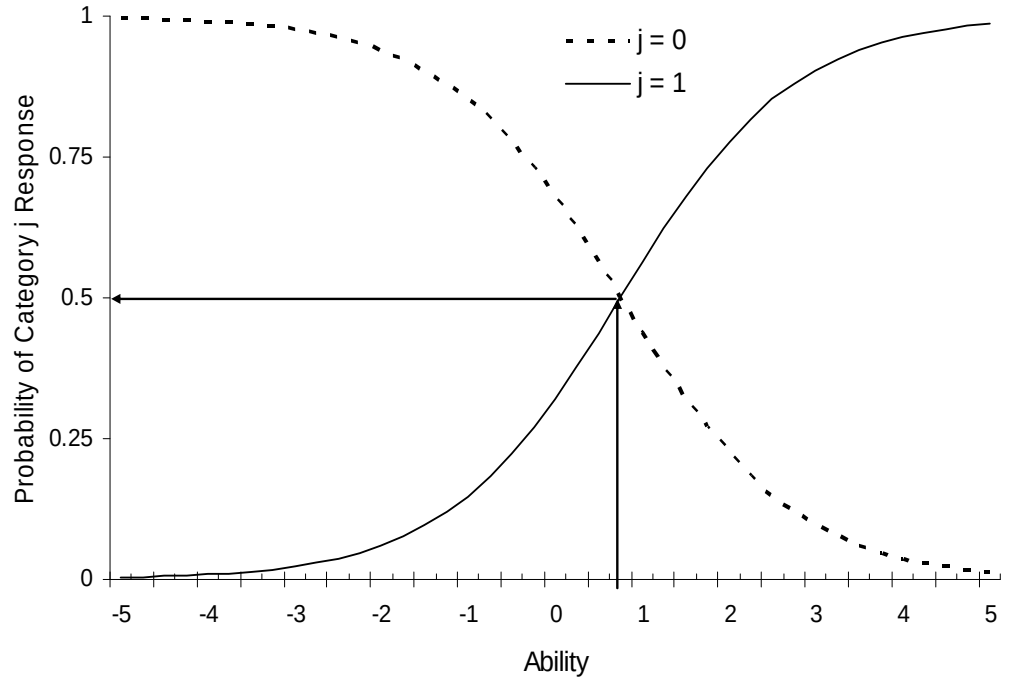


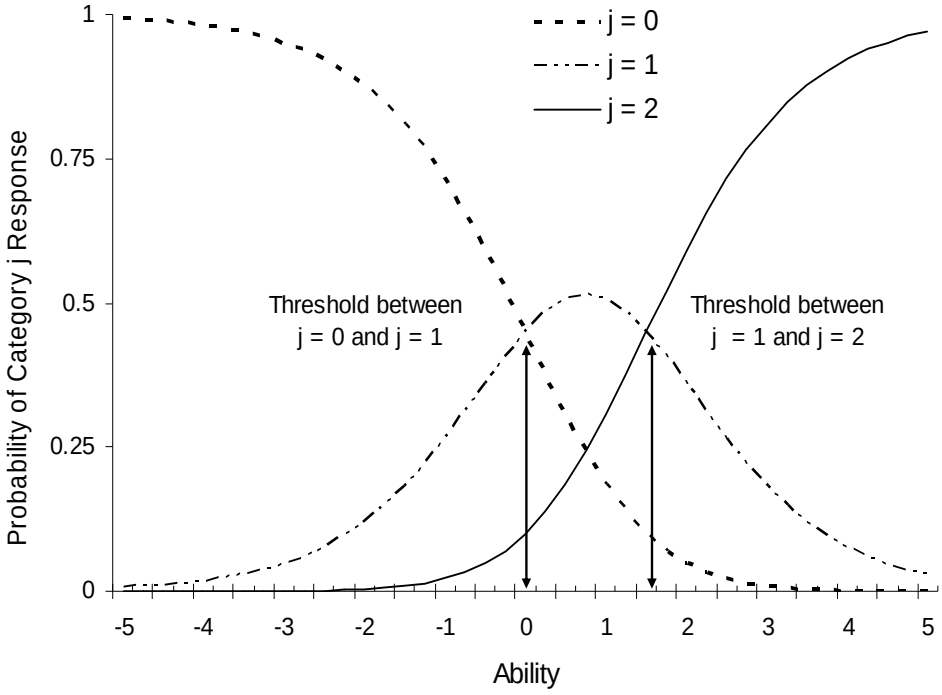
Figure 2: Category Response Curves for a Single-Point Item



The key step in the formulation, and the point at which the Rasch dichotomous model merges with the PCM, comes with the incorporation of additional response categories. Suppose that we add a third category representing responses that, although not totally correct, are still clearly not totally incorrect. An example of the PCM for a polytomous item is illustrated in Figure 3.

The thick dotted curve (j = 0) in Figure 3 represents the probability for examinees getting a score of "0" (completely incorrect) on the item, given their ability. Those of low ability (e.g., below -2) are very likely to be in this category and, in fact, are more likely to be in this category than the other two categories (1 and 2). Those receiving a "1" (partial credit) tend to fall in the middle range of abilities (the long-and-short dotted curve, j = 1). The solid curve (j = 2) represents the probability for those receiving scores of "2" (completely correct). High ability people are clearly more likely to be in category 2 than in any other, but there are still some of low- and average-ability that get full credit on an item.

Figure 3: Category Response Curves for a Two-Point Item



Although the actual computations are more complex, the points at which lines cross in Figure 3 have a similar interpretation as the dichotomous case. Consider the point at which the j=0 line crosses the j=1 line, indicated by the left arrow. For abilities to the left of (or less than) this point, the probability is greatest for a "0" response. To the right of (or above) this point, and up to the point at which the j=1 and j=2 lines cross (marked by the right arrow), the most likely response is a "1." For abilities to the right of this point, the most likely response is a "2." Note that the probability of earning a score of "1" (j=1) decreases as ability either decreases or increases. These points indicated by the two arrows may be thought of as the thresholds of crossing the boundaries between categories.

An important implication of the formulation can be summarized as follows: if the Rasch model for dichotomously-scored items can be thought of as a special case of the PCM, then the act of scaling multiple-choice items together with polytomous items is a straightforward process of

applying the measurement model. The quality of the scaling then can be assessed in terms of known procedures.

One important property of Rasch model and PCM is the separation in estimation of item parameters from person parameters. With either model, total score (given by the sum of the categories in which a person responds) is a sufficient statistic for estimating person ability (i.e., no additional information need be estimated). Additionally, for the PCM, the total number of responses across examinees in a particular category is a sufficient statistic for estimating the step parameter (i.e., category boundary) for that category. Thus with PCM, the same total score will yield the same ability estimate for different examinees.

In terms of the mathematical formulation, the PCM is a direct extension of the expression for the Rasch model. For an item involving M_j score categories, the general expression for the probability of scoring x on item j is given by,

$$P_{xj} = \frac{\exp\left[\sum_{l=0}^{x}\left(\theta - b_{jl}\right)\right]}{\sum_{m=0}^{M_j}\left\{\exp\left[\sum_{k=0}^{m}\left(\theta - b_{jk}\right)\right]\right\}},$$

where $x = 0, 1, ..., M_j$, and,

it is assumed that $\sum_{m=0}^{M_j} b_{jm} = 0$.

According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between θ and b_{jm} of all the completed steps, divided by the sum of the differences of all the steps of a task. Thissen and Steinberg (1986) refer to this model as a divide-by-total model. The parameters estimated by this model are (a) an ability for each person and (b) $M_j - 1$ steps (category boundaries) for each item with M_j score categories.

7.3. Original Calibration, Equating, and Scaling of the WLPT-II

The WLPT-II (Form A) was administered in 2006 and again in 2009. WLPT-II items that were needed to augment the SELP were field tested in quasi-operational status. In 2006, all new items in the Elementary and High School levels were accepted, thus completing Form A for those levels for future administrations. Also in 2006, the Primary level had three items rejected by the data review committee, while the Middle level had one item rejected. This meant that Form A for two of the four levels would require additional calibration, equating, and scaling activities after the 2009 administration. Sections 7.3.1 to 7.3.3 describe the calibration, equating, and scaling activities for Form A using 2006 results. Section 7.4 describes the calibration, equating and scaling of Form A for the 2009 administration.

The WLPT-II (Form A) equating study for 2006 administration results used 100% of the records of the students taking the test within the regular testing window. The data were screened through standardized Quality Control (QC) check processes to ensure that only valid student records were used for the equating study.

7.3.1 Calibration

Calibration, equating, and scaling were based on the overall test score at each grade band (K-2, 3-5, 6-8, and 9-12). An initial set of anchor items from Pearson's SELP item bank was investigated using statistical diagnostic indices that included displacement (Linacre, 2005), Robust-Z (Tenenbaum, Lindsay, Siskind, Wall-Mitchell, & Saunders, 2001), correlation between fixed and free difficulty estimates, the ratio of the standard deviations for fixed and free difficulty estimates, the proportion of anchor items to test length, and b-plots (scatterplots) between fixed versus free difficulty estimates.

The fixed parameter values used for the anchor items were previously obtained from the original calibration of the SELP item bank. During this original calibration of the SELP, the item parameters were adjusted to factor in the appropriate level constant from the SELP vertical scale. For further information on linking the WLPT-II (Form A) to the SELP vertical scale, see the *Washington Language Proficiency Test - II Technical Report* (*2005 - 2006 School Year).*

7.3.2 Equating

Based on the final set of anchor items for each grade span, item parameter estimates and raw score to theta conversion table were obtained from WINSTEPS. Because the fixed parameter values used for the anchor items already incorporated the appropriate SELP vertical scale level constant, the resulting theta estimates from the conversion table were already placed onto the SELP vertical scale. As such, there was no need to add the level constants to the theta estimates.

Item fit statistics (INFIT and OUTFIT) for each grade span, based on the final set of anchor items, are presented in Appendix B. INFIT is a mean square statistic that summarizes the amount of model misfit within ability groups after the misfit from between-ability groups is accounted for. OUTFIT is a mean square statistic summarizing the amount of model misfit between the observed item response function (IRF) and the theoretical IRF under the IRT model. Practically speaking, productive items have INFIT and OUTFIT values between 0.7 and 1.3. Table 15 summarizes the INFIT and OUTFIT values at each grade span.

Table 15: Summary Statistics on the INFIT and OUTFIT Item-Fit Statistics

						Percent of Items Within Productive Range	
	Number	INFIT		OUTFIT			
Grade Span	of Items	M	SD	M	SD	INFIT	OUTFIT
Primary: K-2	84	0.97	0.18	0.95	0.30	93	80
Elementary: 3-5	83	0.99	0.14	1.00	0.23	93	82
Middle Grades: 6-8	92	1.12	0.54	1.12	0.66	79	63
High School: 9-12	94	1.01	0.16	1.02	0.35	97	74

7.3.3 Scaling

In Year 1 of the WLPT-II program, the Lowest Obtainable Scale Score (LOSS), 300, and the Highest Obtainable Scale Score (HOSS), 900, were predetermined by OSPI. Additionally in Year 1, the observed maximum theta (OMXT) and observed minimum theta (OMNT) values in

the raw score to scale score conversion tables across grade bands were identified. The slope and intercept for the linear transformation to convert theta scores to the WLPT-II scale scores were then obtained by solving the following linear system:

$$\text{Slope} = \frac{(\text{HOSS} - \text{LOSS})}{(\text{OMXT} - \text{OMNT})}$$

and

Intercept = LOSS – (SLOPE*OMNT).

The resulting slope and intercept were 36.179 and 603.934, respectively. These slope and intercept values are used to establish the theta (θ) to scale score relationships in all subsequent forms of WLPT-II. Thus, using these slope and intercept values, the final raw score to scale score conversion tables for the total (composite) and modality scores for all grade spans were produced using the following formula:

$$\text{Scaled Score} = 36.179\theta + 603.934,$$

where θ is the theta estimate corresponding to a given total or modality raw score.

7.4. Scaling of the WLPT-II (Form A) for 2009 Administration

The WLPT-II (Form A) was administered in 2006 and 2009. Items needed to augment the SELP were field tested in quasi-operational status during each of these administrations. In 2006, all new items in the Elementary and High School levels were accepted, thus completing Form A for those levels for future administrations. Also in 2006, the Primary level had three items rejected by the data review committee, while the Middle level had one item rejected. This meant that Form A for two of the four levels would require additional calibration, equating, and scaling activities after the 2009 administration.

During analysis of Form B in 2007 the Technical Advisory Committee recommended a change in equating procedures for the WLPT. This change required that the estimation of Rasch item difficulties be weighted by grade level. The weighting was to address the fact the standardization sample for SELP had fairly different proportions of students in each grade as is observed in the Washington population. This change meant that the original WLPT Form A scaling activity needed to be revisited for 2009.

To accomplish this, the original data files used in the 2006 scaling were retrieved, grade-level weights computed, item parameters estimated, and the 2006 scales re-established using the 2007 procedures. For Elementary and High School, the re-post equated 2006 tables were applied directly to the 2009 data. For Primary and Middle levels a few items were new to 2009, and item parameters needed to be estimated. This was accomplished by conducting a single anchored item parameter estimation using 2009 data and 2006 base-lined item parameters. No dropping of anchors was allowed. The raw-to-scale score tables for 2006 were updated to include the new items, and applied to the 2009 student scores in the Primary and Middle levels.

A more detailed and comprehensive description of the WLPT-II (FORM A) equating study for the 2008 – 2009 school year, along with the results and WINSTEPS outputs, can be found in the separate report, *Washington Language Proficiency Test – II Form A Equating Study Report (2008 – 2009 School Year).* Tables A1 – A20 in Appendix A provide the raw to scale score conversion tables for all grade spans.

8. SUMMARY OF OPERATIONAL TEST RESULTS

This section presents scale score and proficiency level summaries of the WLPT-II spring and May administrations.

8.1. Spring Administration of the WLPT-II

Table 16 presents the scale score summary by grade for each modality as well as the overall (composite) test and derived scales. A summary of the conceptual framework of the derived scales is presented below.

It is commonly accepted that language proficiency can be distinguished between comprehension, or the receptive language skills (Listening and Reading), on the one hand, and the productive language skills (Writing and Speaking), on the other hand, based on language use (Canale, 1985; Bachman, 1990). Title III of the federal *No Child Left Behind* (NCLB) Act of 2001 requires assessment of the comprehension proficiency of English Language Learners (ELLs) and reporting a comprehension score, in addition to a score in Listening, Reading, Writing, and Speaking. Accordingly, Comprehension is a reporting category on the Stanford English Language Proficiency (SELP) Exam, the base product of the WLPT II. Further, based on the high-level categorization of language use, Productive is a SELP reporting category as well.

Language use can also be divided into two other broad categories: academic and social. Cummins (1979) introduced the idea of a distinction between language used more commonly in social situations, for which he used the term "basic interpersonal communicative skills" (BICS), and language used more commonly in school/academic settings, which he called "cognitive/academic language proficiency" (CALP). This theoretical distinction has been widely accepted in the field of language acquisition. However, the research on and debate over what constitutes BICS versus what constitutes CALP is still ongoing (e.g., Edelsky, 1990; Edelsky et al., 1983; Martin-Jones & Romaine, 1986; Wiley, 1996).

In the context of assessing the English proficiency of Limited English Proficiency (LEP) students or English Language Learners (ELLs), reading and writing are predominant language use activities and the most essential skills in academic settings (Cheng, 2003). Hence, the Title III of NCLB also requires that Academic language be tested and monitored. Accordingly, the SELP reading and writing subtests assess more academic language and reports an Academic score that include both reading and writing. In contrast to the academic score, the SELP also reports a Social score, which includes Listening and Speaking. To assess the social function of language use, the Listening and Speaking subtests of SELP were designed to test more conversational or functional uses of language. There are no academic lectures that students listen to, or academic topics that students speak on. For this reason, OSPI augmented SELP. The augmented Speaking items that Pearson added to the WLPT-II forms were designed specifically to add school/academic context to the Speaking subtest.

The table includes the following information:

- Number of items (N Items)
- Maximum scale score possible
- Maximum scale score observed
- Minimum scale score observed
- Number of students tested (*N*)
- Average scale score (Mean)
- Standard deviation of scale scores (SD)

Table 16: Descriptive Statistics of the WLPT-II Scale Score (SS) by Grade and Modality[4]

Grade	Modality	N Items	Max SS[a]	Max SS[b]	Min SS[c]	*N*	Mean	SD
K	Composite [d]	84	810	718	300	13153	555.89	30.18
	Listening	20	722	722	305	13153	565.21	50.33
	Reading	24	771	771	415	13153	515.75	55.67
	Speaking	17	734	734	374	13153	578.19	58.02
	Writing	23	779	728	401	13153	550.27	32.17
	Comprehension [e]	44	780	728	303	13153	541.50	36.20
	Social [f]	37	754	754	300	13153	572.32	45.63
	Academic [g]	47	801	750	382	13153	541.06	34.35
	Productive [h]	25	775	698	369	13153	569.32	37.25
1	Composite [d]	84	810	733	300	13183	599.15	31.20
	Listening	20	722	722	305	13183	599.15	41.70
	Reading	24	771	771	415	13183	585.52	48.41
	Speaking	17	734	734	374	13183	617.99	56.65
	Writing	23	779	779	401	13183	604.93	34.16
	Comprehension [e]	44	780	780	303	13183	589.22	36.93
	Social [f]	37	754	754	300	13183	607.76	41.25
	Academic [g]	47	801	801	382	13183	597.06	34.55
	Productive [h]	25	775	775	369	13183	611.25	36.54
2	Composite [d]	84	810	810	300	11320	631.45	33.91
	Listening	20	722	722	305	11320	617.31	42.51
	Reading	24	771	771	415	11320	638.23	53.94
	Speaking	17	734	734	374	11320	641.93	56.69
	Writing	23	779	779	401	11320	638.07	36.15
	Comprehension [e]	44	780	780	303	11320	630.72	42.92
	Social [f]	37	754	754	300	11320	628.64	41.44
	Academic [g]	47	801	801	382	11320	636.17	38.05
	Productive [h]	25	775	775	369	11320	637.12	38.47
3	Composite [d]	83	860	782	370	7605	646.95	28.88
	Listening	20	775	775	430	7605	645.16	40.87
	Reading	24	819	819	419	7605	637.11	41.38
	Speaking	17	773	773	413	7605	679.69	55.71
	Writing	22	834	834	425	7605	644.49	35.05
	Comprehension [e]	44	829	829	398	7605	640.11	34.36
	Social [f]	37	799	799	395	7605	656.66	36.40
	Academic [g]	46	853	801	396	7605	641.30	33.81
	Productive [h]	19	830	830	412	7605	665.03	35.64
4	Composite [d]	83	860	809	370	6732	662.72	30.30
	Listening	20	775	775	430	6732	661.68	42.53
	Reading	24	819	819	419	6732	658.70	44.56
	Speaking	17	773	773	413	6732	692.34	55.85
	Writing	22	834	834	425	6732	662.26	35.31
	Comprehension [e]	44	829	777	398	6732	658.62	36.62
	Social [f]	37	799	799	395	6732	670.22	37.94
	Academic [g]	46	853	801	396	6732	660.15	34.53
	Productive [h]	19	830	830	412	6732	677.79	37.32

[4] [a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Grade	Modality	N Items	Max SSa[5]	Max SSb	Min SSc	N	Mean	SD
	Composite [d]	83	860	809	370	5611	674.31	33.37
	Listening	20	775	775	430	5611	672.59	44.73
	Reading	24	819	819	419	5611	676.19	47.32
	Speaking	17	773	773	413	5611	698.31	56.82
5	Writing	22	834	834	425	5611	675.24	38.92
	Comprehension [e]	44	829	829	398	5611	673.01	40.39
	Social [f]	37	799	799	395	5611	679.03	41.60
	Academic [g]	46	853	853	396	5611	674.51	37.17
	Productive [h]	19	830	830	412	5611	686.13	41.11
	Composite [d]	92	886	794	372	5010	686.98	30.06
	Listening	20	793	793	417	5010	686.37	43.39
	Reading	29	845	845	426	5010	673.69	38.39
	Speaking	17	795	795	417	5010	721.78	56.78
6	Writing	26	862	811	435	5010	693.72	34.58
	Comprehension [e]	49	853	801	395	5010	676.84	34.14
	Social [f]	37	819	819	391	5010	698.40	40.06
	Academic [g]	55	880	829	404	5010	683.42	32.56
	Productive [h]	19	843	843	417	5010	718.57	42.91
	Composite [d]	92	886	783	372	3842	689.69	33.51
	Listening	20	793	793	417	3842	686.20	45.88
	Reading	29	845	845	426	3842	679.25	41.20
	Speaking	17	795	795	417	3842	719.53	60.80
7	Writing	26	862	862	435	3842	698.48	36.49
	Comprehension [e]	49	853	853	395	3842	680.29	37.20
	Social [f]	37	819	819	391	3842	697.52	44.01
	Academic [g]	55	880	803	404	3842	688.33	34.75
	Productive [h]	19	843	843	417	3842	719.54	48.42
	Composite [d]	92	886	809	372	3319	692.27	38.08
	Listening	20	793	793	417	3319	688.37	49.86
	Reading	29	845	845	426	3319	685.13	45.38
	Speaking	17	795	795	417	3319	717.93	63.91
8	Writing	26	862	862	435	3319	701.42	40.10
	Comprehension [e]	49	853	853	395	3319	684.69	42.00
	Social [f]	37	819	819	391	3319	697.79	48.04
	Academic [g]	55	880	880	404	3319	692.51	38.89
	Productive [h]	19	843	843	417	3319	720.84	53.19
	Composite [d]	94	900	807	383	3749	692.75	34.26
	Listening	20	841	841	434	3749	687.84	37.27
	Reading	31	860	860	440	3749	687.64	39.00
	Speaking	17	815	815	417	3749	723.28	70.52
9	Writing	26	863	863	460	3749	698.06	34.23
	Comprehension [e]	51	878	800	411	3749	687.15	33.53
	Social [f]	37	856	856	399	3749	696.81	42.79
	Academic [g]	57	887	808	423	3749	692.64	33.11
	Productive [h]	19	831	831	417	3749	713.53	54.14

[5] [a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Grade	Modality	N Items	Max SSa[6]	Max SSb	Min SSc	N	Mean	SD
10	Composite [d]	94	900	823	383	2923	696.16	33.30
	Listening	20	841	841	434	2923	689.06	37.56
	Reading	31	860	860	440	2923	693.46	40.44
	Speaking	17	815	815	417	2923	722.71	67.77
	Writing	26	863	863	460	2923	702.14	34.63
	Comprehension [e]	51	878	826	411	2923	691.21	34.10
	Social [f]	37	856	856	399	2923	697.80	40.82
	Academic [g]	57	887	887	423	2923	697.35	33.53
	Productive [h]	19	831	831	417	2923	716.09	52.79
11	Composite [d]	94	900	807	383	2287	703.01	32.88
	Listening	20	841	841	434	2287	694.62	37.36
	Reading	31	860	860	440	2287	701.30	39.93
	Speaking	17	815	815	417	2287	729.36	64.02
	Writing	26	863	863	460	2287	710.06	37.20
	Comprehension [e]	51	878	878	411	2287	697.93	33.90
	Social [f]	37	856	856	399	2287	703.73	38.82
	Academic [g]	57	887	887	423	2287	704.75	34.65
	Productive [h]	19	831	831	417	2287	724.22	51.09
12	Composite [d]	94	900	795	383	1864	704.83	30.54
	Listening	20	841	841	434	1864	695.14	34.52
	Reading	31	860	860	440	1864	704.99	38.20
	Speaking	17	815	815	417	1864	728.45	61.40
	Writing	26	863	863	460	1864	712.06	34.13
	Comprehension [e]	51	878	800	411	1864	700.32	31.69
	Social [f]	37	856	804	399	1864	703.70	35.96
	Academic [g]	57	887	887	423	1864	707.57	32.16
	Productive [h]	19	831	831	417	1864	726.05	49.51

Table 17 presents the mean scaled score by level from 2006-2009, as well as, the difference in the mean scaled scores between 2009 and 2008. The same information is presented graphically in
Figure 4. Table 17 presents the mean scaled score by grade from 2006-2009, as well as, the difference in the mean scaled scores between 2009 and 2008. The same information is presented graphically in
Figure 5.

[6] [a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Table 17: Mean Scaled Score by Level from 2006-2009

Level	Mean SS: 2006(A)	Mean SS: 2007(B)	Mean SS: 2008(C)	Mean SS: 2009(A)	SS DIFF (2009-2008)
Level 1	587	589	587	594	7
Level 2	658	647	656	660	4
Level 3	682	690	690	689	-1
Level 4	696	683	703	697	-6

Figure 4: Mean Scaled Score by Level from 2006-2009

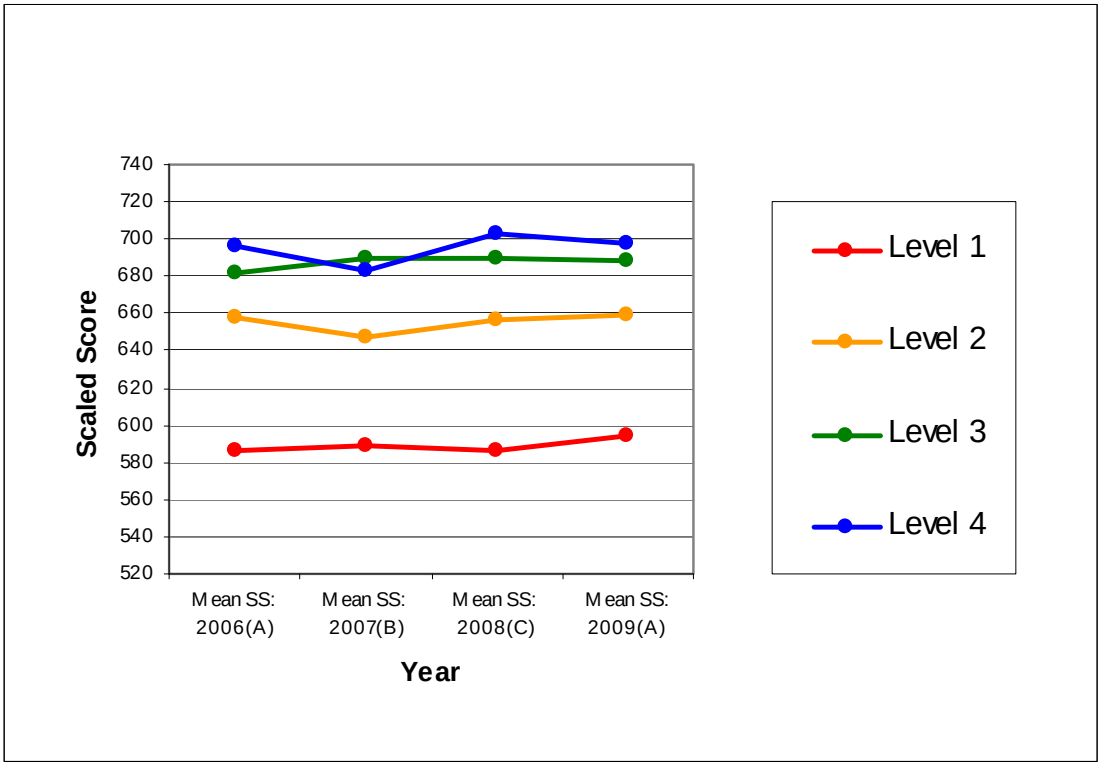


Table 18: Mean Scaled Score by Grade from 2006-2009

Grade	Mean SS: 2006(A)	Mean SS: 2007(B)	Mean SS: 2008(C)	Mean SS: 2009(A)	SS DIFF (2009-2008)
K	552	554	550	556	5
1	588	597	593	599	6
2	620	626	626	631	6
3	648	637	644	647	3
4	663	651	659	662	4
5	668	658	669	674	5
6	676	682	689	687	-2
7	685	691	688	689	1
8	686	698	693	692	-2
9	689	676	695	692	-2
10	700	684	706	695	-11
11	698	688	710	703	-7
12	705	689	709	704	-5

Figure 5: Mean Scaled Score by Grade from 2006-2009

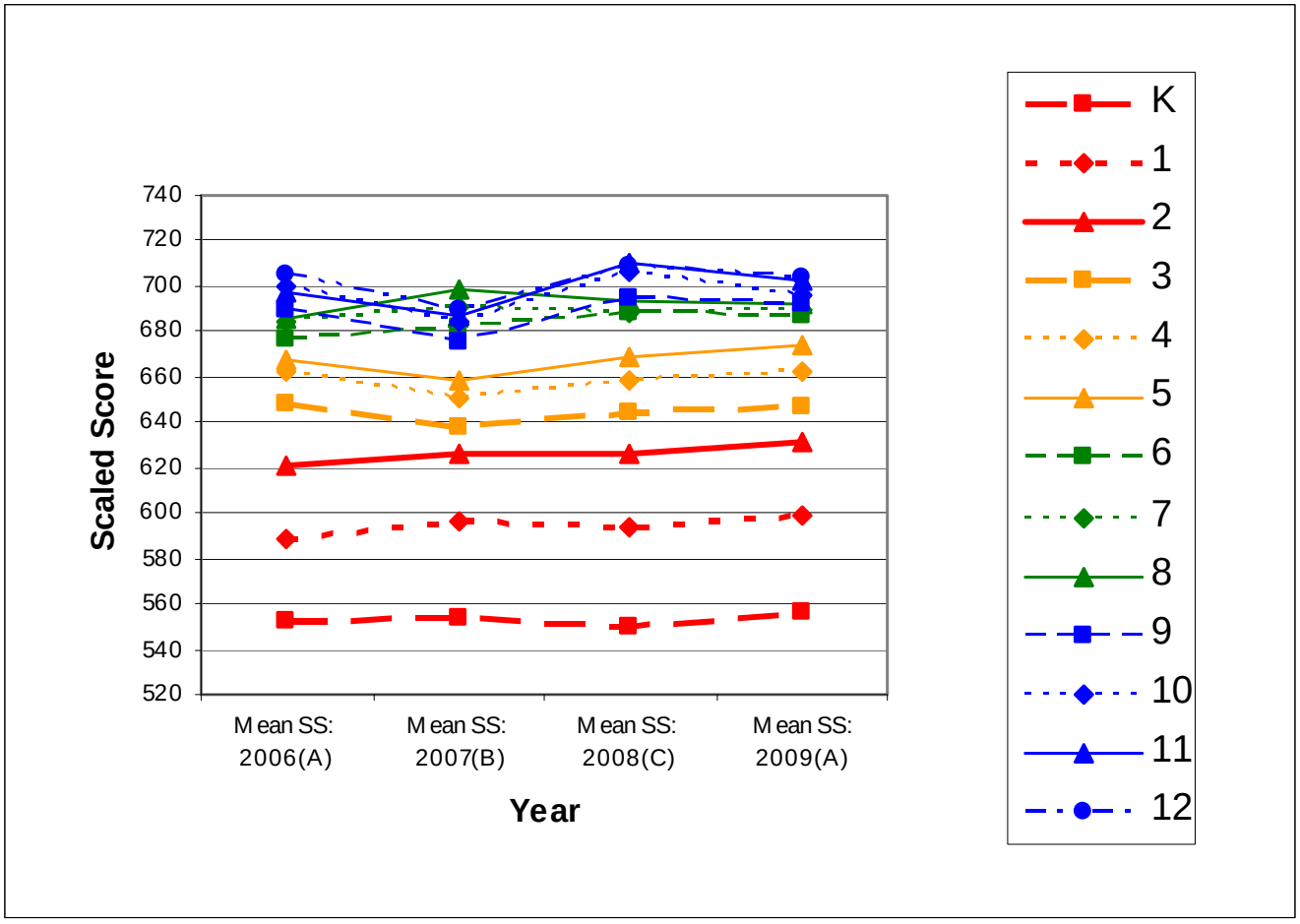


Table 19 contains the percentage of students in each of the proficiency levels by grade for 2006-2009 as well as the change in percentage between 2009 and 2008 for each category. The original adopted WLPT-II overall proficiency cut-scores and the equated 2009 cut-scores can be found in Tables D1 and D2, respectively, in Appendix D.

Table 19: Percentage of Students in Each Proficiency Level by Grade

Grade	Perf. Level	2006 %	2007 %	2008 %	2009 %	Change from 2009 to 2008
Grade K	Transitional	5	7	5	6	1
	Advanced	30	32	27	34	7
	Intermediate	59	54	59	54	-5
	Beg./Adv. Beg.	6	8	9	5	-4
Grade 1	Transitional	11	16	15	17	2
	Advanced	47	52	47	53	6
	Intermediate	40	30	36	28	-8
	Beg./Adv. Beg.	2	2	2	1	-1
Grade 2	Transitional	24	22	26	31	5
	Advanced	50	60	52	51	-1
	Intermediate	24	16	20	17	-3
	Beg./Adv. Beg.	2	2	2	1	-1
Grade 3	Transitional	25	12	20	20	0
	Advanced	60	67	64	66	2
	Intermediate	14	18	15	13	-2
	Beg./Adv. Beg.	1	2	2	1	-1
Grade 4	Transitional	25	12	20	24	4
	Advanced	58	66	62	61	-1
	Intermediate	14	19	16	14	-2
	Beg./Adv. Beg.	2	3	2	1	-1
Grade 5	Transitional	17	8	16	23	7
	Advanced	61	68	65	63	-2
	Intermediate	18	21	16	13	-3
	Beg./Adv. Beg.	4	4	3	2	-1
Grade 6	Transitional	14	20	24	18	-6
	Advanced	69	65	64	71	7
	Intermediate	14	13	10	10	0
	Beg./Adv. Beg.	3	3	2	1	-1
Grade 7	Transitional	14	21	16	15	-1
	Advanced	66	64	65	69	4
	Intermediate	16	12	16	14	-2
	Beg./Adv. Beg.	4	3	3	2	-1
Grade 8	Transitional	11	24	16	16	0
	Advanced	65	58	65	64	-1
	Intermediate	20	14	16	17	1
	Beg./Adv. Beg.	4	3	3	3	0

Grade	Perf. Level	2006 %	2007 %	2008 %	2009 %	Change from 2009 to 2008
Grade 9	Transitional	13	7	18	9	-9
	Advanced	60	60	58	68	10
	Intermediate	22	24	20	20	0
	Beg./Adv. Beg.	5	10	4	2	-2
Grade 10	Transitional	18	10	25	13	-12
	Advanced	59	58	56	65	9
	Intermediate	21	26	17	20	3
	Beg./Adv. Beg.	2	6	2	2	0
Grade 11	Transitional	13	10	24	15	-9
	Advanced	62	58	61	69	8
	Intermediate	22	27	13	15	2
	Beg./Adv. Beg.	2	4	1	1	0
Grade 12	Transitional	14	8	20	10	-10
	Advanced	66	63	64	74	10
	Intermediate	18	26	15	15	0
	Beg./Adv. Beg.	1	3	1	1	0

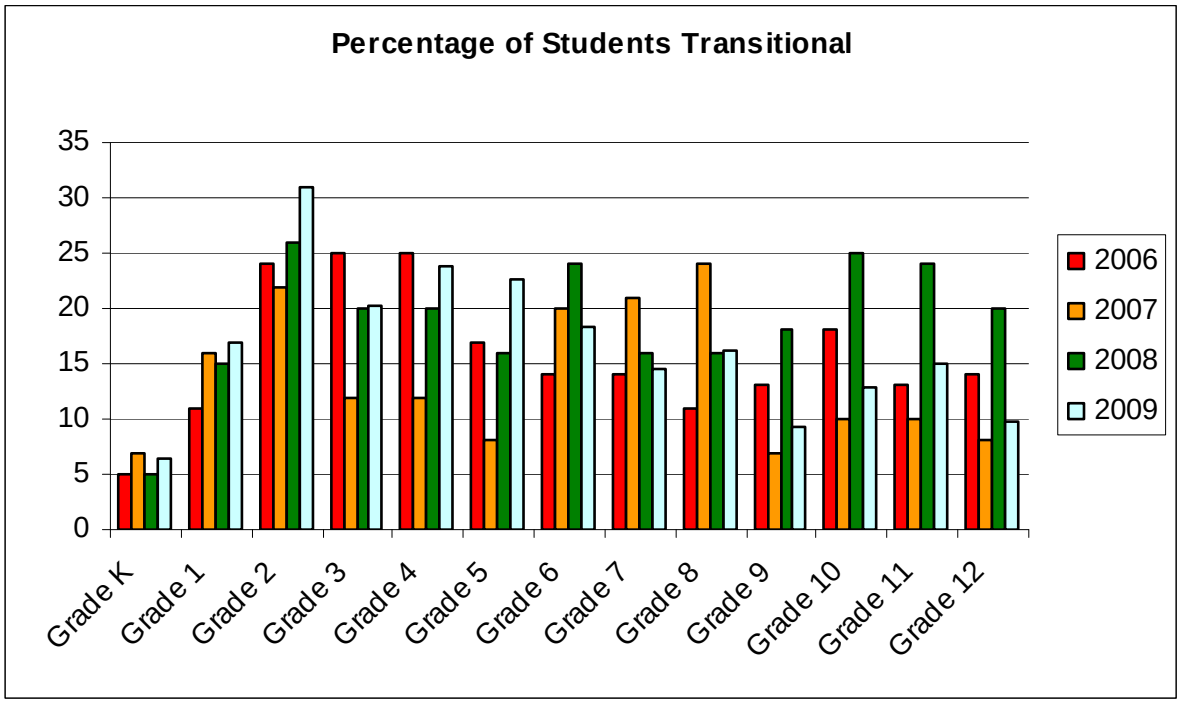
Note. The percentages within a grade may not sum to 100 due to rounding error.

Table 20 presents the percentage of students who were classified in the Transitional performance level for 2006-2009, as well as, the change in the percentage of students who were classified in the Transitional performance category between 2009 and 2008. The percentage of students who were classified in each proficiency level for 2006-2009 is presented graphically in Figure 6. Appendix C contains additional statistical summaries. A stacked bargraph that displays the percentage of students in each proficiency level from 2006-2009 across grades within a level is presented first. The second set of results shows frequency distributions of scaled scores for 2006-2009 for each grade.

Table 20: Percentage of Students in Transitional by Grade

Grade	2006	2007	2008	2009	Change from 2008 to 2009
Grade K	5	7	5	6	1
Grade 1	11	16	15	17	2
Grade 2	24	22	26	31	5
Grade 3	25	12	20	20	0
Grade 4	25	12	20	24	4
Grade 5	17	8	16	23	7
Grade 6	14	20	24	18	-6
Grade 7	14	21	16	15	-1
Grade 8	11	24	16	16	0
Grade 9	13	7	18	9	-9
Grade 10	18	10	25	13	-12
Grade 11	13	10	24	15	-9
Grade 12	14	8	20	10	-10

Figure 6: Percentage of Students Transitional by Grade from 2006-2009



8.2. May Administration of the WLPT-II

The May (Wave 2) test window is intended to be a makeup window for students who were unable to test or complete the test during the annual administration window. These students were tested on the Form C WLPT. For a full discussion and analysis of the Form C WLPT, including the validity, reliability, equating, etc., please refer to the 2008 technical and the equating study reports.

The scale score summaries (as in Table 16) of all students who qualified for ELD services after the last day of the February/March testing window and were tested during the current year's May administration are presented in Appendix E1, while Appendix E2 includes performance classification of these students by grades (as in Table 19).

9. ACCURACY AND CONSISTENCY OF CLASSIFICATIONS

Student performance on the WLPT-II is classified into one of four proficiency levels (Beginner/Advanced Beginner, Intermediate, Advanced, and Transitional). While it is always important to know the reliability of student scores in any examination, it is of even greater importance to assess the reliability of the decisions based on these scores. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of student performance. Methodology from Livingston and Lewis (1995) were applied to derive measures of the accuracy and consistency of the classifications. This methodology allows for any combination of item format within the test. A brief description of the procedure used and results obtained are presented in this section.

9.1. Accuracy of Classification

According to Livingston and Lewis (1995, p. 180), the accuracy of a classification is "…the extent to which the actual classifications of the test takers…agree with those that would be made on the basis of their true score, if their true scores could somehow be known." Accuracy estimates are calculated from cross-tabulations between "classifications based on an observable variable (scores on … a test) and classifications based on an unobservable variable (the test takers' true scores)." True score is equivalent to a hypothetical mean of scores from all possible forms of the test if they were obtainable (Young and Yoon, 1998). Since these true scores are not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed score classifications. An example of a 4×4 cross-tabulation of the true score vs. observed score classifications is given in Figure 7.

Figure 7: An Example of Classification Accuracy Table: Proportions of Students Classified into Proficiency Levels by True Scores vs. Observed Scores

True Score Status	Observed Score Status				
	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional	Total
Beginner/ Advanced Beginner	0.08	0.02	0.00	0.00	0.10
Intermediate	0.03	0.33	0.05	0.00	0.41
Advanced	0.00	0.06	0.38	0.04	0.48
Transitional	0.00	0.00	0.00	0.01	0.01
Total	0.11	0.41	0.43	0.05	1.00

This table shows the proportions of students who were classified into each proficiency category by actual observed scores and by estimated true scores. Diagonal cells represent proportions of students who were correctly classified, whereas off diagonal cells represent proportions of inaccurate classifications. Marginal entries represent total proportions of students classified into each proficiency level by either observed score or estimated true score alone.

For example, the table shows that 48% of students were categorized as *Advanced* by estimated true score status alone, 43% of students were declared as *Advanced* by observed score status alone, and 38% of students were classified as *Advanced* by both true score and observed score status. Also, 6% of students were classified as *Intermediate* by observed score but were *Advanced* by true score (*false negatives*), and 4% of students were classified as *Transitional* by observed score but were *Advanced* by true score (*false positives*).

9.2. Consistency of Classification

Classification consistency is "the agreement between classifications based on two non-overlapping, equally difficult forms of the test" (Livingston & Lewis, 1995). It is estimated using actual response data from a test and the test's reliability. Based on this input information, two parallel forms of the test are statistically modeled and the classifications based on these parallel forms are compared. The example of a 4×4 cross-tabulation between the classifications based on an actual form taken and the classifications based on a hypothetical alternate form is given in

Figure 8. It shows the proportions of student performance classified into each proficiency category by the actual test taken and by the hypothetical alternate test form.

Figure 8: An Example of Classification Consistency Table: Proportions of Students Classified in Proficiency Levels by Test Form Taken vs. Hypothetical Alternate Form

	Status on *Hypothetical Alternate Form*				
Status on *Form Taken*	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional	Total
Beginner/ Advanced Beginner	0.08	0.03	0.00	0.00	0.11
Intermediate	0.03	0.30	0.08	0.00	0.41
Advanced	0.00	0.08	0.32	0.03	0.43
Transitional	0.00	0.00	0.03	0.02	0.05
Total	0.11	0.41	0.43	0.05	1.00

For example, it can be seen that 41% of students are classified into *Intermediate* by the actual test form taken. However, it is estimated that only 30% of students would be consistently classified into the *Intermediate* category if they were to be assessed again by the alternate form of the test.

Note that the proportion of mis-classification in the classification consistency table, in its original form, is symmetric, whereas the proportion of mis-classification in the classification accuracy table is non-symmetrical because it compares classifications based on two different types of scores. Also note that agreement rates are lower in the classification consistency table because both classifications based on both tests contain measurement error, whereas in the accuracy table, true score classification is assumed to be errorless.

9.3. Accuracy and Consistency Indices

Three types of accuracy and consistency indices will be presented: overall, conditional on proficiency level, and by cut point. In order to facilitate the interpretation, a brief outline of computational procedures used to derive accuracy indices are presented using the examples shown in Figure 7 and

Figure 8.

The overall accuracy of proficiency level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels, as indicated by the shaded area in Figure 9 below. It represents a proportion (or percentage) of correct classifications across all the levels. Based on the example shown in Figure 7, the sum of the diagonal cells equals 0.80. This means that 80% of students have their test performance classified in the same proficiency categories based on their observed scores as they would have it classified based on their true scores, if they were known.

Additionally, the overall false positive and false negative rates can be examined. The overall false positive rate equals the sum of the upper right cells above the diagonal in the accuracy table. Based on the example of Figure 7, the overall false positive rate equals .11, which indicates that 11% of students have their test performance classified on a higher proficiency level based on their observed scores as they would have it classified based on their true scores, if they were known. The overall false negative rate equals the sum of the lower left cells below the diagonal in the accuracy table. Based on the example of Figure 7, the overall false negative rate equals .09, which indicates that 9% of students have their test performance classified on a lower proficiency level based on their observed scores as they would have it classified based on their true scores, if they were known.

Likewise, the Transitional false positive and false negative rates can be examined. The Transitional false positive rate is the proportion of students whose classifications based on true scores were levels less than Transitional, but whose classifications based on observed scores were Transitional. The Transitional false negative rate is the proportion of students whose classifications based on true scores were Transitional, but whose classifications based on observed scores were levels less than Transitional.

Figure 9: Overall Classification Accuracy or Consistency as the Sum of the Diagonal Cells (A + B+ C + D)

	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional	Total
Beginner/ Advanced Beginner	A				
Intermediate		B			
Advanced			C		
Transitional				D	
Total					

The overall classification consistency index is computed analogously as the sum of the diagonal cells in the consistency table. Using the data from

Figure 8, it can be determined that the sum of the diagonal cells in the classification consistency table equals 0.72. In other words, 72% of students would be classified in the same proficiency levels based on the alternate form, if they had taken it.

Another way to express overall classification consistency is to use Cohen's *kappa* (κ) coefficient (Cohen, 1960). *Kappa* is a measure of "how much agreement exists beyond chance alone" (Fleiss, 1973, p. 146). In the case of consistency, κ is the proportion of consistent classifications between two forms after removing the proportion of consistent classifications that would be expected by chance alone. Based on the data from

Figure 8, κ equals 0.54. Compared to the previously described overall consistency index, κ has a lower value because it has been corrected for chance.

Classification consistency, conditional on proficiency level, is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all student performance classified into that level (marginal entry, see Figure 10). As an example, the consistency at level *Intermediate* is computed from the data in

Figure 8. The ratio between 0.30 (proportion of the correct classifications at that level) and 0.41 (total proportion of student performance classified into that level) yields 0.73, representing the index of consistency of classification at the level *Intermediate*. It indicates that 73% of all students classified as *Intermediate* would be classified in the same level based on the hypothetical alternate form, if they had taken it.

Figure 10: Accuracy or Consistency Conditional on Level— Intermediate Equals the Ratio of A Over B

	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional	Total
Beginner/ Advanced Beginner					
Intermediate		A			B
Advanced					
Transitional					
Total					

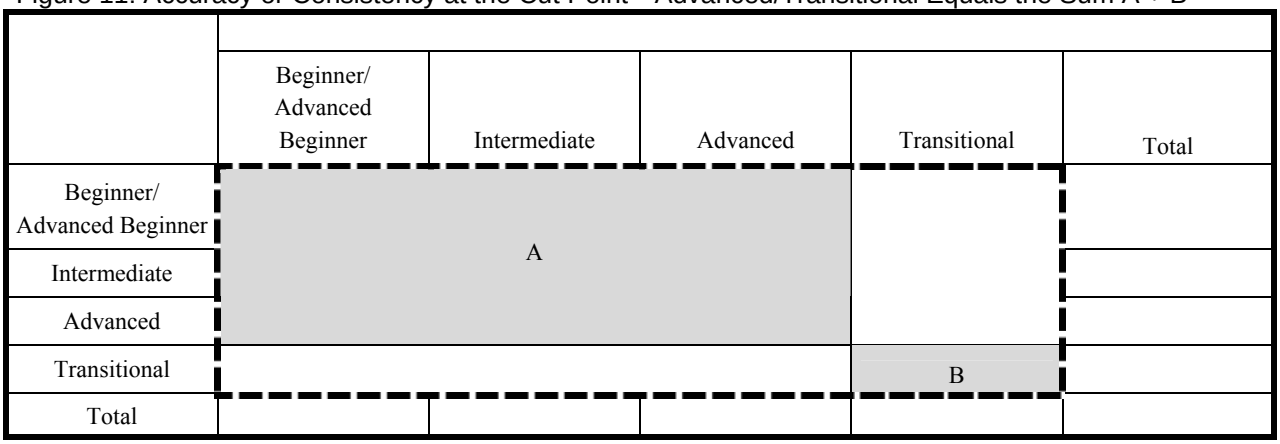
Classification accuracy, conditional on proficiency level, is analogously computed from the accuracy table. The only difference is that the marginal sum based on true status is used as a total for computing accuracy conditional on level. For example, in Figure 7, the proportion of agreement between true score status and observed score status at the *Intermediate* level is 0.33 and the total proportion of student performance with true score status at this level is 0.41. The accuracy conditional on level is equal to the ratio between those two proportions, which yields 0.80. It indicates that 80% of the students who were estimated to have a true score status of *Intermediate* have their performance correctly classified into that category by their observed scores.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points the joint distribution of all the proficiency levels is collapsed into a

dichotomized distribution around that specific cut point. For the purposes of WLPT-II, the dichotomization at the cut point between the *Advanced* and *Transitional* levels is key, since students categorized as *Transitional* are transitioned into English-speaking classrooms.

This dichotomization is depicted in Figure 11. The proportion of correct classifications below this particular cut point is equal to the sum of all the cells at the levels *Beginner/Advanced Beginner*, *Intermediate*, and *Advanced* (upper left shaded area), and the proportion of correct classifications above that particular cut point is equal to the sum of all the cells at the level *Transitional* (lower right shaded area).

Figure 11: Accuracy or Consistency at the Cut Point—Advanced/Transitional Equals the Sum A + B



The classification accuracy index, by cut point, is computed as the sum of the proportions of correct classifications around a selected cut point. Based on the data in Figure 7, the computation of the accuracy index at the cut point between the *Advanced* and *Transitional* levels equals 0.96. This means that 96% of student performance was correctly classified either above or below the particular cut point. The sum of the proportions in the upper right non-shaded area indicates false positives (i.e., 4% of students were classified above the cut point by their observed scores, but fell below the cut point by their true scores). The lower left non-shaded area contains the proportion of false negatives (i.e., 0% of students with observed levels below the cut point whose true levels were above the cut point).

The classification consistency by cut point is obtained in an analogous way. For example, if we take data from

Figure 8 and we dichotomize the distribution at the cut point between the *Advanced* level and the *Transitional* level, the proportion of correct classifications around that cut point equals 0.94. This means that 94% of students would have their test performance classified into either below or above the *Advanced/Transitional* cut consistently by both the actual form taken and by the alternate form (if they had taken it).

9.4. Adjusting the Marginal Proportions

In the classification accuracy table, there is no built-in constraint for the marginal proportions on the observed score status (column marginals) to equal the actual observed marginal proportions of each proficiency level. Similarly in the classification consistency table, there is no built-in constraint for the marginal proportions on the form taken status or the hypothesized alternative form status to equal the observed marginal proportions of each proficiency level. This is because the marginals are based on what is expected under the observed score model. Livingston and

Lewis (1995) proposed adjusting the accuracy and consistency tables so that the column marginals on the accuracy table and both the row and column marginals on the consistency table equal that of the observed marginal proficiency level proportions. In the results presented below, this adjustment was made so that the appropriate marginal proportions equal the observed marginals.

9.5. Summary of Livingston and Lewis (1995) Procedure

Step 1: Estimate effective test length (i.e., the estimated number of hypothetical dichotomous, statistically independent items needed to produce total scores at the observed reliability), using the following:

$$n_{eff} = \frac{\left(\overline{X} - X_{\min}\right)\left(X_{\max} - \overline{X}\right) - r_{XX'} S_X^2}{S_X^2 \left(1 - r_{XX'}\right)},$$

where $\overline{X}$ is the sample mean test score,

$X_{\min}$ is the minimum observed test score,

$X_{\max}$ is the maximum observed test score,

$r_{XX'}$ is the estimated test reliability, and

S_X^2 is the sample test score variance.

In the results presented below, total test (composite) scaled scores were used as the test score. Cronbach's alpha estimate of internal consistency reliability was used as the estimate of test reliability. (Table 4.1 presented the 2008 Form A values for WLPT-II of Cronbach's alpha by grade.) Since Cronbach alpha at each grade was very high (ranging from 0.92 to 0.95), it was unlikely that these were underestimates of reliability. As such, more complex reliability coefficients (e.g., Qualls, 1995) were not needed.

Step 2: Estimate the proportional true score distribution using the four-parameter beta density. Proportional true scores are operationally defined as

$$T_p = \frac{E(X) - X_{\min}}{X_{\max} - X_{\min}},$$

where $E(X)$ is the expected value of an observed score.

The four-parameter beta density for the proportion true score is given by

$$P(T_p | a, b, d, \Delta) = \frac{1}{B(d+1, \Delta+1)} \frac{(T_p - a)^d (b - T_p)^\Delta}{(b-a)^{d+\Delta+1}},$$

where $B(\cdot,\cdot)$ is the two-parameter beta density

d and Δ are the two-parameter beta density parameters, and

a and *b* are transformational parameters to place the two-parameter beta density onto a $(0,1)$ metric.

Step 3: Estimate the conditional classification distribution for an alternative form of the test at each level of the proportional true score; i.e., estimate $P(X < x_j^* | T_p)$, where x_j^* is the *j*-th cut score or cut point. For the results to be presented, scaled cut scores were used.

Step 4: Estimate the joint classification distribution of true scores and scores on an alternate form. This is then used to form a two-way classification table.

Step 5: Estimate the joint classification distribution of true scores and scores on the form that was taken by adjusting the two-way table from Step 4 using multipliers formed via the observed proficiency level frequencies. This adjusted table is then used for examining decision accuracy.

Step 6: Estimate the joint classification distribution of scores on two alternate forms. Then form a two-way classification table using this joint distribution.

Step 7: Adjust the two-way table formed in Step 6 using multipliers formed via the observed proficiency level frequencies. This adjusted table is then used for examining decision consistency.

9.6. Accuracy and Consistency Results

Table 21 presents the overall classification accuracy results by grade. The overall classification accuracies ranged from 0.84 to .95. The overall false positive rates ranged from 0.02 to 0.12, while the false negative rates ranged from 0.02 to 0.08. The Transitional false positive rates ranged from 0.01 to 0.10, while the transitional false negative rates ranged from 0.00 to a maximum of 0.05. The accuracy results for Transitional included five grades with values = 0.00, indicated in the table by '*'. This, however, is not an actual indication of accuracy but an artifact of the method as explained in the footnote.

Table 21: Overall Accuracy Results by Grade

	Diagonals					Overall		Transitional	
Grade	B/AB	I	A	T*	Overall	False Positive	False Negative	False Positive	False Negative
K	0.07	0.38	0.39	*	0.84	0.09	0.08	*	*
1	0.01	0.28	0.41	0.16	0.85	0.07	0.08	0.03	0.02
2	0.01	0.15	0.43	0.27	0.86	0.07	0.07	0.05	0.03
3	0.01	0.12	0.55	0.17	0.85	0.07	0.08	0.06	0.05
4	0.01	0.12	0.54	0.19	0.86	0.08	0.07	0.05	0.04
5	0.01	0.13	0.54	0.17	0.85	0.08	0.07	0.05	0.05
6	0.01	0.10	0.62	0.11	0.84	0.09	0.07	0.07	0.05
7	0.02	0.14	0.63	0.09	0.87	0.07	0.06	0.05	0.03
8	0.04	0.14	0.68	*	0.85	0.12	0.03	*	*
9	0.03	0.16	0.73	*	0.92	0.05	0.03	*	*
10	0.02	0.17	0.63	0.05	0.87	0.07	0.06	0.05	0.03
11	0.03	0.11	0.76	*	0.89	0.08	0.02	*	*
12	0.02	0.10	0.84	*	0.95	0.02	0.02	*	*

Note. 1. B/AB is Beginner/Advanced Beginner, I is Intermediate, A is Advanced, and T is Transitional.

2. Overall is the sum across these four proficiency levels.

* The proportional true score associated with score X is expressed on a scale of 0 to 1. The four-parameter beta density for the proportional true scores is a function of a location parameter, a scale parameter, and two parameters for the upper and lower bounds on X . There are times, however, when the upper bound parameter is less than 1. Under these circumstances, it is quite likely that the proportional *true* score cut may never reach the *observed* proportional score cut. Because of this, the correct accuracy classification at the highest level cut (Transitional) may not be achieved, and the proportion of students at this level will have a "0" for correct classification, and have no False Negatives. In addition, the observed proportions at this level are then classified as False Positives. Thus, both of these outcomes are artifacts of the procedure used to calculate accuracy classification.

Table 22 presents the overall classification consistency results. Overall classification consistency ranged from 0.76 to 0.93 across grades.

Table 22: Overall Consistency Results by Grade

	Diagonals					
Grade	B/AB	I	A	T	Overall	Kappa
K	0.07	0.36	0.34	0.00	0.76	0.61
1	0.01	0.26	0.37	0.15	0.79	0.67
2	0.01	0.14	0.39	0.26	0.80	0.68
3	0.01	0.12	0.52	0.15	0.79	0.61
4	0.01	0.12	0.50	0.18	0.80	0.64
5	0.01	0.12	0.50	0.16	0.79	0.63
6	0.01	0.09	0.58	0.10	0.78	0.55
7	0.02	0.13	0.59	0.08	0.81	0.62
8	0.04	0.13	0.61	0.03	0.80	0.58
9	0.03	0.15	0.70	0.00	0.88	0.69
10	0.02	0.16	0.59	0.05	0.82	0.64
11	0.02	0.10	0.70	0.01	0.84	0.57
12	0.02	0.09	0.82	0.00	0.93	0.73

Note. 1. B/AB is Beginner/Advanced Beginner, I is Intermediate, A is Advanced, and T is Transitional.

2. Overall is the sum across these four proficiency levels.

Table 23 presents the conditional accuracy and classification consistency results. The accuracy results for the Intermediate and Advanced proficiency levels were largely in the .80s and .90s,

while the Beginner/Advanced Beginner level results ranged from .56 to .86. On the other hand, the consistency results for the Intermediate and Advanced proficiency levels were largely in the .70s and .80s, while the Beginner/Advanced Beginner level results ranged from .57 to .83. Conditional accuracy results for Transitional included five grades with values = 0.00, indicated by '*' in the table. This, similar to the results in Table 21, however, is not an actual indication of accuracy but an artifact of the method as explained in the footnote.

Table 23: Conditional Accuracy and Consistency Results by Grade

	Accuracy				Consistency			
Grade	B/AB	I	A	T*	B/AB	I	A	T
K	0.85	0.84	0.89	*	0.82	0.79	0.77	0.12
1	0.56	0.86	0.87	0.82	0.57	0.81	0.79	0.78
2	0.63	0.82	0.88	0.85	0.62	0.76	0.80	0.82
3	0.75	0.82	0.89	0.75	0.73	0.78	0.83	0.67
4	0.74	0.83	0.90	0.77	0.71	0.78	0.84	0.72
5	0.79	0.83	0.89	0.76	0.76	0.78	0.83	0.69
6	0.79	0.82	0.91	0.61	0.75	0.77	0.85	0.55
7	0.79	0.83	0.92	0.64	0.75	0.78	0.87	0.58
8	0.86	0.84	0.97	*	0.83	0.78	0.87	0.32
9	0.86	0.85	0.98	*	0.82	0.80	0.93	0.12
10	0.82	0.85	0.93	0.52	0.79	0.80	0.88	0.49
11	0.85	0.82	0.98	*	0.82	0.78	0.90	0.19
12	0.85	0.83	0.98	*	0.82	0.77	0.96	0.02

Note. 1. B/AB is Beginner/Advanced Beginner, I is Intermediate, A is Advanced, and T is Transitional.
2. Overall is the sum across these four proficiency levels.

* The proportional true score associated with score X is expressed on a scale of 0 to 1. The four-parameter beta density for the proportional true scores is a function of a location parameter, a scale parameter, and two parameters for the upper and lower bounds on X . There are times, however, when the upper bound parameter is less than 1. Under these circumstances, it is quite likely that the proportional *true* score cut may never reach the *observed* proportional score cut. Because of this, the correct accuracy classification at the highest level cut (Transitional) may not be achieved, and the proportion of students at this level will have a "0" for correct classification, and have no False Negatives. In addition, the observed proportions at this level are then classified as False Positives. Thus, both of these outcomes are artifacts of the procedure used to calculate accuracy classification. Since conditional accuracy is a ratio in which the numerator is the proportion of correct accuracy classification, the conditional accuracy at this level will also be zero.

Table 24 presents the cut point classification accuracy and classification consistency results. Accuracy ranged from 0.88 to 0.99 and consistency ranged from 0.84 to 0.98.

Table 24: Cut Point Accuracy and Consistency by Grade

Grade	Accuracy	Consistency
K	0.97	0.95
1	0.94	0.92
2	0.92	0.89
3	0.90	0.85
4	0.90	0.86
5	0.90	0.86
6	0.88	0.84
7	0.92	0.89
8	0.90	0.87
9	0.97	0.95
10	0.93	0.90
11	0.93	0.89
12	0.99	0.98

10. REFERENCES

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. CA: SAGE Publications.

Canale, M. (1985). *A Theory of Strategy-Oriented* Language *Development.* ED273147.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37- 46.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika*, 16, 297 - 334.

Cummins, J. (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism,* No. 19, 121-129.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed-response and differential item functioning: a pragmatic approach. *ETS Research Report* No. 91-49. Princeton, NJ: Educational Testing Service.

Edelsky, C. (1990). *With literacy and justice for all: Rethinking the social in language and education.* London: The Falmer Press.

Edelsky, C, Hudelson, S., Altwerger, B., Flores, B., Barkin, F., & Jilbert, K. (1983). Semilingualism and language deficit. *Applied Linguistics, 4*(1), 1-22.

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). NY: McGraw-Hill.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items applied. *Psychological Measurement, 9,* 139-164.

Kolen, M. J., & Brennan R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). NY: Springer-Verlag.

Linacre, J. M. (2006). WINSTEPS (Version 3.63) [Computer software]. Chicago, IL: Winsteps.

Linacre, J. M. (2005). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: Winsteps.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

Martin-Jones, M., & Romaine, S. (1986) Semilingualism: A half-baked theory of communicative competence. *Applied Linguistics,* 7, 26-38.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Morgan, D. L., & Perie, M. (2004). *Setting standards in education: choosing the best method for your assessment and population*. Unpublished Paper. NJ: Educational Testing Service.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). NJ: Person Education Inc.

Qualls, A. L. (1995). Estimating the Reliability of a Test Containing Multiple Item Formats. *Applied Measurement in Education, 8*, 111-120.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* IL: University of Chicago Press.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). NJ: Lawrence Erlbaum Associates, Inc.

Tenenbaum, I., Lindsay, S., Siskind, T., Wall-Mitchell, M. E., & Saunders, J. (2001). *Technical documentation for the 2000 palmetto achievement challenge tests of English language arts and mathematics.* SC: South Carolina Department of Education.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51,* 567-577.

Wiley, T. G. (1996). *Literacy and language diversity in the United States.* Washington, DC: Center for Applied Linguistics and Delta Systems.

Young, M. J., & Yoon, B. (1998). Estimating the consistency and accuracy of classification in a standards-referenced assessment. CSE Technical Report 475. UCLA Center for the Study of Evaluation: Los Angeles, CA.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

APPENDIX A: WLPT-II (FORM A) RAW SCORE TO SCALE SCORE CONVERSION TABLES

Table A1: Form A Total Raw Score to Scale Score Conversion Table for Primary (Grades K-2)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-8.5045	2.0193	300	73
1	-7.0606	1.0371	348	38
2	-6.2935	0.7564	376	27
3	-5.8179	0.6336	393	23
4	-5.4644	0.56	406	20
5	-5.18	0.5089	417	18
6	-4.9409	0.4703	425	17
7	-4.7343	0.4396	433	16
8	-4.5523	0.4142	439	15
9	-4.3897	0.3929	445	14
10	-4.2426	0.3747	450	14
11	-4.108	0.3591	455	13
12	-3.984	0.3456	460	13
13	-3.8687	0.3338	464	12
14	-3.7607	0.3236	468	12
15	-3.659	0.3146	472	11
16	-3.5625	0.3067	475	11
17	-3.4707	0.2997	478	11
18	-3.3828	0.2934	482	11
19	-3.2983	0.2878	485	10
20	-3.217	0.2827	488	10
21	-3.1384	0.2781	490	10
22	-3.0622	0.2739	493	10
23	-2.9882	0.27	496	10
24	-2.9163	0.2663	498	10
25	-2.8464	0.2628	501	10
26	-2.7782	0.2596	503	9
27	-2.7115	0.2565	506	9
28	-2.6465	0.2535	509	9
29	-2.583	0.2507	510	9
30	-2.5208	0.248	513	9
31	-2.46	0.2454	515	9
32	-2.4003	0.2429	517	9
33	-2.3419	0.2405	519	9
34	-2.2847	0.2383	521	9
35	-2.2283	0.2361	523	9
36	-2.1731	0.234	525	8
37	-2.1188	0.2321	527	8
38	-2.0653	0.2302	529	8
39	-2.0128	0.2285	531	8
40	-1.9609	0.2268	533	8
41	-1.9098	0.2252	535	8
42	-1.8595	0.2238	537	8
43	-1.8097	0.2224	538	8
44	-1.7606	0.2211	540	8
45	-1.7119	0.22	542	8
46	-1.6637	0.2189	544	8
47	-1.616	0.2179	545	8
48	-1.5688	0.217	547	8
49	-1.5219	0.2162	549	8
50	-1.4753	0.2155	551	8
51	-1.429	0.2148	552	8
52	-1.383	0.2143	554	8
53	-1.3372	0.2138	556	8
54	-1.2915	0.2134	557	8
55	-1.2461	0.2131	559	8
56	-1.2008	0.2128	560	8
57	-1.1555	0.2127	562	8

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
58	-1.1103	0.2126	564	8
59	-1.0651	0.2126	566	8
60	-1.0198	0.2127	567	8
61	-0.9746	0.2128	569	8
62	-0.9293	0.213	570	8
63	-0.8839	0.2133	572	8
64	-0.8383	0.2136	574	8
65	-0.7926	0.214	575	8
66	-0.7467	0.2145	577	8
67	-0.7005	0.2151	579	8
68	-0.6541	0.2157	580	8
69	-0.6075	0.2164	582	8
70	-0.5605	0.2172	584	8
71	-0.5132	0.218	586	8
72	-0.4654	0.2189	587	8
73	-0.4173	0.2199	589	8
74	-0.3687	0.221	591	8
75	-0.3196	0.2222	592	8
76	-0.2699	0.2234	594	8
77	-0.2197	0.2248	596	8
78	-0.1688	0.2262	598	8
79	-0.1173	0.2277	600	8
80	-0.065	0.2294	602	8
81	-0.0121	0.2312	603	8
82	0.0419	0.233	605	8
83	0.0966	0.235	607	9
84	0.1523	0.2372	609	9
85	0.2091	0.2395	611	9
86	0.267	0.2419	614	9
87	0.3262	0.2445	616	9
88	0.3867	0.2473	618	9
89	0.4485	0.2503	620	9
90	0.5119	0.2534	622	9
91	0.577	0.2568	625	9
92	0.6438	0.2604	627	9
93	0.7127	0.2643	630	10
94	0.7837	0.2685	632	10
95	0.857	0.273	635	10
96	0.9328	0.2779	638	10
97	1.0116	0.2832	641	10
98	1.0933	0.2889	643	10
99	1.1787	0.2952	647	11
100	1.2678	0.3021	650	11
101	1.3614	0.3098	653	11
102	1.4601	0.3184	657	12
103	1.5644	0.3281	661	12
104	1.6756	0.3391	665	12
105	1.795	0.3519	669	13
106	1.924	0.3669	674	13
107	2.0651	0.3849	679	14
108	2.2216	0.4069	684	15
109	2.3981	0.4345	691	16
110	2.6023	0.4707	698	17
111	2.8466	0.5204	707	19
112	3.1548	0.5945	718	22
113	3.5799	0.7206	733	26
114	4.2919	1.0091	759	37
115	5.6917	2.0044	810	73

Table A2: Form A Listening Raw Score to Scale Score Conversion Table for Primary (Grades K-2)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-8.2675	2.0316	305	74
1	-6.784	1.0639	358	38
2	-5.9557	0.7989	388	29
3	-5.4088	0.6914	408	25
4	-4.9734	0.6333	424	23
5	-4.596	0.5978	438	22
6	-4.253	0.575	450	21
7	-3.9315	0.5602	462	20
8	-3.6229	0.5515	473	20
9	-3.3212	0.548	484	20
10	-3.0206	0.5494	495	20
11	-2.7155	0.5561	506	20
12	-2.4	0.5685	517	21
13	-2.0664	0.5877	529	21
14	-1.7058	0.615	542	22
15	-1.3051	0.6529	557	24
16	-0.8451	0.7063	573	26
17	-0.2919	0.7865	593	28
18	0.4276	0.9214	619	33
19	1.516	1.2012	659	43
20	3.2502	2.1146	722	77

Table A3: Form A Speaking Raw Score to Scale Score Conversion Table for Primary (Grades K-2)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-6.343	1.9641	374	71
1	-5.0403	0.9556	422	35
2	-4.4124	0.6727	444	24
3	-4.0415	0.5569	458	20
4	-3.7681	0.4938	468	18
5	-3.5443	0.4547	476	16
6	-3.3501	0.4284	483	15
7	-3.1749	0.4095	489	15
8	-3.0133	0.3952	495	14
9	-2.8618	0.3835	500	14
10	-2.7186	0.3737	506	14
11	-2.5823	0.365	511	13
12	-2.4519	0.3572	515	13
13	-2.3268	0.3503	520	13
14	-2.2063	0.3443	524	12
15	-2.0895	0.3392	528	12
16	-1.9759	0.335	532	12
17	-1.8648	0.3319	536	12
18	-1.7554	0.3297	540	12
19	-1.6471	0.3286	544	12
20	-1.5393	0.3285	548	12
21	-1.431	0.3296	552	12
22	-1.3217	0.3317	556	12
23	-1.2106	0.3352	560	12
24	-1.0968	0.3399	564	12
25	-0.9793	0.346	569	13
26	-0.8569	0.3538	573	13
27	-0.7285	0.3635	578	13
28	-0.5921	0.3754	583	14
29	-0.4458	0.39	588	14
30	-0.2868	0.4081	594	15
31	-0.1112	0.4307	600	16
32	0.0864	0.4594	607	17
33	0.3143	0.4969	615	18
34	0.586	0.5481	625	20
35	0.9262	0.6232	637	23
36	1.3895	0.749	654	27
37	2.147	1.0333	682	37
38	3.5864	2.0183	734	73

Table A4: Form A Reading Raw Score to Scale Score Conversion Table for Primary (Grades K-2)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.2187	2.0285	415	73
1	-3.748	1.0538	468	38
2	-2.9469	0.778	497	28
3	-2.4385	0.659	516	24
4	-2.0517	0.5896	530	21
5	-1.7321	0.5435	541	20
6	-1.455	0.5109	551	18
7	-1.2066	0.4869	560	18
8	-0.9786	0.4691	569	17
9	-0.7649	0.456	576	16
10	-0.5614	0.4468	584	16
11	-0.3646	0.4408	591	16
12	-0.1718	0.4379	598	16
13	0.0197	0.4378	605	16
14	0.2124	0.4407	612	16
15	0.4091	0.4468	619	16
16	0.6128	0.4567	626	17
17	0.8278	0.4712	634	17
18	1.0592	0.4919	642	18
19	1.3149	0.5214	652	19
20	1.6083	0.5645	662	20
21	1.9633	0.6318	675	23
22	2.4334	0.7506	692	27
23	3.1886	1.0301	719	37
24	4.6201	2.0148	771	73

Table A5: Form A Writing Raw Score to Scale Score Conversion Table for Primary (Grades K-2)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.6049	2.022	401	73
1	-4.1546	1.0404	454	38
2	-3.3826	0.7586	482	27
3	-2.905	0.6344	499	23
4	-2.5509	0.5602	512	20
5	-2.2662	0.5094	522	18
6	-2.026	0.4722	531	17
7	-1.8166	0.444	538	16
8	-1.6293	0.4224	545	15
9	-1.4581	0.4057	551	15
10	-1.2989	0.393	557	14
11	-1.1483	0.3834	562	14
12	-1.0041	0.3765	568	14
13	-0.8642	0.3718	573	13
14	-0.7272	0.3689	578	13
15	-0.5917	0.3675	583	13
16	-0.4566	0.3677	587	13
17	-0.3211	0.3692	592	13
18	-0.1837	0.3722	597	13
19	-0.0435	0.3767	602	14
20	0.1007	0.3829	608	14
21	0.2501	0.3907	613	14
22	0.4065	0.4004	619	14
23	0.5714	0.4121	625	15
24	0.7469	0.426	631	15
25	0.9352	0.4423	638	16
26	1.1392	0.4616	645	17
27	1.3629	0.4852	653	18
28	1.6128	0.5158	662	19
29	1.8999	0.5583	673	20
30	2.2466	0.6238	685	23
31	2.7041	0.7403	702	27
32	3.4406	1.019	728	37
33	4.8524	2.0072	779	73

Table A6: Form A Total Raw Score to Scale Score Conversion Table for Elementary (Grades 3-5)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-6.4551	2.0039	370	72
1	-5.057	1.0077	421	36
2	-4.3487	0.7175	447	26
3	-3.9286	0.5897	462	21
4	-3.6267	0.5138	473	19
5	-3.3897	0.4625	481	17
6	-3.1935	0.4249	488	15
7	-3.0253	0.396	494	14
8	-2.8778	0.373	500	13
9	-2.7457	0.3543	505	13
10	-2.6258	0.3387	509	12
11	-2.5156	0.3255	513	12
12	-2.4134	0.3142	517	11
13	-2.3178	0.3045	520	11
14	-2.2277	0.296	523	11
15	-2.1423	0.2885	526	10
16	-2.0609	0.2819	529	10
17	-1.9832	0.276	532	10
18	-1.9084	0.2707	535	10
19	-1.8364	0.2659	537	10
20	-1.7669	0.2616	540	9
21	-1.6995	0.2576	542	9
22	-1.6341	0.2539	545	9
23	-1.5705	0.2505	547	9
24	-1.5085	0.2474	549	9
25	-1.4481	0.2445	552	9
26	-1.3889	0.2419	554	9
27	-1.331	0.2394	556	9
28	-1.2742	0.237	558	9
29	-1.2186	0.2348	560	8
30	-1.164	0.2328	562	8
31	-1.1101	0.2309	564	8
32	-1.0572	0.2292	566	8
33	-1.0051	0.2276	568	8
34	-0.9536	0.2261	569	8
35	-0.9029	0.2247	572	8
36	-0.8527	0.2234	573	8
37	-0.803	0.2222	575	8
38	-0.7539	0.2212	577	8
39	-0.7052	0.2202	578	8
40	-0.6569	0.2193	580	8
41	-0.609	0.2186	582	8
42	-0.5614	0.2179	584	8
43	-0.514	0.2173	585	8
44	-0.4669	0.2168	587	8
45	-0.42	0.2163	589	8
46	-0.3733	0.2159	590	8
47	-0.3268	0.2157	592	8
48	-0.2802	0.2154	594	8
49	-0.2339	0.2153	595	8
50	-0.1876	0.2152	597	8
51	-0.1413	0.2151	599	8
52	-0.0951	0.2151	600	8
53	-0.0488	0.2152	602	8
54	-0.0024	0.2153	604	8
55	0.0439	0.2155	606	8
56	0.0904	0.2157	607	8
57	0.137	0.216	609	8
58	0.1838	0.2163	611	8

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
59	0.2306	0.2167	612	8
60	0.2777	0.2171	614	8
61	0.325	0.2176	616	8
62	0.3724	0.2182	617	8
63	0.4202	0.2188	619	8
64	0.4682	0.2195	621	8
65	0.5166	0.2202	623	8
66	0.5652	0.221	624	8
67	0.6142	0.2219	626	8
68	0.6637	0.2229	628	8
69	0.7137	0.224	630	8
70	0.764	0.2251	632	8
71	0.815	0.2264	633	8
72	0.8665	0.2278	635	8
73	0.9188	0.2292	637	8
74	0.9717	0.2308	639	8
75	1.0254	0.2326	641	8
76	1.0799	0.2345	644	8
77	1.1353	0.2365	645	9
78	1.1918	0.2386	647	9
79	1.2492	0.241	649	9
80	1.3079	0.2435	651	9
81	1.3679	0.2462	653	9
82	1.4292	0.2491	656	9
83	1.4921	0.2522	658	9
84	1.5565	0.2555	660	9
85	1.6226	0.2591	663	9
86	1.6907	0.2629	665	10
87	1.7609	0.267	668	10
88	1.8334	0.2714	670	10
89	1.9083	0.2761	673	10
90	1.986	0.2812	676	10
91	2.0666	0.2867	679	10
92	2.1505	0.2927	682	11
93	2.238	0.2991	686	11
94	2.3295	0.3061	688	11
95	2.4256	0.3139	692	11
96	2.5268	0.3224	695	12
97	2.6338	0.3319	701	12
98	2.7474	0.3425	703	12
99	2.8687	0.3545	708	13
100	2.9992	0.3682	712	13
101	3.1406	0.3842	718	14
102	3.2953	0.403	723	15
103	3.4669	0.4258	729	15
104	3.6601	0.4542	736	16
105	3.8825	0.4908	744	18
106	4.1473	0.5406	754	20
107	4.478	0.6141	766	22
108	4.9281	0.7387	782	27
109	5.6679	1.0236	809	37
110	7.0904	2.0122	860	73

Table A7: Form A Listening Raw Score to Scale Score Conversion Table for Elementary (Grades 3-5)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.8088	2.0339	430	74
1	-3.3226	1.0629	484	38
2	-2.5037	0.7888	513	29
3	-1.9787	0.6711	532	24
4	-1.5754	0.6036	547	22
5	-1.2386	0.5599	559	20
6	-0.9424	0.5303	570	19
7	-0.6726	0.5098	580	18
8	-0.4201	0.4961	589	18
9	-0.1784	0.4879	597	18
10	0.0575	0.4843	606	18
11	0.2921	0.4851	615	18
12	0.5297	0.4906	623	18
13	0.7753	0.5013	632	18
14	1.0347	0.5186	641	19
15	1.3166	0.5449	652	20
16	1.6343	0.585	663	21
17	2.0121	0.6492	677	23
18	2.5038	0.7645	695	28
19	3.2798	1.0398	723	38
20	4.7261	2.0197	775	73

Table A8: Form A Speaking Raw Score to Scale Score Conversion Table for Elementary (Grades 3-5)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.286	1.9819	413	72
1	-3.9448	0.9749	461	35
2	-3.2905	0.6861	485	25
3	-2.907	0.5638	499	20
4	-2.6294	0.4949	509	18
5	-2.4071	0.4507	517	16
6	-2.2182	0.4201	524	15
7	-2.0513	0.3978	530	14
8	-1.9001	0.3808	535	14
9	-1.7603	0.3676	540	13
10	-1.6291	0.3569	545	13
11	-1.5049	0.3482	549	13
12	-1.3863	0.341	554	12
13	-1.2721	0.335	558	12
14	-1.1617	0.3301	562	12
15	-1.0539	0.3262	566	12
16	-0.9486	0.3233	570	12
17	-0.8448	0.3213	573	12
18	-0.7419	0.3204	577	12
19	-0.6392	0.3204	581	12
20	-0.5363	0.3215	585	12
21	-0.4323	0.3237	588	12
22	-0.3266	0.327	592	12
23	-0.2181	0.3316	596	12
24	-0.1062	0.3376	600	12
25	0.0101	0.345	604	12
26	0.1322	0.3541	609	13
27	0.2614	0.3651	613	13
28	0.3995	0.3783	618	14
29	0.5485	0.3943	624	14
30	0.7114	0.4136	630	15
31	0.8921	0.4373	636	16
32	1.0961	0.467	644	17
33	1.3316	0.5052	652	18
34	1.6123	0.5569	662	20
35	1.9631	0.6322	675	23
36	2.4386	0.7576	692	27
37	3.2103	1.0408	720	38
38	4.6619	2.0227	773	73

Table A9: Form A Reading Raw Score to Scale Score Conversion Table for Elementary (Grades 3-5)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.1146	2.0389	419	74
1	-3.6073	1.0793	473	39
2	-2.7442	0.8213	505	30
3	-2.1604	0.7177	526	26
4	-1.6889	0.6595	543	24
5	-1.2812	0.6192	558	22
6	-0.9178	0.5872	571	21
7	-0.5889	0.5604	583	20
8	-0.2878	0.5379	594	19
9	-0.0085	0.5197	604	19
10	0.2539	0.5055	613	18
11	0.504	0.4952	622	18
12	0.7456	0.4885	631	18
13	0.9823	0.4852	639	18
14	1.2175	0.4852	648	18
15	1.4542	0.4887	657	18
16	1.6964	0.496	665	18
17	1.9479	0.508	674	18
18	2.2147	0.526	684	19
19	2.5045	0.5525	695	20
20	2.8308	0.5924	706	21
21	3.2173	0.6559	720	24
22	3.7177	0.7701	738	28
23	4.5022	1.0438	767	38
24	5.9546	2.0215	819	73

Table A10: Form A Writing Raw Score to Scale Score Conversion Table for Elementary (Grades 3-5)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.9411	2.0339	425	74
1	-3.4517	1.0667	479	39
2	-2.6211	0.7979	509	29
3	-2.0803	0.6835	529	25
4	-1.661	0.6159	544	22
5	-1.3111	0.5691	556	21
6	-1.0078	0.5335	567	19
7	-0.7386	0.5049	577	18
8	-0.4959	0.4812	586	17
9	-0.274	0.4614	594	17
10	-0.0687	0.445	601	16
11	0.1232	0.4318	608	16
12	0.3052	0.4219	615	15
13	0.4802	0.4153	621	15
14	0.6512	0.4123	627	15
15	0.8211	0.4129	634	15
16	0.9933	0.4175	640	15
17	1.1709	0.4262	646	15
18	1.3579	0.4394	653	16
19	1.5586	0.4573	660	17
20	1.7781	0.4803	668	17
21	2.0221	0.5085	677	18
22	2.2976	0.5422	687	20
23	2.6133	0.5824	698	21
24	2.9811	0.6322	712	23
25	3.4226	0.7006	728	25
26	3.9881	0.8136	748	29
27	4.8441	1.0786	779	39
28	6.3531	2.0404	834	74

Table A11: Form A Total Raw Score to Scale Score Conversion Table for Middle Grades (Grades 6-8)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-6.415	2.0085	372	73
1	-5.0035	1.0166	423	37
2	-4.2771	0.73	449	26
3	-3.8388	0.6049	465	22
4	-3.5188	0.5311	477	19
5	-3.2637	0.4814	486	17
6	-3.0499	0.445	494	16
7	-2.8646	0.417	500	15
8	-2.7001	0.3946	506	14
9	-2.5518	0.3762	512	14
10	-2.4161	0.3608	517	13
11	-2.2909	0.3476	521	13
12	-2.174	0.3362	525	12
13	-2.0644	0.3262	529	12
14	-1.9609	0.3173	533	11
15	-1.8628	0.3094	537	11
16	-1.7693	0.3023	540	11
17	-1.6798	0.2958	543	11
18	-1.5942	0.2899	546	10
19	-1.5117	0.2846	549	10
20	-1.432	0.2796	552	10
21	-1.3553	0.275	555	10
22	-1.2808	0.2707	558	10
23	-1.2086	0.2667	560	10
24	-1.1385	0.263	563	10
25	-1.0702	0.2595	565	9
26	-1.0036	0.2562	568	9
27	-0.9389	0.2531	570	9
28	-0.8756	0.2501	572	9
29	-0.8138	0.2473	574	9
30	-0.7533	0.2447	577	9
31	-0.6939	0.2422	579	9
32	-0.6359	0.2398	581	9
33	-0.5789	0.2375	583	9
34	-0.523	0.2353	585	9
35	-0.4682	0.2333	587	8
36	-0.4142	0.2314	589	8
37	-0.361	0.2295	591	8
38	-0.3088	0.2278	593	8
39	-0.2573	0.2262	595	8
40	-0.2065	0.2247	596	8
41	-0.1563	0.2232	598	8
42	-0.1067	0.2219	600	8
43	-0.0578	0.2207	602	8
44	-0.0093	0.2195	604	8
45	0.0386	0.2185	605	8
46	0.0862	0.2175	607	8
47	0.1332	0.2167	609	8
48	0.18	0.2159	610	8
49	0.2265	0.2152	612	8
50	0.2727	0.2146	614	8
51	0.3186	0.214	615	8
52	0.3643	0.2136	617	8
53	0.4098	0.2132	619	8
54	0.4552	0.2129	620	8
55	0.5004	0.2126	622	8
56	0.5456	0.2124	624	8
57	0.5907	0.2123	625	8
58	0.6358	0.2122	627	8

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
59	0.6807	0.2122	629	8
60	0.7258	0.2123	630	8
61	0.7708	0.2123	632	8
62	0.816	0.2125	633	8
63	0.8612	0.2127	635	8
64	0.9065	0.2129	637	8
65	0.9518	0.2132	638	8
66	0.9974	0.2135	640	8
67	1.043	0.2139	642	8
68	1.0889	0.2144	643	8
69	1.135	0.2149	645	8
70	1.1813	0.2154	647	8
71	1.2278	0.216	648	8
72	1.2746	0.2167	650	8
73	1.3218	0.2175	652	8
74	1.3693	0.2183	654	8
75	1.4171	0.2192	655	8
76	1.4653	0.2202	657	8
77	1.5141	0.2213	659	8
78	1.5633	0.2225	660	8
79	1.6131	0.2237	662	8
80	1.6634	0.2251	664	8
81	1.7144	0.2266	666	8
82	1.7662	0.2283	668	8
83	1.8186	0.2301	670	8
84	1.872	0.232	672	8
85	1.9263	0.234	674	8
86	1.9817	0.2362	676	9
87	2.038	0.2386	678	9
88	2.0955	0.2411	680	9
89	2.1543	0.2438	682	9
90	2.2143	0.2467	684	9
91	2.276	0.2497	686	9
92	2.3391	0.253	689	9
93	2.404	0.2564	691	9
94	2.4707	0.2601	693	9
95	2.5393	0.264	696	10
96	2.6101	0.2681	698	10
97	2.6832	0.2725	701	10
98	2.7587	0.2772	704	10
99	2.8369	0.2822	707	10
100	2.9181	0.2875	710	10
101	3.0023	0.2933	713	11
102	3.0902	0.2994	716	11
103	3.1818	0.3061	719	11
104	3.2776	0.3133	723	11
105	3.3783	0.3212	728	12
106	3.4843	0.33	730	12
107	3.5963	0.3397	734	12
108	3.7153	0.3506	738	13
109	3.8426	0.3631	743	13
110	3.9797	0.3776	748	14
111	4.1286	0.3949	753	14
112	4.2927	0.416	759	15
113	4.4766	0.4426	766	16
114	4.6877	0.4775	774	17
115	4.9382	0.5259	783	19
116	5.2516	0.5985	794	22
117	5.681	0.7232	809	26
118	6.396	1.0103	835	37
119	7.7973	2.0049	886	73

Table A12: Form A Listening Raw Score to Scale Score Conversion Table for Middle Grades (Grades 6-8)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.1712	2.0356	417	74
1	-3.6756	1.0718	471	39
2	-2.83	0.81	502	29
3	-2.2647	0.7049	522	26
4	-1.8099	0.6485	538	23
5	-1.4132	0.6135	553	22
6	-1.052	0.5896	566	21
7	-0.7153	0.5717	578	21
8	-0.3967	0.5577	590	20
9	-0.092	0.5469	601	20
10	0.2028	0.5394	611	20
11	0.4913	0.5354	622	19
12	0.7777	0.5358	632	19
13	1.0673	0.5414	643	20
14	1.3664	0.5537	653	20
15	1.684	0.5752	665	21
16	2.0339	0.6107	678	22
17	2.441	0.6703	692	24
18	2.9591	0.7808	711	28
19	3.7592	1.0508	740	38
20	5.2222	2.025	793	73

Table A13: Form A Speaking Raw Score to Scale Score Conversion Table for Middle Grades (Grades 6-8)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.1656	2.0108	417	73
1	-3.7466	1.0217	468	37
2	-3.0092	0.7379	495	27
3	-2.5587	0.6154	511	22
4	-2.2252	0.5442	523	20
5	-1.9555	0.497	533	18
6	-1.7257	0.4631	541	17
7	-1.5235	0.4373	549	16
8	-1.3415	0.4169	555	15
9	-1.1746	0.4002	561	14
10	-1.0202	0.3862	567	14
11	-0.8758	0.3743	572	14
12	-0.7396	0.364	577	13
13	-0.6103	0.3551	582	13
14	-0.4869	0.3475	586	13
15	-0.3685	0.341	591	12
16	-0.2541	0.3356	595	12
17	-0.1429	0.3314	599	12
18	-0.0342	0.3283	603	12
19	0.073	0.3264	607	12
20	0.1792	0.3257	610	12
21	0.2853	0.3262	614	12
22	0.3922	0.3281	618	12
23	0.5009	0.3314	622	12
24	0.6123	0.3362	626	12
25	0.7273	0.3426	630	12
26	0.8475	0.3509	635	13
27	0.9742	0.3612	639	13
28	1.1091	0.3738	644	14
29	1.2545	0.3893	649	14
30	1.4132	0.4081	655	15
31	1.589	0.4313	661	16
32	1.7873	0.4603	669	17
33	2.0162	0.4979	677	18
34	2.2887	0.5486	687	20
35	2.6291	0.6229	699	23
36	3.0914	0.7476	716	27
37	3.8457	1.0313	743	37
38	5.2812	2.0168	795	73

Table A14: Form A Reading Raw Score to Scale Score Conversion Table for Middle Grades (Grades 6-8)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.9214	2.0551	426	74
1	-3.3691	1.1038	482	40
2	-2.4637	0.8399	515	30
3	-1.8602	0.7239	537	26
4	-1.389	0.6531	554	24
5	-0.9956	0.6036	568	22
6	-0.6542	0.5663	580	20
7	-0.3504	0.5371	591	19
8	-0.0747	0.5136	601	19
9	0.1791	0.4946	610	18
10	0.4158	0.4791	619	17
11	0.6392	0.4665	627	17
12	0.8519	0.4564	635	17
13	1.0564	0.4485	642	16
14	1.2549	0.4427	649	16
15	1.449	0.4387	656	16
16	1.6404	0.4367	663	16
17	1.8309	0.4365	670	16
18	2.0221	0.4384	677	16
19	2.2159	0.4427	684	16
20	2.4148	0.4496	691	16
21	2.6214	0.46	699	17
22	2.8394	0.4746	707	17
23	3.0741	0.4953	715	18
24	3.3332	0.5245	725	19
25	3.6297	0.5673	735	21
26	3.9877	0.6341	748	23
27	4.4605	0.7523	765	27
28	5.2182	1.0312	793	37
29	6.6512	2.0152	845	73

Table A15: Form A Writing Raw Score to Scale Score Conversion Table for Middle Grades (Grades 6-8)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.6714	2.0419	435	74
1	-3.1583	1.0809	490	39
2	-2.2988	0.8148	521	29
3	-1.7328	0.7001	541	25
4	-1.2923	0.6314	557	23
5	-0.9246	0.5835	570	21
6	-0.6059	0.547	582	20
7	-0.3228	0.5178	592	19
8	-0.0675	0.4934	601	18
9	0.1654	0.4723	610	17
10	0.3797	0.4538	618	16
11	0.5782	0.4375	625	16
12	0.7633	0.4236	632	15
13	0.9378	0.412	638	15
14	1.1037	0.4032	644	15
15	1.2638	0.3974	650	14
16	1.4205	0.3949	655	14
17	1.5765	0.3958	661	14
18	1.7347	0.4003	667	14
19	1.8981	0.4085	673	15
20	2.0697	0.4206	679	15
21	2.253	0.4362	685	16
22	2.4512	0.4548	693	16
23	2.6675	0.4755	700	17
24	2.9039	0.4969	709	18
25	3.1615	0.5179	718	19
26	3.4406	0.5388	728	19
27	3.7435	0.5625	739	20
28	4.0774	0.5956	751	22
29	4.4629	0.6509	765	24
30	4.9509	0.758	783	27
31	5.7104	1.0287	811	37
32	7.1356	2.0113	862	73

Table A16: Form A Total Raw Score to Scale Score Conversion Table for High School (Grades 9-12)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-6.0933	2.0117	383	73
1	-4.6723	1.0229	435	37
2	-3.9334	0.7385	462	27
3	-3.4831	0.6146	478	22
4	-3.1514	0.5418	490	20
5	-2.8851	0.4927	500	18
6	-2.6605	0.4566	508	17
7	-2.4649	0.4288	515	16
8	-2.2908	0.4064	521	15
9	-2.1332	0.3879	527	14
10	-1.9889	0.3721	532	13
11	-1.8556	0.3586	537	13
12	-1.7313	0.3467	541	13
13	-1.6148	0.3361	546	12
14	-1.5051	0.3266	549	12
15	-1.4013	0.318	553	12
16	-1.3027	0.3101	557	11
17	-1.2087	0.3029	560	11
18	-1.119	0.2963	563	11
19	-1.0331	0.2901	567	10
20	-0.9506	0.2844	570	10
21	-0.8713	0.279	572	10
22	-0.7948	0.274	575	10
23	-0.721	0.2694	578	10
24	-0.6496	0.265	580	10
25	-0.5804	0.2609	583	9
26	-0.5134	0.2571	585	9
27	-0.4482	0.2536	588	9
28	-0.3848	0.2502	590	9
29	-0.3229	0.2471	592	9
30	-0.2626	0.2441	594	9
31	-0.2037	0.2414	597	9
32	-0.146	0.2388	599	9
33	-0.0896	0.2365	601	9
34	-0.0342	0.2342	603	8
35	0.0202	0.2321	605	8
36	0.0737	0.2302	607	8
37	0.1262	0.2284	608	8
38	0.178	0.2267	610	8
39	0.229	0.2252	613	8
40	0.2795	0.2237	614	8
41	0.3292	0.2224	616	8
42	0.3783	0.2211	618	8
43	0.427	0.22	619	8
44	0.4752	0.2189	621	8
45	0.5229	0.2179	623	8
46	0.5701	0.217	625	8
47	0.6171	0.2161	626	8
48	0.6636	0.2153	628	8
49	0.7098	0.2146	630	8
50	0.7556	0.2139	631	8
51	0.8013	0.2133	633	8
52	0.8466	0.2127	635	8
53	0.8918	0.2121	636	8
54	0.9366	0.2116	638	8
55	0.9813	0.2112	639	8
56	1.0258	0.2107	641	8
57	1.0702	0.2103	643	8
58	1.1144	0.21	644	8

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
59	1.1584	0.2096	646	8
60	1.2023	0.2093	647	8
61	1.246	0.2091	649	8
62	1.2896	0.2088	651	8
63	1.3332	0.2086	652	8
64	1.3767	0.2085	654	8
65	1.4202	0.2084	655	8
66	1.4636	0.2083	657	8
67	1.5069	0.2083	658	8
68	1.5504	0.2084	660	8
69	1.5938	0.2085	662	8
70	1.6373	0.2086	663	8
71	1.6809	0.2089	665	8
72	1.7246	0.2092	666	8
73	1.7685	0.2096	668	8
74	1.8125	0.2101	670	8
75	1.8568	0.2107	671	8
76	1.9013	0.2114	673	8
77	1.9461	0.2122	675	8
78	1.9914	0.2131	676	8
79	2.0369	0.2141	678	8
80	2.083	0.2152	679	8
81	2.1295	0.2164	681	8
82	2.1767	0.2178	683	8
83	2.2245	0.2193	684	8
84	2.2729	0.2208	686	8
85	2.3221	0.2226	688	8
86	2.372	0.2244	690	8
87	2.4227	0.2264	692	8
88	2.4745	0.2285	693	8
89	2.5272	0.2307	695	8
90	2.581	0.2331	697	8
91	2.6359	0.2356	699	9
92	2.692	0.2383	701	9
93	2.7495	0.2411	703	9
94	2.8083	0.2441	706	9
95	2.8686	0.2472	708	9
96	2.9306	0.2506	710	9
97	2.9943	0.2542	712	9
98	3.0598	0.258	715	9
99	3.1275	0.2621	717	9
100	3.1973	0.2664	720	10
101	3.2695	0.2712	722	10
102	3.3445	0.2763	725	10
103	3.4223	0.2818	728	10
104	3.5033	0.2878	732	10
105	3.5881	0.2944	735	11
106	3.6769	0.3018	737	11
107	3.7704	0.3099	740	11
108	3.8692	0.319	744	12
109	3.9742	0.3293	748	12
110	4.0866	0.3411	752	12
111	4.2076	0.3547	756	13
112	4.339	0.3707	761	13
113	4.4833	0.3896	766	14
114	4.644	0.4127	772	15
115	4.8259	0.4415	779	16
116	5.037	0.4788	786	17
117	5.2899	0.5296	795	19
118	5.6089	0.6045	807	22
119	6.0473	0.7308	823	26

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
120	6.7758	1.0181	849	37
121	8.1905	2.0098	900	73

Table A17: Form A Listening Raw Score to Scale Score Conversion Table for High School (Grades 9-12)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.6927	2.0477	434	74
1	-3.155	1.1003	490	40
2	-2.2392	0.8569	523	31
3	-1.589	0.7666	546	28
4	-1.0403	0.7181	566	26
5	-0.5508	0.682	584	25
6	-0.1077	0.6497	600	24
7	0.2953	0.6206	615	22
8	0.665	0.5962	628	22
9	1.0088	0.5775	640	21
10	1.3344	0.5647	652	20
11	1.6487	0.5577	664	20
12	1.9586	0.5564	675	20
13	2.2701	0.5611	686	20
14	2.5908	0.5727	698	21
15	2.9298	0.5935	710	21
16	3.3011	0.6278	723	23
17	3.7291	0.6856	739	25
18	4.2677	0.7936	758	29
19	5.0874	1.0598	788	38
20	6.5642	2.0293	841	73

Table A18: Form A Speaking Raw Score to Scale Score Conversion Table for High School (Grades 9-12)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-5.1547	2.0299	417	73
1	-3.6814	1.0542	471	38
2	-2.8827	0.7746	500	28
3	-2.3822	0.6509	518	24
4	-2.0084	0.5764	531	21
5	-1.7065	0.5248	542	19
6	-1.4518	0.4861	551	18
7	-1.2305	0.4557	559	16
8	-1.0343	0.4309	567	16
9	-0.8575	0.4105	573	15
10	-0.696	0.3935	579	14
11	-0.5469	0.3793	584	14
12	-0.4076	0.3675	589	13
13	-0.2763	0.3577	594	13
14	-0.1512	0.3498	598	13
15	-0.0311	0.3435	603	12
16	0.0851	0.3387	607	12
17	0.1986	0.3353	611	12
18	0.3103	0.3332	615	12
19	0.4209	0.3323	619	12
20	0.5313	0.3326	623	12
21	0.6424	0.3341	627	12
22	0.7548	0.3368	631	12
23	0.8695	0.3408	635	12
24	0.9874	0.3461	640	13
25	1.1095	0.3529	644	13
26	1.2369	0.3614	649	13
27	1.3712	0.3718	654	13
28	1.514	0.3845	659	14
29	1.6676	0.3998	664	14
30	1.8348	0.4186	670	15
31	2.0195	0.4417	677	16
32	2.2273	0.4709	685	17
33	2.4664	0.5087	693	18
34	2.7509	0.5603	703	20
35	3.1057	0.6358	716	23
36	3.5868	0.7621	734	28
37	4.3672	1.0463	762	38
38	5.8295	2.0269	815	73

Table A19: Form A Reading Raw Score to Scale Score Conversion Table for High School (Grades 9-12)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-4.5376	2.0627	440	75
1	-2.9641	1.1148	497	40
2	-2.0409	0.8461	530	31
3	-1.434	0.7216	552	26
4	-0.9716	0.6426	569	23
5	-0.5954	0.5867	582	21
6	-0.2761	0.545	594	20
7	0.003	0.513	604	19
8	0.2532	0.4882	613	18
9	0.4817	0.4686	621	17
10	0.6938	0.4532	629	16
11	0.8935	0.4411	636	16
12	1.0838	0.4317	643	16
13	1.2671	0.4246	650	15
14	1.445	0.4195	656	15
15	1.6194	0.4162	663	15
16	1.7919	0.4145	669	15
17	1.9636	0.4145	675	15
18	2.1359	0.416	681	15
19	2.3102	0.4192	688	15
20	2.4879	0.4242	694	15
21	2.6708	0.4313	701	16
22	2.8608	0.4409	707	16
23	3.0605	0.4535	715	16
24	3.2735	0.4702	722	17
25	3.5048	0.4926	731	18
26	3.762	0.5232	740	19
27	4.0579	0.5672	751	21
28	4.4161	0.6348	764	23
29	4.8904	0.7535	781	27
30	5.6503	1.0324	808	37
31	7.0854	2.0161	860	73

Table A20: Form A Writing Raw Score to Scale Score Conversion Table for High School (Grades 9-12)

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
0	-3.9761	2.0379	460	74
1	-2.4761	1.0724	514	39
2	-1.6361	0.8022	545	29
3	-1.0912	0.6843	564	25
4	-0.6729	0.6134	580	22
5	-0.3277	0.5638	592	20
6	-0.0314	0.526	603	19
7	0.229	0.4955	612	18
8	0.4615	0.4695	621	17
9	0.6713	0.4468	628	16
10	0.8618	0.4265	635	15
11	1.0361	0.4086	641	15
12	1.1966	0.3932	647	14
13	1.3461	0.3803	653	14
14	1.4866	0.3701	658	13
15	1.6208	0.3629	663	13
16	1.7509	0.3587	667	13
17	1.8789	0.3574	672	13
18	2.007	0.3592	677	13
19	2.1376	0.3639	681	13
20	2.2727	0.3718	686	13
21	2.4148	0.3827	691	14
22	2.5666	0.3968	697	14
23	2.7309	0.4141	703	15
24	2.9108	0.435	709	16
25	3.1109	0.4601	716	17
26	3.3365	0.4909	725	18
27	3.5963	0.5301	734	19
28	3.9048	0.5833	745	21
29	4.289	0.6612	759	24
30	4.8083	0.7908	778	29
31	5.6421	1.0762	808	39
32	7.1582	2.0464	863	74

APPENDIX B: WLPT-II (FORM A) ITEM DIFFICULTY, FIT STATISTICS, AND CLASSICAL ITEM STATISTICS

Table B1: Form A Primary (Grades K-2)

	Primary	Grade K	Grade 1	Grade 2
N-Count	37,656	13,153	13,183	11,320

		Primary			Grade K		Grade 1		Grade 2	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Listening	1	-3.3113	1.93	2.03	0.73	0.35	0.89	0.31	0.94	0.27
Listening	2	-3.0921	1.09	0.75	0.73	0.43	0.93	0.34	0.97	0.29
Listening	3	-4.6047	1.16	0.68	0.93	0.33	0.98	0.22	0.99	0.24
Listening	4	-4.3906	0.68	0.38	0.95	0.31	0.99	0.17	1.00	0.19
Listening	5	-4.8684	1.03	0.79	0.95	0.30	0.99	0.16	1.00	0.14
Listening	6	-3.3033	0.78	0.41	0.82	0.47	0.96	0.34	0.98	0.35
Listening	7	-5.3779	0.93	0.41	0.97	0.29	0.99	0.17	1.00	0.20
Listening	8	-5.2033	0.95	0.57	0.96	0.30	0.99	0.16	1.00	0.20
Listening	9	-3.7251	0.92	0.89	0.89	0.35	0.97	0.21	0.98	0.16
Listening	10	-4.5286	1.07	0.80	0.93	0.37	0.98	0.21	0.99	0.22
Listening	11	-2.7992	0.90	0.88	0.83	0.29	0.93	0.25	0.96	0.26
Listening	12	-0.7660	1.23	1.37	0.58	0.13	0.65	0.16	0.74	0.21
Listening	13	-2.7328	1.03	1.25	0.81	0.26	0.91	0.20	0.95	0.21
Listening	14	-2.6201	0.83	0.59	0.73	0.46	0.92	0.37	0.96	0.36
Listening	15	-2.9102	1.08	1.11	0.76	0.39	0.92	0.26	0.96	0.21
Listening	16	-2.2514	0.94	1.11	0.73	0.35	0.89	0.25	0.93	0.21
Listening	17	-1.0106	1.47	1.87	0.64	0.12	0.65	0.08	0.66	0.13
Listening	18	-1.6885	1.36	1.71	0.58	0.26	0.71	0.18	0.84	0.21
Listening	19	-0.1341	1.21	1.32	0.32	0.19	0.45	0.23	0.55	0.25
Listening	20	1.5705	1.21	1.94	0.18	0.01	0.17	0.03	0.20	0.06
Writing Conventions	21	-1.8684	0.98	0.83	0.54	0.29	0.82	0.42	0.92	0.40
Writing Conventions	22	-2.8011	1.08	0.76	0.69	0.36	0.91	0.36	0.96	0.31
Writing Conventions	23	-1.3359	0.83	0.67	0.34	0.35	0.75	0.50	0.91	0.46
Writing Conventions	24	-0.3731	0.90	0.90	0.22	0.13	0.52	0.45	0.76	0.44
Writing Conventions	25	-0.6812	0.94	0.86	0.27	0.22	0.53	0.42	0.80	0.50
Writing Conventions	26	-1.2360	0.87	0.72	0.24	0.33	0.71	0.46	0.90	0.48
Writing Conventions	27	0.6686	1.15	1.19	0.12	0.22	0.32	0.17	0.36	0.14

		Primary			Grade K		Grade 1		Grade 2	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Writing Conventions	28	0.5776	1.10	1.18	0.16	0.17	0.28	0.12	0.40	0.25
Writing Conventions	29	0.6260	1.13	1.27	0.20	0.12	0.30	0.16	0.44	0.27
Writing Conventions	30	1.0837	1.02	1.04	0.07	0.20	0.26	0.27	0.40	0.25
Writing Conventions	31	-0.5787	0.97	0.94	0.22	0.21	0.53	0.40	0.75	0.44
Writing Conventions	32	-0.8258	0.86	0.78	0.14	0.31	0.59	0.51	0.83	0.49
Writing Conventions	33	-0.7468	0.93	0.84	0.27	0.23	0.57	0.42	0.82	0.50
Writing Conventions	34	-0.6747	0.87	0.81	0.11	0.29	0.52	0.52	0.80	0.48
Writing Conventions	35	-0.8324	0.99	0.89	0.17	0.28	0.51	0.45	0.79	0.50
Reading	36	-2.5161	0.95	0.74	0.57	0.41	0.92	0.30	0.97	0.20
Reading	37	-1.4154	0.84	0.71	0.36	0.44	0.77	0.46	0.91	0.43
Reading	38	-0.8084	0.96	0.87	0.30	0.29	0.58	0.38	0.83	0.43
Reading	39	-2.0121	0.98	0.70	0.35	0.41	0.84	0.45	0.96	0.41
Reading	40	-0.1488	0.87	0.83	0.17	0.23	0.43	0.40	0.73	0.51
Reading	41	-1.1836	1.00	0.89	0.25	0.22	0.64	0.46	0.87	0.42
Reading	42	-0.3600	0.89	0.86	0.17	0.20	0.47	0.42	0.77	0.45
Reading	43	-0.5787	0.90	0.83	0.13	0.23	0.47	0.47	0.78	0.50
Reading	44	-0.3792	1.08	1.09	0.23	0.20	0.47	0.31	0.66	0.36
Reading	45	-0.2815	1.06	1.08	0.21	0.20	0.43	0.30	0.66	0.36
Reading	46	-0.1789	0.95	0.90	0.15	0.23	0.37	0.35	0.66	0.48
Reading	47	-0.1909	0.83	0.77	0.10	0.22	0.40	0.45	0.72	0.55
Reading	48	-0.1728	1.11	1.14	0.24	0.19	0.42	0.24	0.60	0.33
Reading	49	0.6686	0.84	0.83	0.08	0.18	0.22	0.33	0.44	0.44
Reading	50	-0.2150	0.80	0.74	0.06	0.24	0.37	0.50	0.75	0.52
Reading	51	0.0656	0.94	0.92	0.13	0.23	0.38	0.34	0.64	0.39
Reading	52	0.5776	1.05	1.16	0.18	0.13	0.30	0.21	0.46	0.33
Reading	53	0.5234	0.93	0.93	0.10	0.19	0.24	0.25	0.47	0.37
Reading	54	0.6657	0.82	0.78	0.05	0.20	0.26	0.40	0.54	0.44
Reading	55	0.3610	0.97	0.97	0.13	0.24	0.33	0.28	0.56	0.37
Reading	56	0.3119	0.91	0.91	0.12	0.23	0.32	0.31	0.59	0.42
Reading	57	0.7774	0.95	0.98	0.09	0.19	0.29	0.29	0.54	0.37
Reading	58	1.0213	0.88	0.90	0.07	0.18	0.23	0.32	0.51	0.43
Reading	59	0.3968	0.91	0.91	0.12	0.21	0.34	0.33	0.65	0.43
Writing	60	-1.9049	0.57	0.50	0.89	0.22	0.96	0.13	0.98	0.09

		Primary			Grade K		Grade 1		Grade 2	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Writing	61	-1.7508	0.76	0.70	0.68	0.27	0.90	0.21	0.95	0.16
Writing	62	-0.7242	0.92	1.02	1.14	0.27	1.48	0.25	1.57	0.19
Writing	63	-1.1132	0.94	1.08	1.45	0.44	1.84	0.33	1.93	0.27
Writing	64	-0.7983	1.05	0.94	0.59	0.48	1.48	0.54	1.77	0.50
Writing	65	0.2782	0.54	0.57	0.38	0.49	1.02	0.56	1.27	0.49
Writing	66	0.6819	0.72	0.75	0.41	0.51	1.61	0.64	2.34	0.61
Writing	67	0.7127	0.71	0.71	0.31	0.50	1.62	0.71	2.50	0.70
Speaking	68	-3.2996	1.13	1.42	1.90	0.35	1.96	0.25	1.97	0.26
Speaking	69	-3.3070	1.13	1.45	1.90	0.35	1.96	0.25	1.97	0.25
Speaking	70	-3.1399	1.03	1.17	1.79	0.41	1.91	0.29	1.94	0.30
Speaking	71	-2.2678	0.92	0.82	1.58	0.50	1.82	0.41	1.91	0.40
Speaking	72	-2.3182	0.82	0.82	1.59	0.53	1.83	0.44	1.91	0.41
Speaking	73	-1.2700	0.88	0.87	1.15	0.62	1.51	0.55	1.70	0.51
Speaking	74	-1.5686	0.90	0.93	1.24	0.61	1.53	0.52	1.69	0.49
Speaking	75	-1.5141	0.85	0.87	1.18	0.65	1.57	0.58	1.75	0.55
Speaking	76	-1.0683	0.89	0.94	1.13	0.60	1.45	0.50	1.62	0.47
Speaking	77	-1.4579	0.83	0.86	1.28	0.64	1.62	0.54	1.75	0.50
Speaking	78	-0.7564	1.03	1.04	2.13	0.70	2.84	0.63	3.22	0.59
Speaking	79	-0.9059	1.08	1.13	2.08	0.69	2.74	0.61	3.09	0.56
Speaking	80	-0.9587	1.05	1.03	1.13	0.58	1.51	0.49	1.69	0.48
Speaking	81	-1.8095	0.86	0.89	1.39	0.58	1.79	0.48	1.87	0.47
Speaking	82	-0.8139	1.04	1.03	1.03	0.60	1.46	0.51	1.64	0.49
Speaking	83	-0.9666	0.91	0.87	1.13	0.63	1.57	0.54	1.74	0.50
Speaking	84	-1.5877	0.99	1.21	1.35	0.57	1.70	0.46	1.80	0.44

Table B2: Form A Elementary (Grades 3-5)

	Elementary	Grade 3	Grade 4	Grade 5
N-Count	19,948	7,605	6,732	5,611

		Elementary			Grade 3		Grade 4		Grade 5	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Listening	1	-2.2877	0.98	0.86	0.95	0.25	0.98	0.22	0.98	0.26
Listening	2	-0.3914	0.95	0.93	0.80	0.36	0.85	0.38	0.87	0.40
Listening	3	-0.6153	1.07	1.13	0.83	0.21	0.84	0.23	0.88	0.28
Listening	4	-1.4495	0.92	0.81	0.93	0.34	0.95	0.34	0.94	0.38
Listening	5	-0.2230	1.17	1.37	0.66	0.37	0.73	0.33	0.76	0.36
Listening	6	0.9455	0.96	0.96	0.66	0.40	0.76	0.40	0.81	0.43
Listening	7	-1.0220	1.07	1.21	0.71	0.30	0.76	0.29	0.80	0.29
Listening	8	1.4211	1.09	1.14	0.51	0.34	0.62	0.35	0.70	0.38
Listening	9	-0.0843	1.10	1.22	0.61	0.35	0.71	0.38	0.77	0.41
Listening	10	0.2045	1.13	1.18	0.60	0.32	0.69	0.34	0.74	0.32
Listening	11	0.4647	0.96	0.95	0.77	0.47	0.84	0.44	0.86	0.46
Listening	12	-0.1931	1.21	1.29	0.67	0.36	0.73	0.36	0.76	0.38
Listening	13	0.6603	1.11	1.15	0.42	0.29	0.52	0.29	0.58	0.33
Listening	14	0.5817	0.99	0.96	0.79	0.43	0.85	0.41	0.88	0.42
Listening	15	0.6829	1.05	1.08	0.53	0.32	0.60	0.28	0.65	0.31
Listening	16	-0.4832	0.98	0.94	0.76	0.18	0.79	0.17	0.81	0.17
Listening	17	0.3613	1.05	1.10	0.58	0.37	0.68	0.37	0.74	0.41
Listening	18	1.2579	1.06	1.12	0.86	0.23	0.90	0.23	0.92	0.25
Listening	19	-0.2852	0.93	0.89	0.44	0.26	0.53	0.28	0.60	0.31
Listening	20	1.0147	1.08	1.13	0.76	0.31	0.83	0.29	0.85	0.27
Writing Conventions	21	-1.6205	0.97	0.85	0.91	0.34	0.95	0.32	0.96	0.30
Writing Conventions	22	-2.2616	0.97	0.78	0.95	0.28	0.98	0.27	0.98	0.27
Writing Conventions	23	-1.7352	0.68	0.44	0.95	0.38	0.97	0.34	0.98	0.35
Writing Conventions	24	-1.5157	0.88	0.59	0.91	0.44	0.95	0.44	0.96	0.42
Writing Conventions	25	0.4838	0.89	0.85	0.64	0.42	0.75	0.42	0.79	0.45
Writing Conventions	26	0.7020	0.96	0.94	0.55	0.38	0.66	0.39	0.73	0.41
Writing Conventions	27	-0.1449	1.01	0.97	0.71	0.42	0.79	0.42	0.83	0.41
Writing Conventions	28	-0.0375	0.92	0.86	0.74	0.45	0.81	0.43	0.85	0.46

Modality	Item Sequence	Elementary			Grade 3		Grade 4		Grade 5	
		Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Writing Conventions	29	0.1321	0.93	0.85	0.65	0.44	0.76	0.45	0.83	0.50
Writing Conventions	30	-0.4593	0.81	0.71	0.81	0.49	0.87	0.46	0.91	0.49
Writing Conventions	31	0.4342	0.94	0.96	0.66	0.43	0.74	0.38	0.79	0.43
Writing Conventions	32	0.8200	1.00	1.04	0.60	0.38	0.67	0.34	0.72	0.32
Writing Conventions	33	1.0098	1.08	1.10	0.40	0.29	0.53	0.33	0.62	0.34
Writing Conventions	34	0.0262	1.07	1.03	0.63	0.44	0.77	0.46	0.83	0.46
Writing Conventions	35	0.5327	1.05	1.03	0.55	0.35	0.68	0.38	0.77	0.44
Writing Conventions	36	1.4211	1.04	1.08	0.40	0.31	0.49	0.32	0.54	0.37
Writing Conventions	37	-0.0843	1.01	0.88	0.63	0.51	0.79	0.55	0.87	0.60
Writing Conventions	38	1.0589	1.19	1.27	0.43	0.22	0.49	0.22	0.56	0.26
Writing Conventions	39	2.5328	1.13	1.63	0.22	0.01	0.26	0.09	0.31	0.15
Writing Conventions	40	1.0658	1.19	1.27	0.49	0.14	0.55	0.16	0.63	0.23
Reading	41	-1.7352	1.03	1.15	0.92	0.27	0.94	0.24	0.95	0.21
Reading	42	-2.4489	1.10	1.29	0.95	0.27	0.97	0.26	0.98	0.22
Reading	43	-1.8845	0.85	0.65	0.95	0.40	0.96	0.37	0.97	0.36
Reading	44	-1.9452	1.11	0.94	0.93	0.39	0.95	0.39	0.97	0.34
Reading	45	-1.6497	1.22	1.02	0.90	0.47	0.94	0.40	0.96	0.41
Reading	46	-0.1245	1.10	1.10	0.68	0.42	0.77	0.40	0.83	0.38
Reading	47	0.9355	1.08	1.11	0.46	0.29	0.57	0.35	0.65	0.37
Reading	48	0.9455	0.94	0.93	0.44	0.39	0.57	0.45	0.68	0.49
Reading	49	0.0703	0.86	0.72	0.66	0.52	0.80	0.53	0.86	0.55
Reading	50	0.5917	0.86	0.80	0.56	0.44	0.71	0.50	0.80	0.51
Reading	51	0.2810	0.86	0.76	0.62	0.47	0.76	0.53	0.84	0.56
Reading	52	0.8200	0.91	0.86	0.48	0.45	0.61	0.50	0.72	0.53
Reading	53	0.4892	0.87	0.81	0.64	0.42	0.73	0.44	0.81	0.48
Reading	54	2.4213	1.00	1.23	0.19	0.12	0.27	0.21	0.35	0.29
Reading	55	1.1806	1.10	1.15	0.46	0.25	0.54	0.28	0.61	0.33
Reading	56	2.1192	0.99	1.15	0.25	0.22	0.40	0.34	0.54	0.40
Reading	57	2.0336	1.08	1.22	0.31	0.18	0.40	0.24	0.49	0.30
Reading	58	2.1635	1.05	1.18	0.29	0.19	0.36	0.24	0.43	0.30
Reading	59	2.7676	1.05	1.39	0.19	0.09	0.25	0.19	0.35	0.27
Reading	60	0.9372	0.95	0.92	0.51	0.37	0.64	0.43	0.73	0.46
Reading	61	0.9878	0.90	0.87	0.47	0.40	0.63	0.46	0.73	0.47

Modality	Item Sequence	Elementary			Grade 3		Grade 4		Grade 5	
		Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Reading	62	1.5127	0.93	0.96	0.39	0.33	0.52	0.42	0.61	0.45
Reading	63	2.3823	1.11	1.35	0.26	0.15	0.33	0.20	0.39	0.23
Reading	64	2.0697	1.12	1.33	0.32	0.16	0.38	0.21	0.44	0.24
Writing	65	1.9932	1.34	1.41	1.83	0.53	2.14	0.53	2.36	0.53
Writing	66	2.0181	1.40	1.51	1.82	0.55	2.09	0.53	2.28	0.51
Speaking	67	-1.9361	1.05	1.27	1.97	0.26	1.98	0.23	1.97	0.23
Speaking	68	-1.9193	1.01	1.09	1.95	0.29	1.96	0.27	1.96	0.26
Speaking	69	-1.7695	0.92	0.86	1.94	0.36	1.95	0.34	1.96	0.33
Speaking	70	-1.3027	0.65	0.54	1.92	0.41	1.94	0.39	1.94	0.41
Speaking	71	-0.8569	0.67	0.62	1.85	0.47	1.89	0.47	1.89	0.50
Speaking	72	-0.5531	0.73	0.68	1.78	0.49	1.81	0.50	1.82	0.52
Speaking	73	-0.4945	0.83	0.83	1.69	0.50	1.74	0.48	1.77	0.53
Speaking	74	0.2341	0.81	0.80	1.54	0.52	1.62	0.51	1.66	0.56
Speaking	75	-0.5712	0.62	0.52	1.84	0.54	1.88	0.54	1.89	0.56
Speaking	76	-0.4241	0.77	0.74	1.73	0.47	1.78	0.47	1.81	0.48
Speaking	77	0.3302	0.94	0.93	3.15	0.59	3.29	0.58	3.35	0.62
Speaking	78	0.1102	0.97	1.02	3.10	0.56	3.24	0.53	3.32	0.58
Speaking	79	-0.4901	1.07	1.12	1.85	0.29	1.88	0.26	1.90	0.27
Speaking	80	-0.7866	1.01	1.08	1.84	0.40	1.89	0.38	1.90	0.39
Speaking	81	0.0437	0.89	0.84	1.68	0.45	1.78	0.46	1.79	0.48
Speaking	82	-0.7907	0.80	0.69	1.86	0.46	1.90	0.47	1.90	0.49
Speaking	83	-0.0680	0.80	0.71	1.74	0.49	1.82	0.52	1.83	0.55

Table B3: Form A Middle Grades (Grades 6-8)

	Middle Grades	Grade 6	Grade 7	Grade 8
N-Count	12,171	5,010	3,842	3,319

		Middle Grades			Grade 6		Grade 7		Grade 8	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Listening	1	-1.2028	0.84	0.54	0.97	0.39	0.96	0.42	0.95	0.48
Listening	2	-2.4595	1.28	0.63	0.99	0.31	0.98	0.34	0.98	0.38
Listening	3	-1.9277	0.78	0.48	0.99	0.34	0.98	0.38	0.98	0.37
Listening	4	-1.5247	0.84	0.45	0.98	0.38	0.97	0.42	0.96	0.48
Listening	5	-1.9732	0.97	0.81	0.98	0.21	0.98	0.29	0.98	0.30
Listening	6	1.1174	1.57	1.78	0.62	0.23	0.57	0.25	0.59	0.28
Listening	7	1.2194	0.98	0.97	0.75	0.28	0.76	0.32	0.78	0.36
Listening	8	0.7769	1.88	2.11	0.56	0.31	0.59	0.33	0.64	0.37
Listening	9	2.1617	1.00	1.08	0.94	0.37	0.94	0.39	0.93	0.46
Listening	10	-1.4636	3.52	3.66	0.88	0.36	0.88	0.42	0.88	0.48
Listening	11	1.5095	0.80	0.77	0.82	0.34	0.82	0.36	0.82	0.40
Listening	12	0.7581	1.91	2.22	0.58	0.32	0.59	0.32	0.60	0.35
Listening	13	1.7649	0.86	0.86	0.84	0.32	0.82	0.32	0.83	0.34
Listening	14	-0.7846	3.16	3.20	0.81	0.41	0.80	0.47	0.80	0.48
Listening	15	-0.0423	3.69	4.24	0.57	0.40	0.58	0.40	0.59	0.41
Listening	16	0.4702	1.80	2.16	0.71	0.28	0.69	0.30	0.69	0.37
Listening	17	1.6625	0.83	0.82	0.78	0.35	0.77	0.42	0.77	0.45
Listening	18	0.3312	1.36	1.50	0.81	0.27	0.79	0.32	0.79	0.37
Listening	19	0.5065	2.83	3.60	0.49	0.20	0.53	0.21	0.54	0.25
Listening	20	1.6571	0.74	0.73	0.96	0.31	0.96	0.32	0.95	0.39
Writing Conventions	21	-2.1547	1.14	1.28	0.98	0.21	0.98	0.18	0.98	0.22
Writing Conventions	22	-1.7863	1.22	0.90	0.96	0.28	0.98	0.25	0.97	0.30
Writing Conventions	23	-1.2476	0.95	0.68	0.96	0.36	0.95	0.35	0.96	0.39
Writing Conventions	24	-0.6700	0.72	0.46	0.94	0.44	0.94	0.50	0.94	0.52
Writing Conventions	25	0.1350	1.27	1.31	0.84	0.35	0.83	0.37	0.84	0.39
Writing Conventions	26	-0.2090	0.96	0.76	0.90	0.45	0.90	0.48	0.89	0.49
Writing Conventions	27	-0.2145	1.26	1.22	0.86	0.39	0.87	0.40	0.88	0.45
Writing Conventions	28	1.1080	0.86	0.80	0.76	0.47	0.77	0.48	0.77	0.51

		Middle Grades			Grade 6		Grade 7		Grade 8	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Writing Conventions	29	1.2926	1.16	1.18	0.63	0.32	0.69	0.30	0.71	0.29
Writing Conventions	30	0.6704	1.02	0.92	0.78	0.48	0.79	0.52	0.77	0.52
Writing Conventions	31	2.7500	1.09	1.25	0.31	0.11	0.36	0.14	0.40	0.18
Writing Conventions	32	1.5662	1.20	1.25	0.55	0.31	0.56	0.33	0.58	0.37
Writing Conventions	33	0.8510	0.88	0.77	0.79	0.52	0.77	0.56	0.77	0.55
Writing Conventions	34	1.6844	1.01	1.02	0.64	0.36	0.64	0.37	0.62	0.41
Writing Conventions	35	1.4935	1.09	1.09	0.59	0.36	0.63	0.36	0.68	0.36
Writing Conventions	36	0.5696	1.38	1.45	0.73	0.27	0.77	0.30	0.79	0.35
Writing Conventions	37	0.7426	0.96	0.88	0.80	0.45	0.80	0.44	0.79	0.46
Writing Conventions	38	1.6727	1.07	1.10	0.59	0.34	0.63	0.33	0.63	0.36
Writing Conventions	39	1.3889	1.13	1.16	0.65	0.31	0.67	0.34	0.66	0.32
Writing Conventions	40	2.3536	1.17	1.28	0.41	0.15	0.43	0.13	0.46	0.16
Writing Conventions	41	1.7848	1.04	1.06	0.57	0.35	0.61	0.37	0.61	0.37
Writing Conventions	42	0.8589	1.01	0.92	0.75	0.50	0.75	0.53	0.74	0.49
Writing Conventions	43	3.8566	0.99	1.40	0.16	0.06	0.19	0.08	0.19	0.14
Writing Conventions	44	4.0644	1.04	1.63	0.15	0.02	0.16	0.03	0.19	0.05
Reading	45	-2.6812	1.05	0.85	0.99	0.22	0.99	0.21	0.99	0.26
Reading	46	0.3997	1.14	1.19	0.83	0.28	0.84	0.28	0.84	0.30
Reading	47	-0.5830	1.12	1.15	0.92	0.30	0.91	0.34	0.92	0.42
Reading	48	-1.0492	0.79	0.47	0.96	0.42	0.96	0.43	0.95	0.46
Reading	49	-2.0169	0.89	0.68	0.98	0.30	0.98	0.32	0.98	0.35
Reading	50	-0.1186	1.30	1.13	0.84	0.47	0.85	0.48	0.86	0.51
Reading	51	-1.1664	1.16	0.85	0.94	0.39	0.95	0.40	0.95	0.42
Reading	52	1.0612	0.94	0.89	0.74	0.45	0.75	0.50	0.75	0.51
Reading	53	2.5031	0.96	0.99	0.40	0.32	0.44	0.37	0.47	0.38
Reading	54	1.8123	1.19	1.22	0.45	0.29	0.47	0.32	0.51	0.36
Reading	55	1.5571	1.16	1.17	0.56	0.37	0.57	0.36	0.59	0.37
Reading	56	2.0531	1.02	1.02	0.47	0.34	0.52	0.36	0.54	0.40
Reading	57	2.5652	1.12	1.21	0.38	0.14	0.39	0.18	0.43	0.22
Reading	58	0.8994	1.06	0.98	0.70	0.47	0.74	0.50	0.75	0.53
Reading	59	2.3740	0.94	0.96	0.38	0.34	0.43	0.39	0.51	0.43
Reading	60	2.7540	0.95	0.99	0.26	0.24	0.28	0.26	0.31	0.28
Reading	61	1.4208	1.15	1.14	0.57	0.35	0.61	0.35	0.68	0.42

		Middle Grades			Grade 6		Grade 7		Grade 8	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Reading	62	0.2839	1.26	1.26	0.82	0.33	0.83	0.29	0.84	0.29
Reading	63	1.1154	1.12	1.05	0.66	0.43	0.68	0.45	0.70	0.48
Reading	64	2.4525	0.89	0.89	0.45	0.43	0.49	0.43	0.53	0.46
Reading	65	2.6972	1.07	1.18	0.36	0.19	0.36	0.18	0.36	0.20
Reading	66	1.6236	1.24	1.30	0.56	0.24	0.58	0.24	0.62	0.22
Reading	67	0.5586	1.10	0.95	0.79	0.45	0.78	0.45	0.81	0.47
Reading	68	1.1574	1.04	0.98	0.67	0.47	0.69	0.47	0.72	0.50
Reading	69	2.2362	1.01	1.02	0.41	0.33	0.47	0.35	0.50	0.37
Reading	70	2.5502	1.07	1.17	0.38	0.20	0.40	0.21	0.45	0.24
Reading	71	3.0129	1.07	1.22	0.29	0.12	0.33	0.17	0.37	0.20
Reading	72	2.8095	1.03	1.12	0.30	0.19	0.35	0.23	0.39	0.23
Reading	73	2.5537	0.96	1.02	0.36	0.30	0.44	0.32	0.47	0.38
Writing	74	2.7305	1.29	1.45	2.59	0.60	2.65	0.65	2.72	0.68
Writing	75	2.7640	1.73	1.91	2.68	0.61	2.81	0.64	2.90	0.68
Speaking	76	-1.2281	0.68	0.54	1.96	0.28	1.95	0.39	1.95	0.39
Speaking	77	-1.1747	0.84	0.84	1.95	0.35	1.94	0.40	1.93	0.40
Speaking	78	-1.0646	0.68	0.54	1.94	0.44	1.92	0.52	1.91	0.54
Speaking	79	-1.0997	0.64	0.48	1.95	0.42	1.94	0.46	1.93	0.49
Speaking	80	-0.8851	0.73	0.65	1.93	0.44	1.90	0.52	1.86	0.57
Speaking	81	0.1868	0.65	0.61	1.84	0.55	1.81	0.61	1.77	0.65
Speaking	82	0.3473	0.76	0.76	1.76	0.50	1.74	0.60	1.70	0.63
Speaking	83	0.1189	0.77	0.76	1.81	0.45	1.76	0.55	1.74	0.58
Speaking	84	0.2581	0.71	0.69	1.79	0.51	1.75	0.59	1.73	0.63
Speaking	85	0.1112	0.67	0.60	1.86	0.47	1.85	0.55	1.83	0.56
Speaking	86	0.8545	0.90	0.93	3.33	0.62	3.27	0.69	3.22	0.72
Speaking	87	0.9745	0.85	0.89	3.22	0.63	3.23	0.70	3.17	0.73
Speaking	88	0.1924	0.71	0.62	1.87	0.40	1.87	0.45	1.87	0.51
Speaking	89	0.3033	0.58	0.46	1.89	0.51	1.86	0.60	1.84	0.65
Speaking	90	0.7067	0.80	0.75	1.76	0.50	1.72	0.57	1.71	0.61
Speaking	91	0.3084	0.64	0.54	1.86	0.53	1.83	0.60	1.81	0.64
Speaking	92	0.2021	0.68	0.55	1.89	0.45	1.86	0.51	1.86	0.58

Table B4: Form A High School (Grades 9-12)

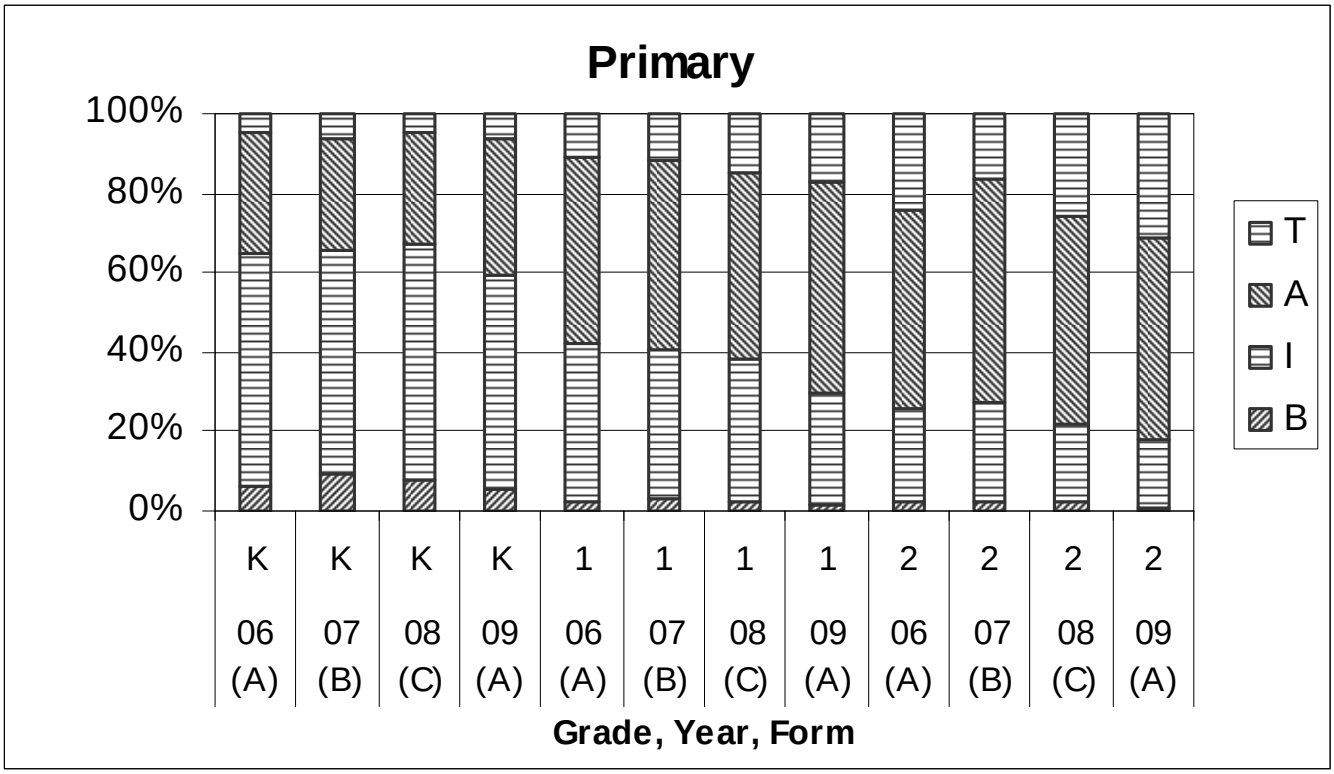
	High School	Grade 9	Grade 10	Grade 11	Grade 12
N-Count	10,823	3,749	2,923	2,287	1,864

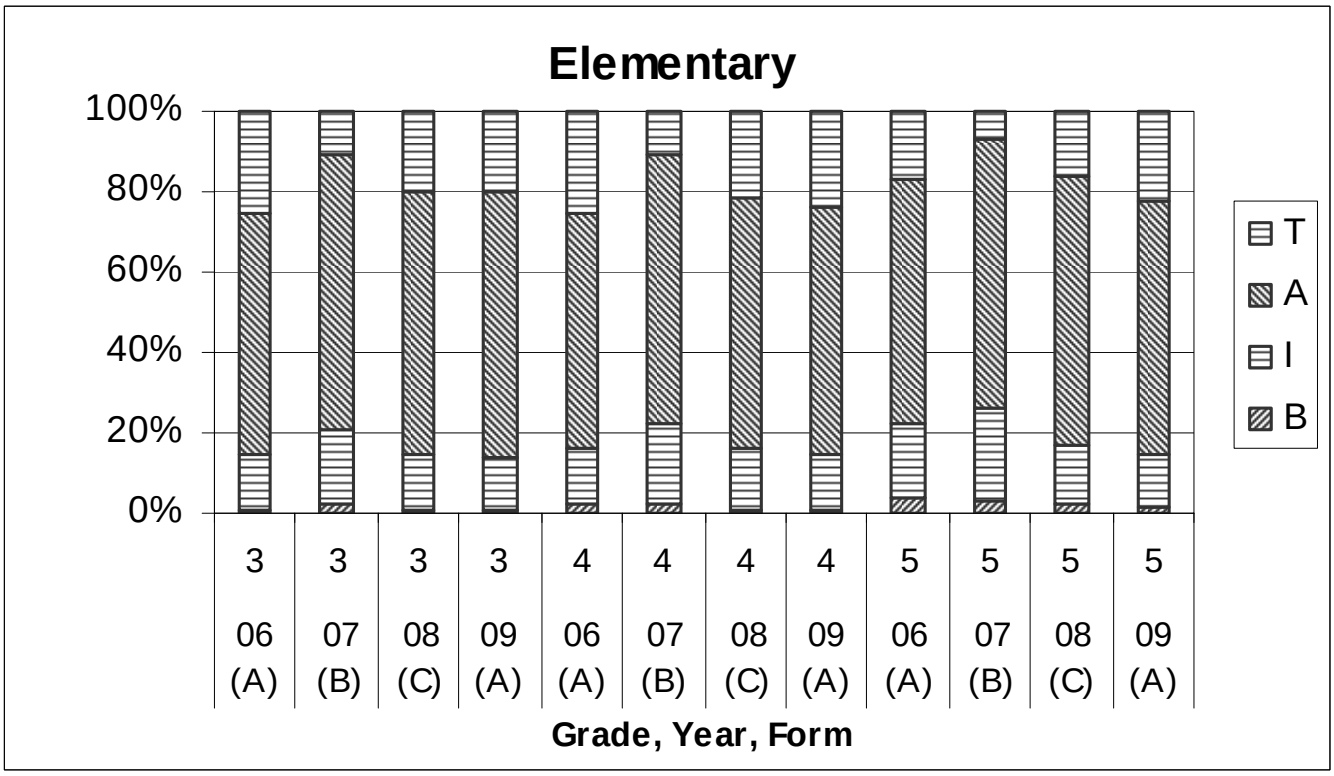
		High School			Grade 9		Grade 10		Grade 11		Grade 12	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Listening	1	-2.0921	1.02	1.13	0.99	0.21	0.98	0.20	0.99	0.24	0.99	0.21
Listening	2	-1.8107	0.98	1.09	0.98	0.26	0.98	0.24	0.99	0.25	0.99	0.22
Listening	3	-1.6080	0.93	0.77	0.98	0.29	0.98	0.27	0.99	0.28	0.99	0.23
Listening	4	-1.5477	0.92	0.80	0.98	0.28	0.98	0.25	0.97	0.20	0.97	0.16
Listening	5	-0.5205	0.96	0.84	0.91	0.35	0.92	0.38	0.95	0.32	0.95	0.28
Listening	6	1.3782	1.00	1.00	0.69	0.22	0.71	0.23	0.74	0.18	0.74	0.18
Listening	7	1.3603	1.02	1.05	0.52	0.35	0.49	0.36	0.52	0.37	0.54	0.33
Listening	8	0.8708	0.83	0.74	0.23	0.01	0.24	-0.01	0.24	0.05	0.23	0.05
Listening	9	1.7053	1.19	1.24	0.39	0.22	0.40	0.25	0.44	0.25	0.44	0.23
Listening	10	0.1481	0.86	0.79	0.37	0.11	0.36	0.10	0.36	0.14	0.37	0.09
Listening	11	1.2575	1.23	1.36	0.62	0.35	0.65	0.36	0.69	0.34	0.69	0.33
Listening	12	2.1602	1.06	1.11	0.40	0.23	0.41	0.23	0.46	0.28	0.45	0.26
Listening	13	3.5380	1.28	1.96	0.27	0.23	0.28	0.22	0.30	0.23	0.30	0.22
Listening	14	2.3439	1.14	1.21	0.50	0.24	0.52	0.24	0.54	0.27	0.53	0.27
Listening	15	2.9195	1.26	1.51	0.31	0.25	0.29	0.26	0.34	0.30	0.32	0.29
Listening	16	1.3782	1.18	1.24	0.69	0.42	0.74	0.40	0.76	0.43	0.78	0.39
Listening	17	2.7506	1.14	1.33	0.80	0.44	0.80	0.44	0.83	0.40	0.85	0.33
Listening	18	3.4437	1.07	1.34	0.83	0.57	0.83	0.57	0.87	0.52	0.90	0.43
Listening	19	2.3494	1.13	1.26	0.68	0.28	0.69	0.34	0.70	0.35	0.67	0.32
Listening	20	3.3125	1.04	1.26	0.87	0.44	0.89	0.45	0.91	0.42	0.92	0.32
Writing Conventions	21	-1.2675	0.81	0.51	0.97	0.36	0.97	0.36	0.97	0.34	0.98	0.32
Writing Conventions	22	0.4588	1.01	0.98	0.81	0.36	0.84	0.33	0.85	0.35	0.86	0.33
Writing Conventions	23	0.0112	0.96	0.79	0.88	0.47	0.88	0.43	0.89	0.41	0.89	0.31
Writing Conventions	24	0.8158	0.82	0.68	0.84	0.48	0.83	0.46	0.85	0.38	0.85	0.39
Writing Conventions	25	-1.2675	0.82	0.51	0.96	0.37	0.96	0.35	0.97	0.34	0.98	0.33
Writing Conventions	26	0.5531	0.84	0.74	0.84	0.42	0.87	0.35	0.87	0.35	0.88	0.30
Writing Conventions	27	1.2079	0.94	0.91	0.71	0.49	0.72	0.48	0.78	0.46	0.77	0.40
Writing Conventions	28	1.9071	1.04	1.05	0.54	0.32	0.59	0.34	0.67	0.35	0.68	0.32

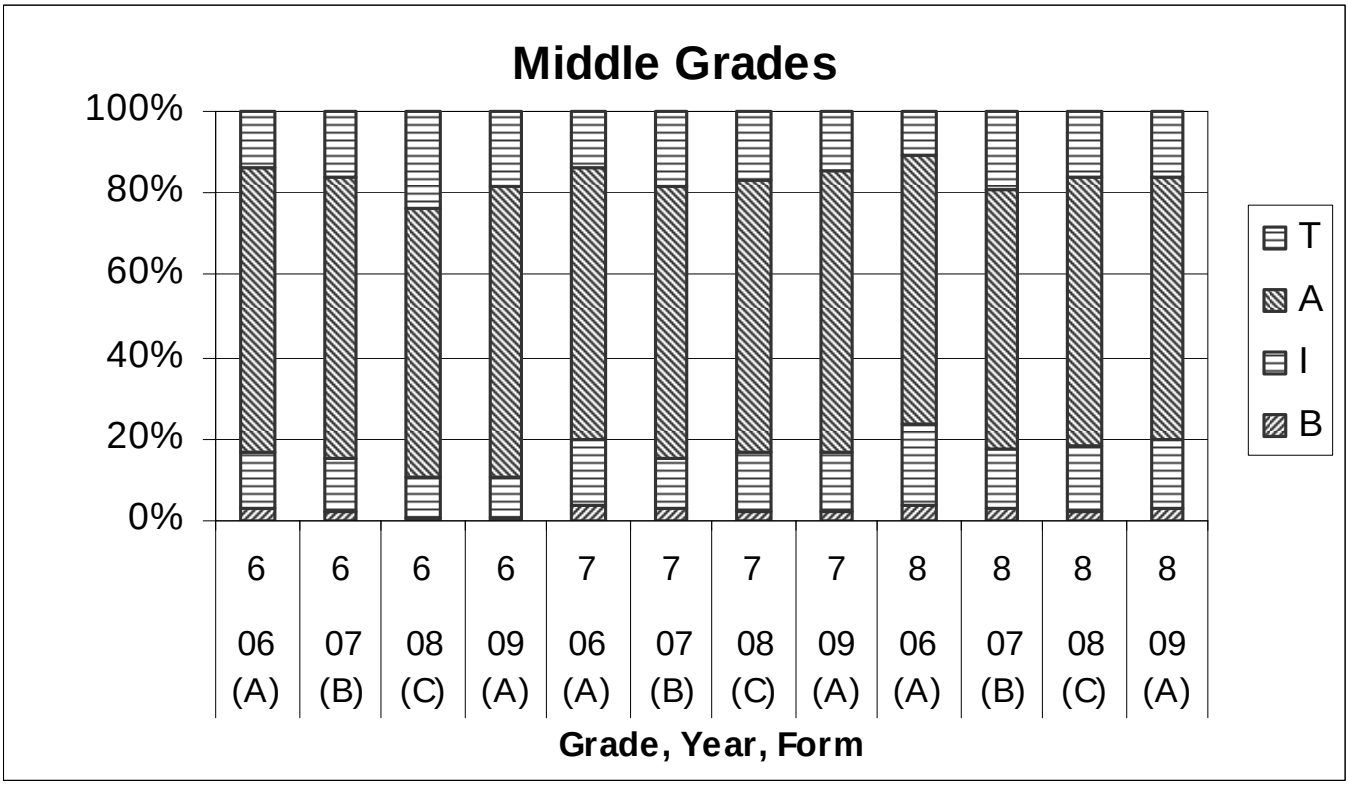
Modality	Item Sequence	High School			Grade 9		Grade 10		Grade 11		Grade 12	
		Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Writing Conventions	29	3.1137	1.09	1.20	0.40	0.24	0.36	0.25	0.39	0.23	0.41	0.21
Writing Conventions	30	2.5792	1.11	1.22	0.48	0.25	0.53	0.27	0.59	0.30	0.63	0.30
Writing Conventions	31	2.7004	1.07	1.15	0.38	0.25	0.44	0.29	0.51	0.31	0.52	0.27
Writing Conventions	32	1.0820	1.10	1.03	0.67	0.38	0.73	0.39	0.78	0.38	0.81	0.44
Writing Conventions	33	2.0692	1.16	1.19	0.50	0.25	0.57	0.28	0.63	0.31	0.63	0.26
Writing Conventions	34	4.9609	1.08	2.78	0.08	-0.04	0.09	0.00	0.13	0.07	0.13	0.13
Writing Conventions	35	1.6471	0.91	0.87	0.65	0.52	0.65	0.51	0.67	0.54	0.69	0.51
Writing Conventions	36	2.3208	1.08	1.13	0.48	0.29	0.50	0.29	0.54	0.32	0.54	0.26
Writing Conventions	37	1.3060	0.73	0.63	0.75	0.57	0.77	0.58	0.81	0.53	0.82	0.48
Writing Conventions	38	1.7511	0.95	0.93	0.71	0.42	0.67	0.41	0.70	0.37	0.70	0.38
Writing Conventions	39	-0.2454	0.88	0.59	0.91	0.45	0.92	0.43	0.94	0.42	0.95	0.35
Writing Conventions	40	1.4732	0.77	0.68	0.74	0.52	0.74	0.51	0.76	0.53	0.78	0.46
Writing Conventions	41	1.9558	1.02	1.05	0.65	0.36	0.64	0.32	0.65	0.33	0.65	0.34
Writing Conventions	42	1.8877	1.08	1.12	0.60	0.32	0.61	0.33	0.63	0.33	0.63	0.35
Writing Conventions	43	1.2079	1.00	0.98	0.76	0.36	0.77	0.33	0.80	0.33	0.81	0.29
Writing Conventions	44	3.0325	1.21	1.45	0.38	0.14	0.39	0.14	0.42	0.16	0.44	0.15
Reading	45	-1.6081	1.04	1.01	0.97	0.37	0.97	0.32	0.98	0.26	0.98	0.29
Reading	46	-0.0492	0.91	0.65	0.89	0.51	0.90	0.52	0.93	0.47	0.95	0.38
Reading	47	-2.4163	0.97	0.86	0.99	0.24	0.99	0.27	0.99	0.28	0.99	0.26
Reading	48	-0.8572	0.98	0.79	0.94	0.33	0.95	0.35	0.95	0.27	0.97	0.32
Reading	49	0.3265	0.95	0.94	0.87	0.29	0.88	0.29	0.91	0.31	0.90	0.27
Reading	50	0.7520	0.99	0.94	0.79	0.39	0.82	0.38	0.83	0.42	0.85	0.37
Reading	51	1.1070	0.79	0.63	0.76	0.59	0.74	0.63	0.81	0.59	0.83	0.53
Reading	52	1.8195	0.90	0.87	0.65	0.45	0.68	0.42	0.74	0.45	0.76	0.43
Reading	53	1.7305	0.80	0.73	0.63	0.56	0.62	0.57	0.68	0.57	0.71	0.54
Reading	54	1.1580	1.18	1.17	0.63	0.33	0.69	0.34	0.73	0.36	0.77	0.30
Reading	55	1.3060	0.92	0.85	0.69	0.43	0.74	0.46	0.79	0.44	0.82	0.42
Reading	56	1.8195	1.06	1.05	0.59	0.32	0.62	0.35	0.64	0.35	0.67	0.37
Reading	57	1.5900	1.02	0.99	0.67	0.31	0.72	0.33	0.77	0.33	0.78	0.35
Reading	58	1.6594	1.00	0.96	0.66	0.40	0.68	0.38	0.71	0.36	0.72	0.39
Reading	59	1.8940	0.98	0.98	0.57	0.43	0.63	0.43	0.67	0.44	0.71	0.38
Reading	60	2.0238	1.10	1.12	0.55	0.28	0.60	0.34	0.65	0.32	0.64	0.28
Reading	61	1.1789	0.81	0.69	0.72	0.56	0.75	0.57	0.79	0.55	0.82	0.48

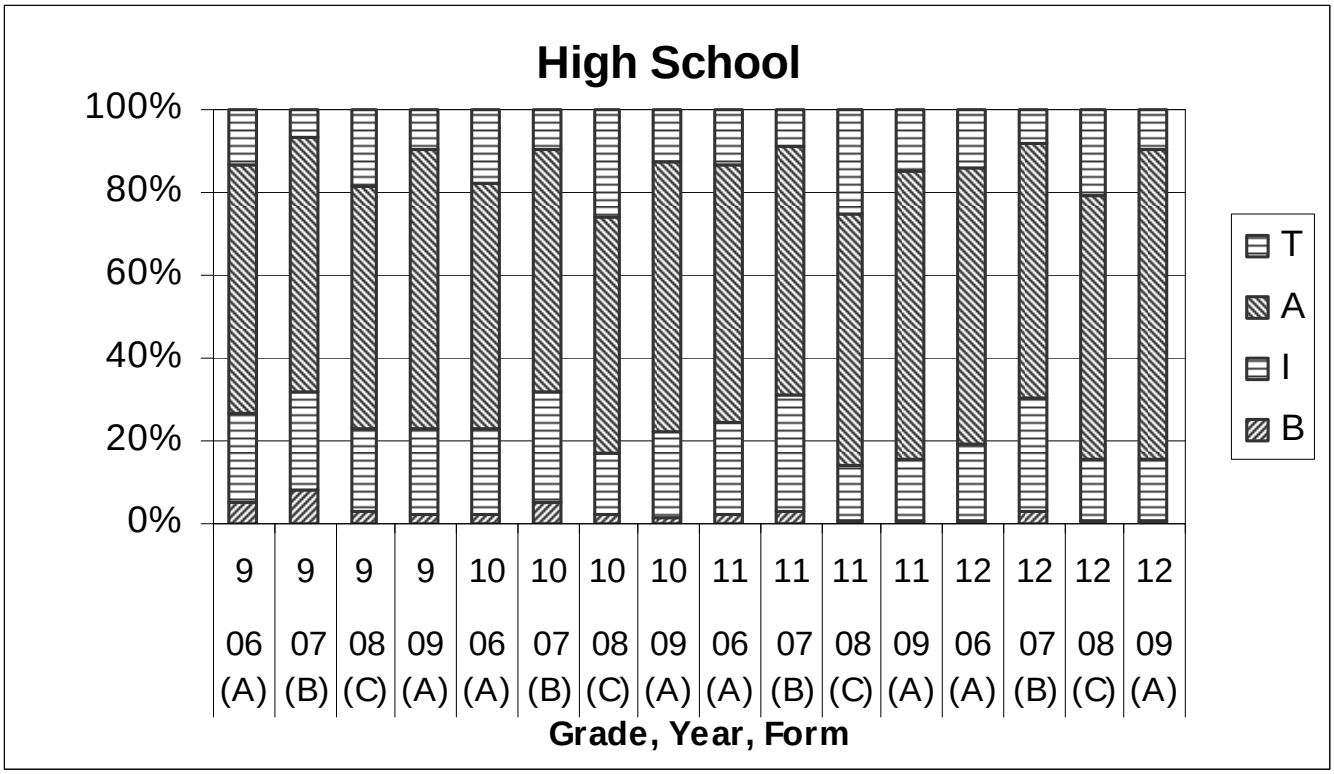
		High School			Grade 9		Grade 10		Grade 11		Grade 12	
Modality	Item Sequence	Difficulty	INFIT	OUTFIT	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation	Item Mean	Item-Total Correlation
Reading	62	0.4938	0.98	0.91	0.82	0.39	0.85	0.39	0.88	0.36	0.90	0.35
Reading	63	1.3317	0.89	0.82	0.70	0.48	0.72	0.49	0.77	0.48	0.78	0.44
Reading	64	2.7088	1.09	1.17	0.39	0.23	0.41	0.23	0.43	0.28	0.44	0.23
Reading	65	3.0638	1.24	1.51	0.40	0.08	0.40	0.08	0.40	0.12	0.39	0.15
Reading	66	1.4411	0.94	0.90	0.67	0.43	0.70	0.42	0.75	0.46	0.77	0.44
Reading	67	3.5196	1.07	1.41	0.23	0.09	0.29	0.15	0.32	0.18	0.34	0.21
Reading	68	2.9192	1.00	1.10	0.39	0.33	0.42	0.34	0.43	0.36	0.46	0.31
Reading	69	2.7549	1.23	1.41	0.43	0.13	0.45	0.13	0.47	0.12	0.46	0.15
Reading	70	3.4586	1.06	1.24	0.28	0.17	0.28	0.20	0.31	0.21	0.34	0.21
Reading	71	3.0207	1.06	1.17	0.35	0.23	0.40	0.26	0.42	0.27	0.45	0.26
Reading	72	3.4010	1.02	1.18	0.26	0.25	0.30	0.30	0.33	0.29	0.35	0.26
Reading	73	3.0818	1.16	1.42	0.34	0.17	0.38	0.18	0.41	0.21	0.40	0.20
Reading	74	2.7652	1.15	1.26	0.41	0.20	0.41	0.18	0.46	0.21	0.48	0.22
Reading	75	2.1663	0.96	0.95	0.54	0.41	0.57	0.45	0.66	0.39	0.69	0.36
Writing	76	2.2947	1.67	1.94	2.90	0.65	2.98	0.64	3.14	0.64	3.24	0.58
Writing	77	2.2965	1.66	1.98	2.87	0.69	2.96	0.68	3.15	0.64	3.25	0.61
Speaking	78	-1.2360	1.06	0.74	1.95	0.41	1.95	0.35	1.95	0.35	1.96	0.28
Speaking	79	-0.9508	0.93	0.78	1.87	0.48	1.87	0.40	1.87	0.39	1.87	0.40
Speaking	80	-0.4603	0.98	0.88	1.87	0.57	1.88	0.48	1.88	0.45	1.90	0.38
Speaking	81	-0.3893	0.90	0.79	1.81	0.57	1.81	0.53	1.82	0.46	1.81	0.48
Speaking	82	-0.2631	0.88	0.81	1.85	0.57	1.86	0.50	1.87	0.44	1.87	0.45
Speaking	83	0.6681	0.86	0.83	1.63	0.67	1.63	0.62	1.68	0.54	1.68	0.52
Speaking	84	1.2166	0.80	0.76	1.56	0.70	1.56	0.67	1.64	0.61	1.64	0.59
Speaking	85	0.7996	0.90	0.84	1.63	0.69	1.65	0.65	1.70	0.61	1.72	0.53
Speaking	86	0.6000	0.98	0.97	1.65	0.64	1.67	0.62	1.71	0.58	1.73	0.50
Speaking	87	0.9738	0.66	0.59	1.68	0.74	1.68	0.69	1.75	0.62	1.77	0.59
Speaking	88	1.2239	1.04	1.03	3.01	0.74	3.00	0.71	3.12	0.65	3.12	0.62
Speaking	89	1.3437	0.76	0.78	3.09	0.76	3.10	0.72	3.23	0.68	3.23	0.62
Speaking	90	0.1480	0.88	0.59	1.87	0.55	1.89	0.51	1.93	0.45	1.94	0.39
Speaking	91	-0.2359	0.98	0.72	1.83	0.62	1.87	0.55	1.90	0.48	1.91	0.44
Speaking	92	0.8823	0.84	0.71	1.70	0.69	1.70	0.67	1.76	0.60	1.78	0.55
Speaking	93	0.1225	0.89	0.70	1.82	0.62	1.85	0.53	1.88	0.50	1.90	0.42
Speaking	94	0.5095	0.87	0.73	1.76	0.71	1.77	0.66	1.82	0.60	1.86	0.49

APPENDIX C: WLPT-II ADDITIONAL STATISTICAL SUMMARIES

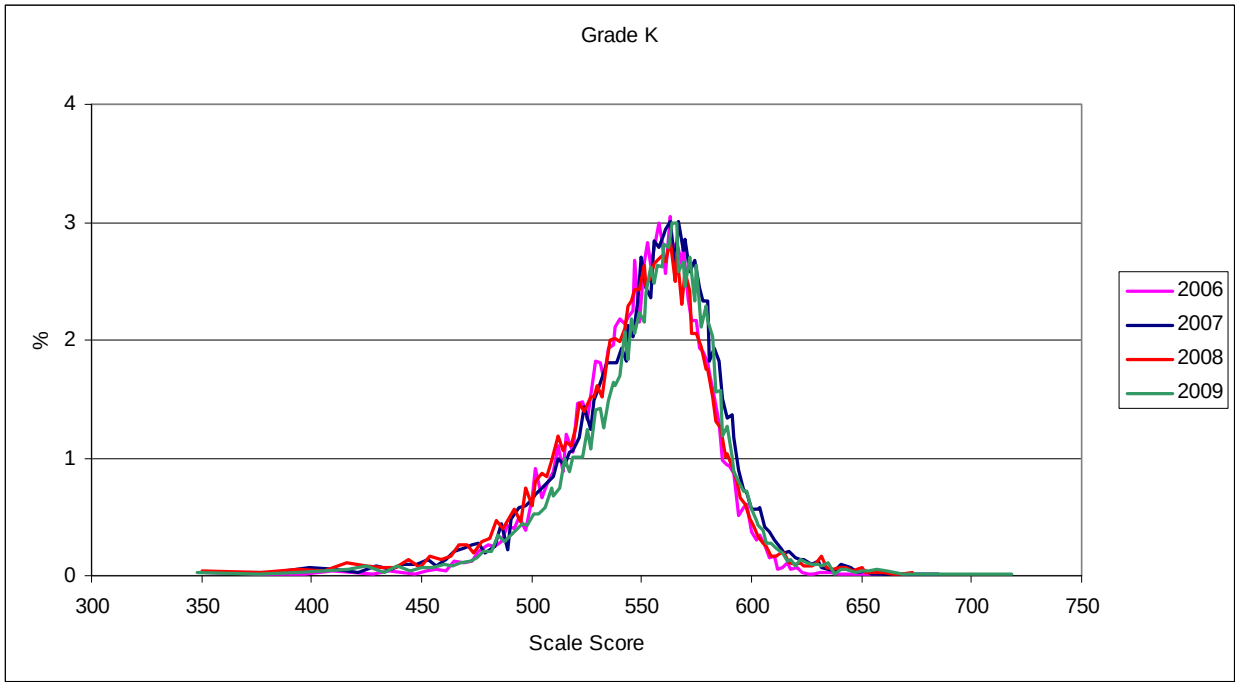


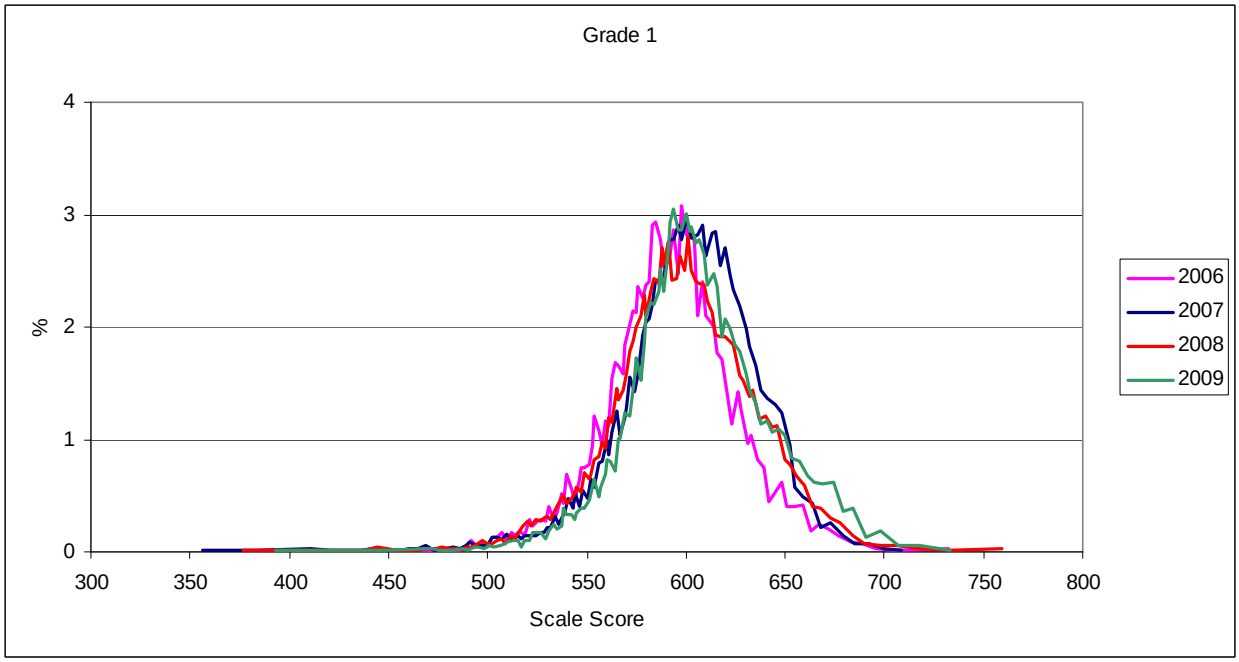



Middle Grades
100%
80%
60%
40%
20%
0%
T
A
I
B
6 06 (A)
6 07 (B)
6 08 (C)
6 09 (A)
7 06 (A)
7 07 (B)
7 08 (C)
7 09 (A)
8 06 (A)
8 07 (B)
8 08 (C)
8 09 (A)
Grade, Year, Form

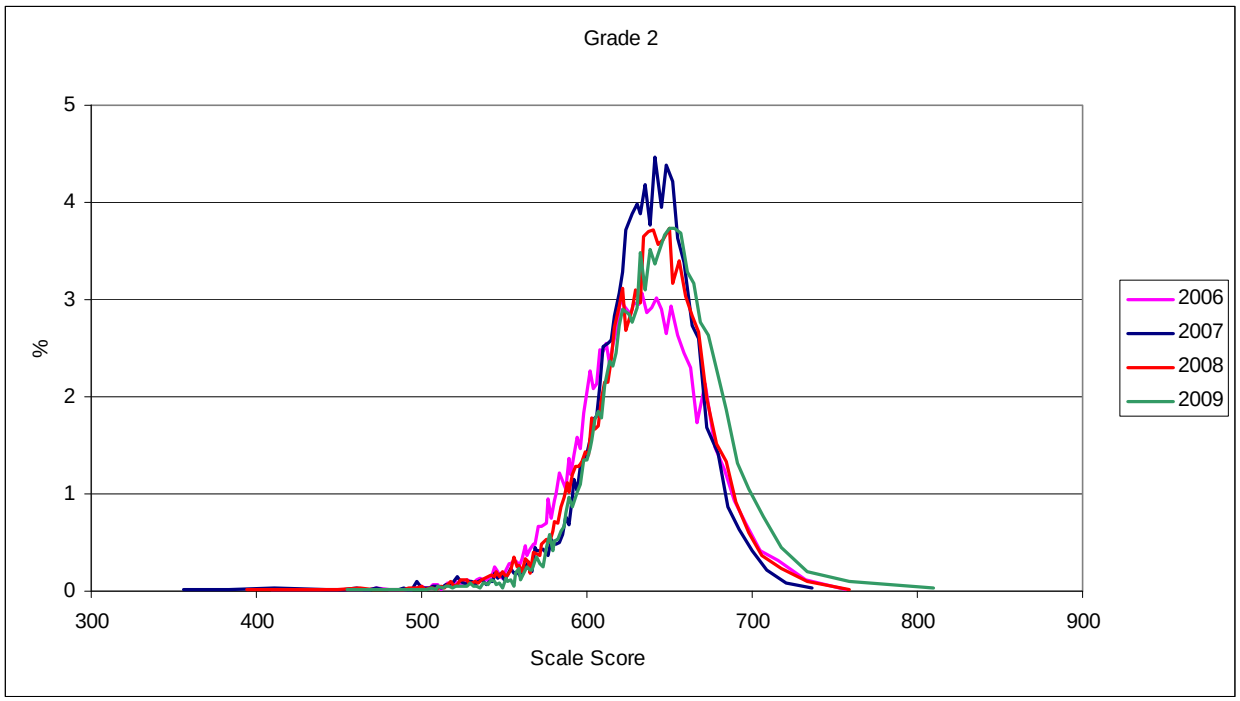

High School
100%
80%
60%
40%
20%
0%
T
A
I
B
9 06 (A)
9 07 (B)
9 08 (C)
9 09 (A)
10 06 (A)
10 07 (B)
10 08 (C)
10 09 (A)
11 06 (A)
11 07 (B)
11 08 (C)
11 09 (A)
12 06 (A)
12 07 (B)
12 08 (C)
12 09 (A)
Grade, Year, Form

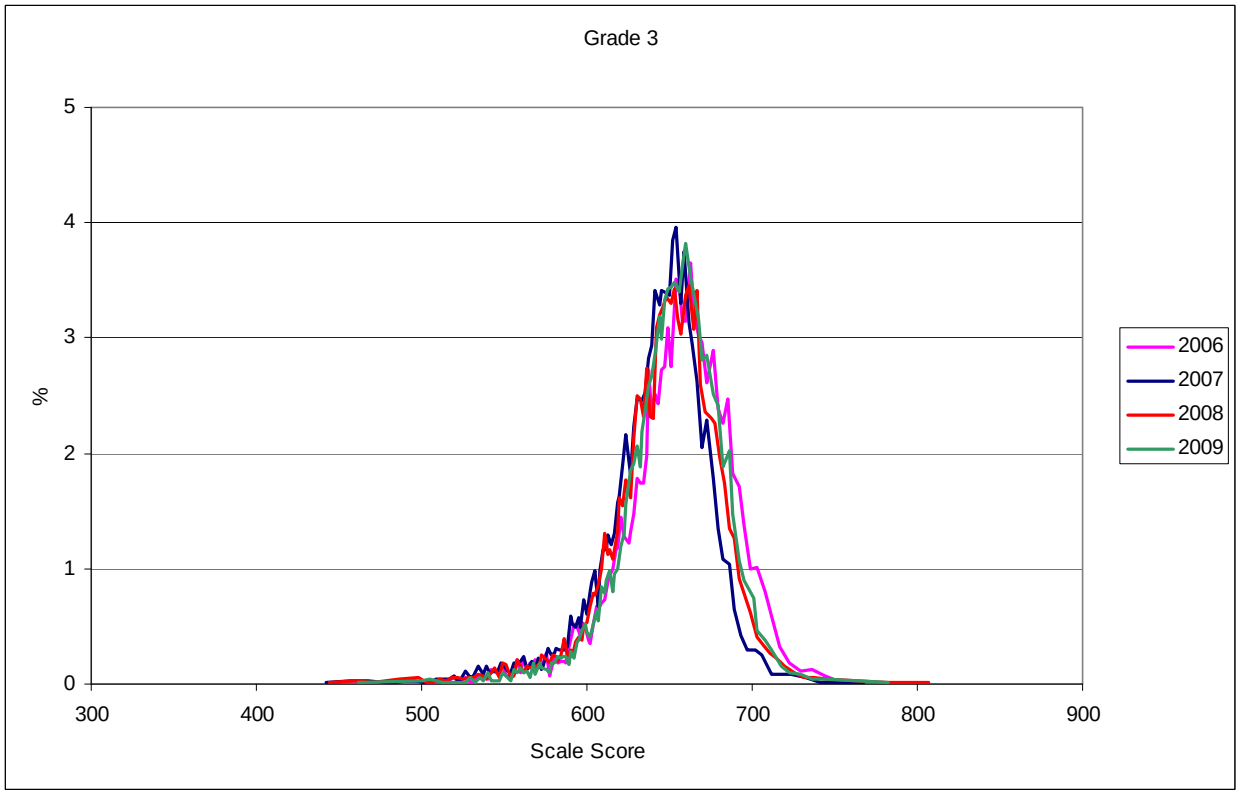
Scaled Score Summaries



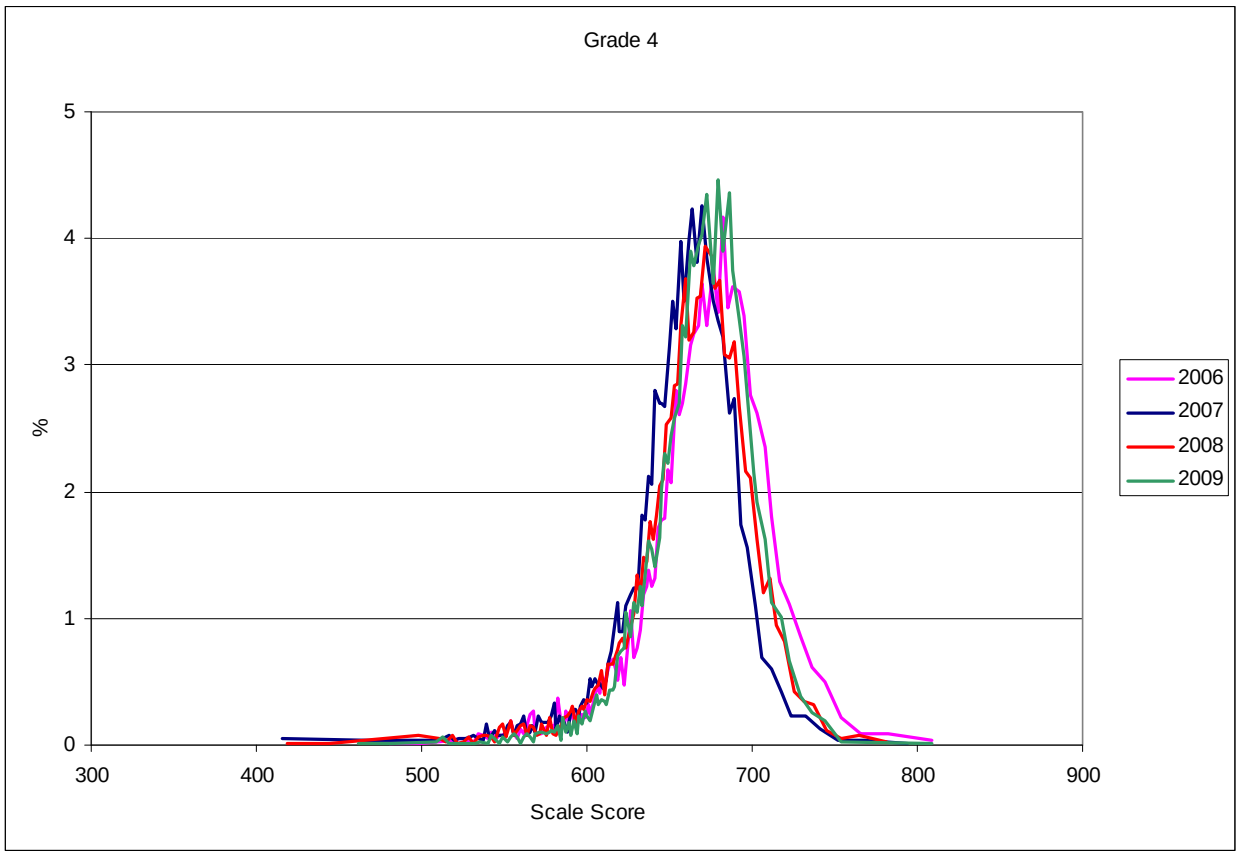


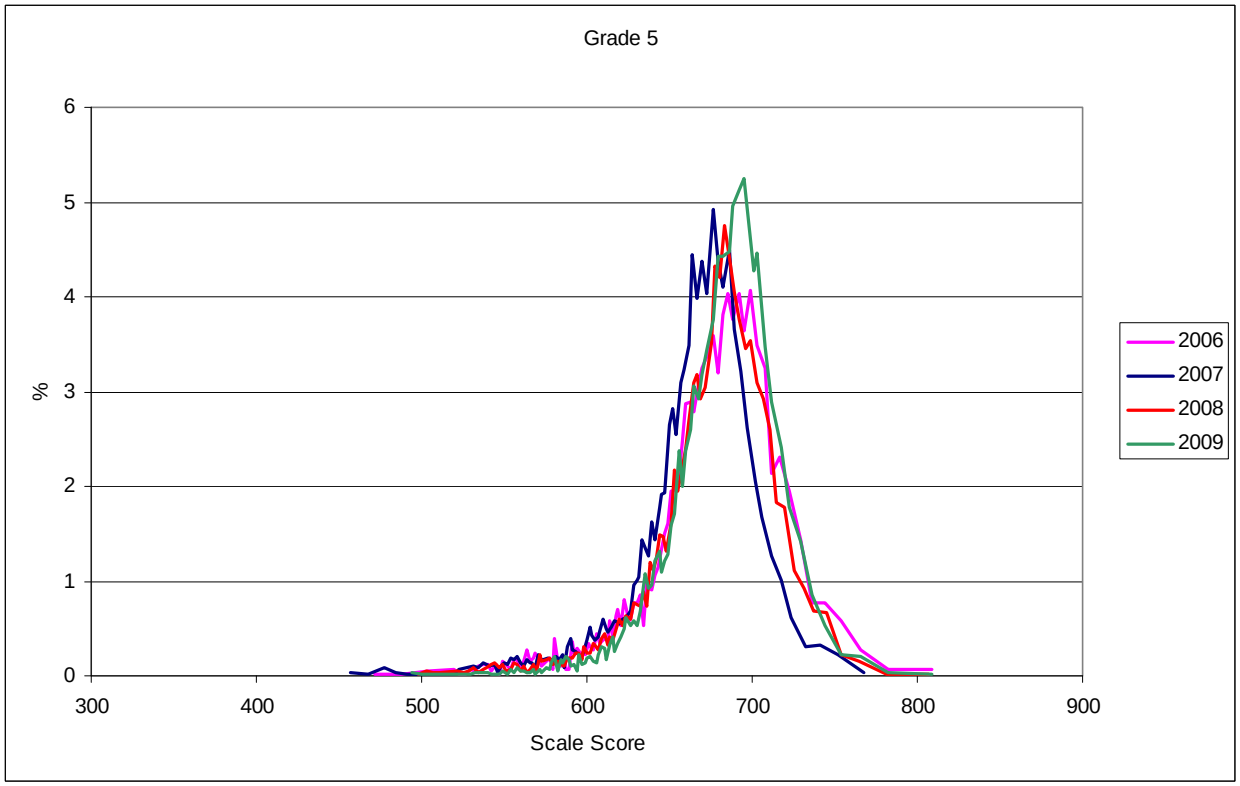
Scaled Score Summaries



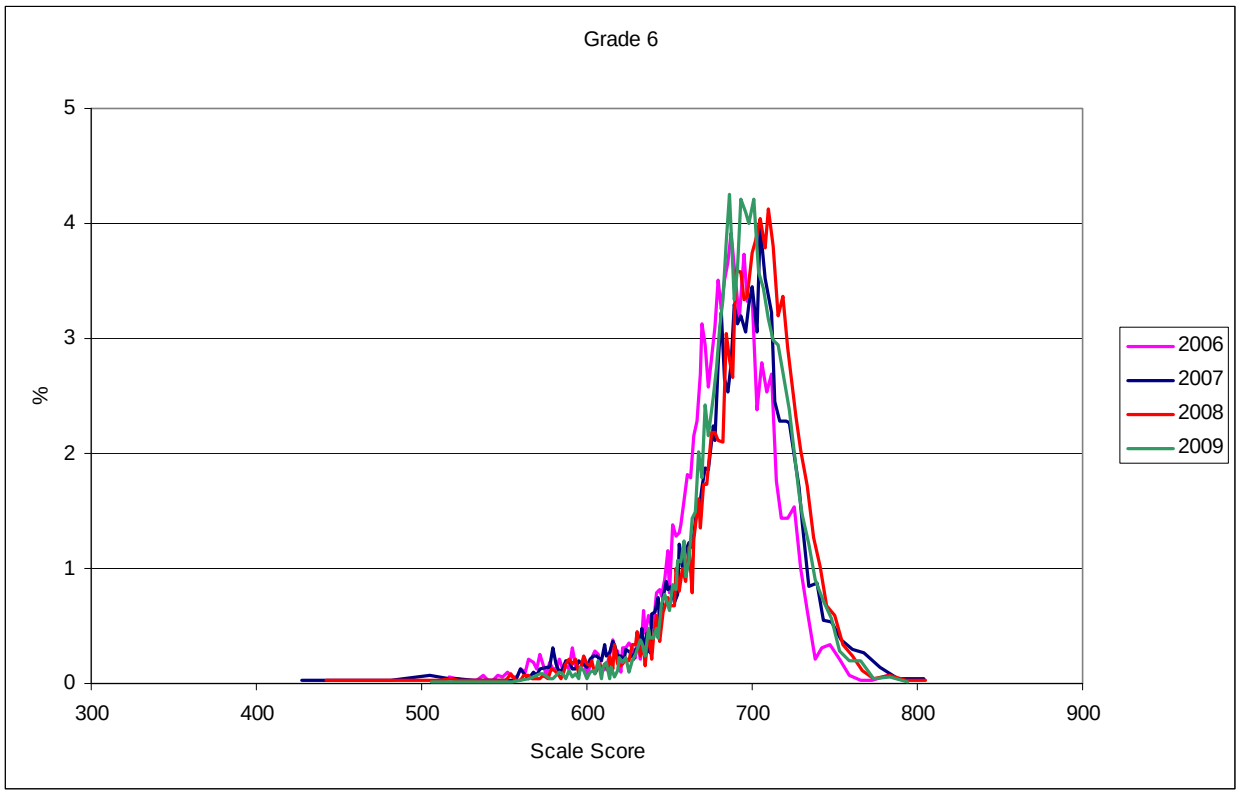


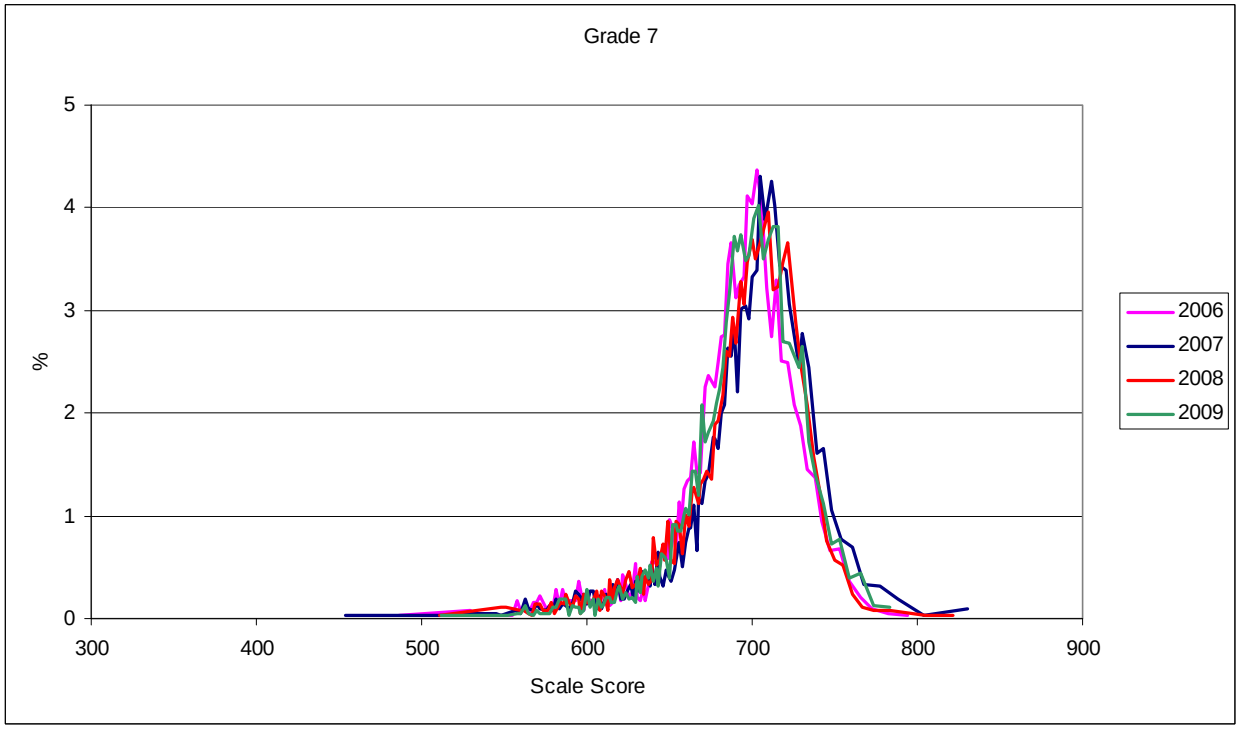
Scaled Score Summaries



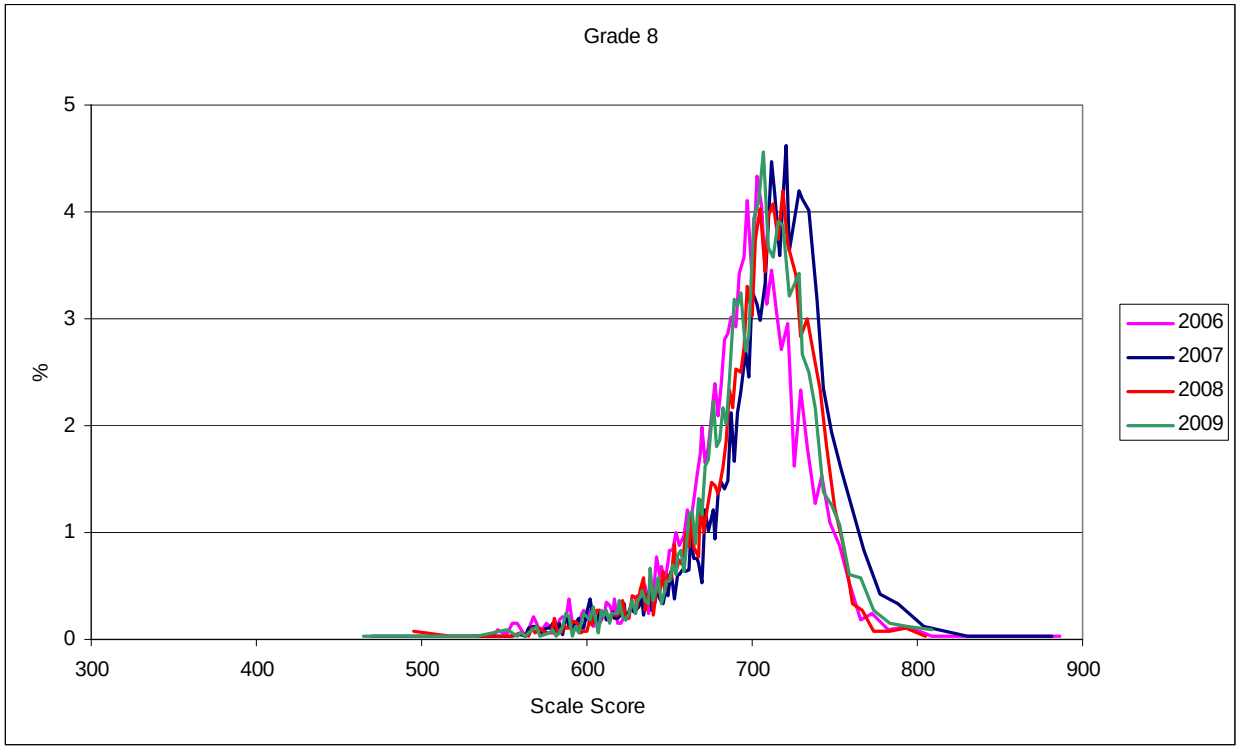


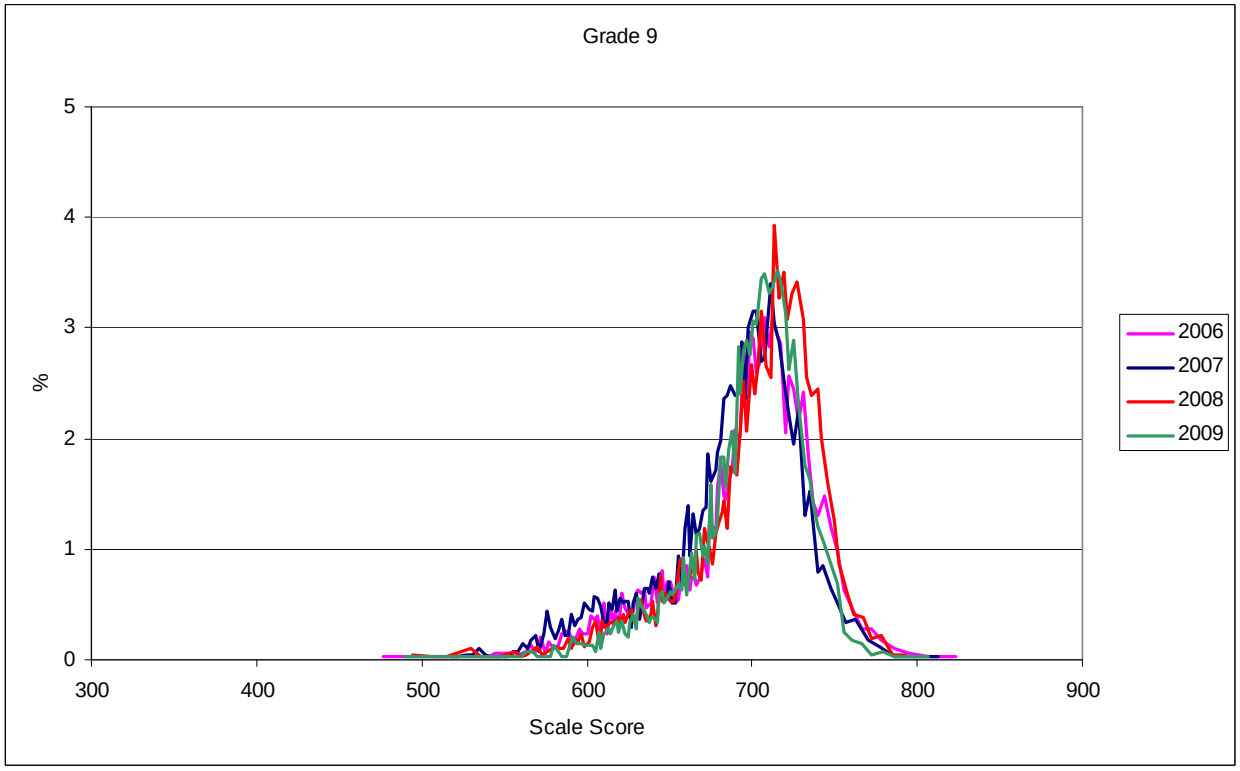
Scaled Score Summaries



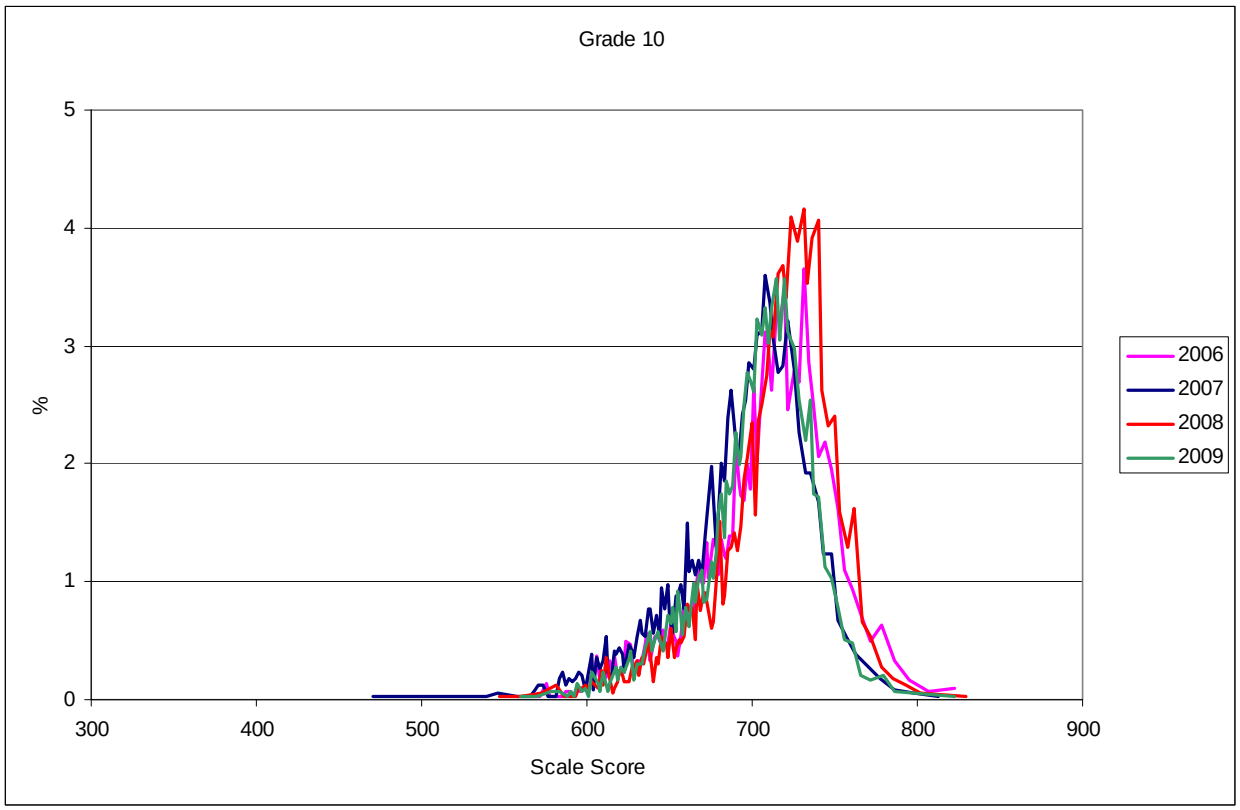


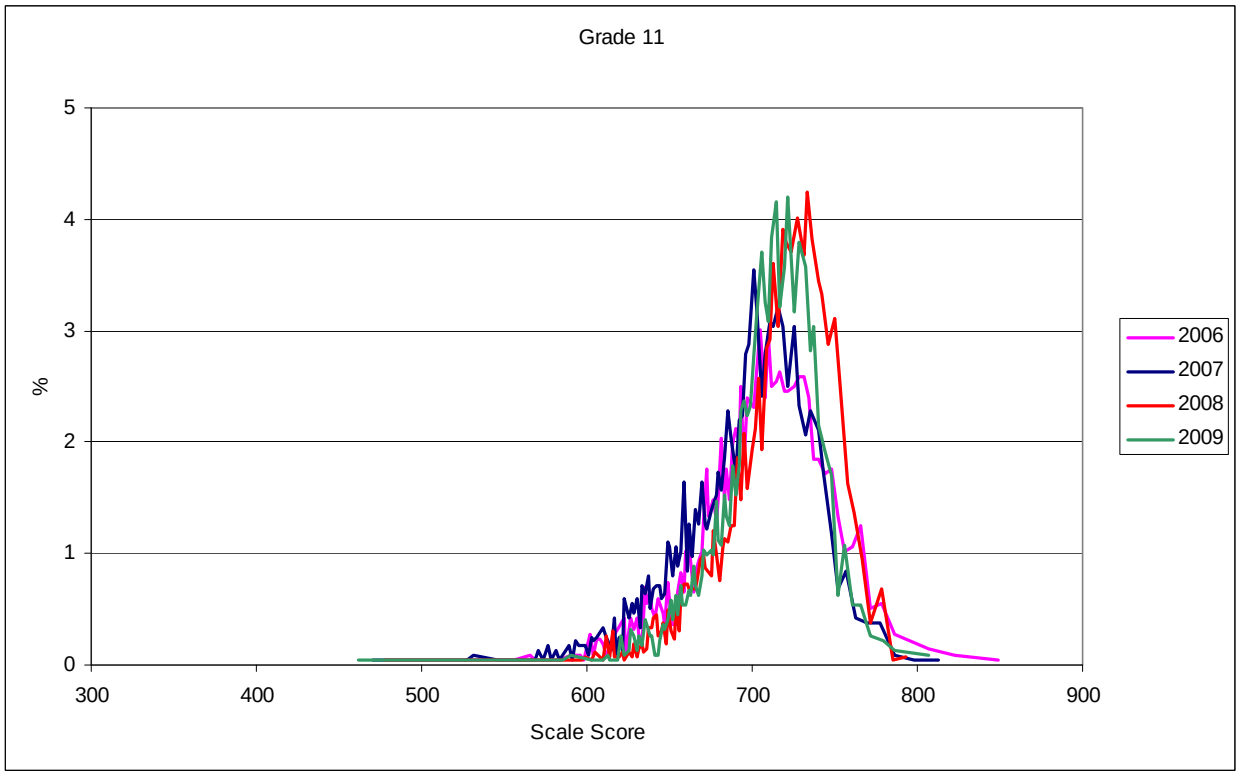
Scaled Score Summaries



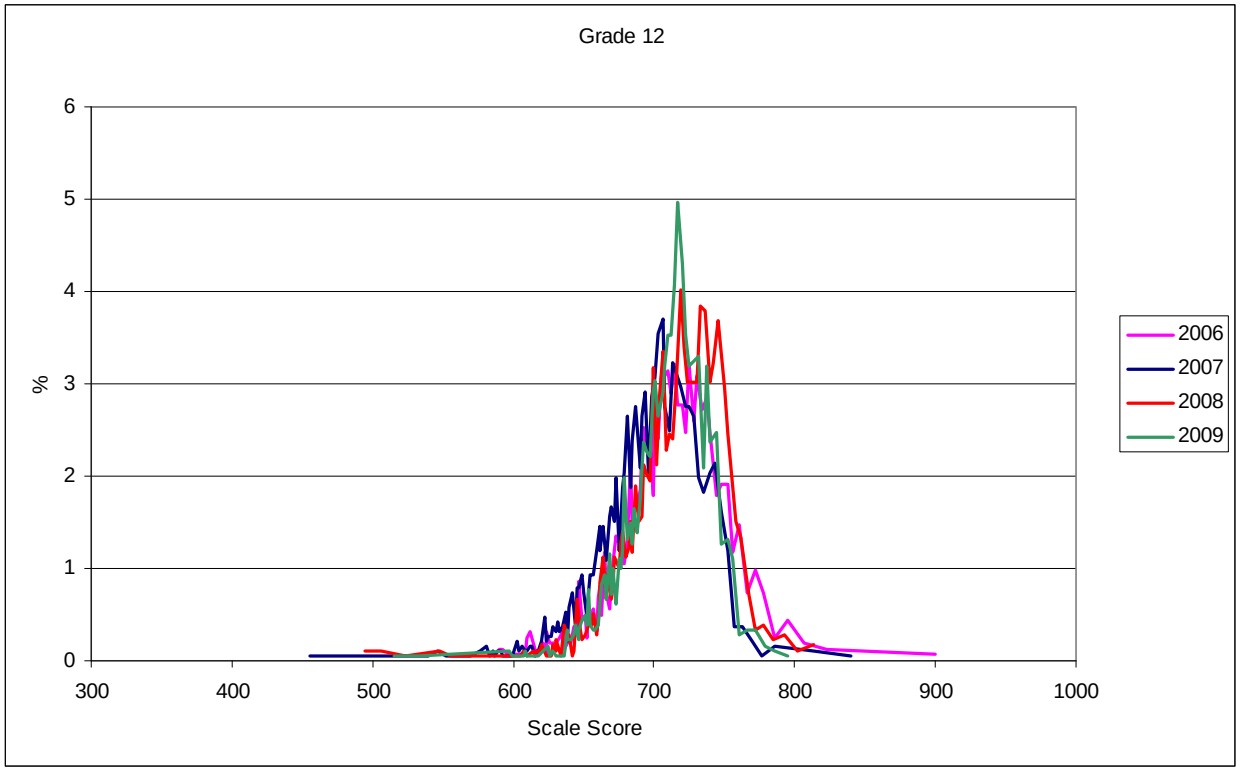


Scaled Score Summaries





Scaled Score Summaries



APPENDIX D: WLPT-II PROFICIENCY LEVEL CUT SCORES

Table D1: WLPT-II Overall Performance Level Cut Scores

	Scale Score			Theta		
Grade	I	A	T	I	A	T
K	509	566	594	-2.6240	-1.0485	-0.2746
1	527	586	627	-2.1265	-0.4957	0.6376
2	544	603	650	-1.6566	-0.0258	1.2733
3	559	619	669	-1.2420	0.4164	1.7984
4	572	633	686	-0.8827	0.8034	2.2683
5	584	644	701	-0.5510	1.1074	2.6829
6	594	654	712	-0.2746	1.3838	2.9870
7	602	662	721	-0.0535	1.6050	3.2357
8	608	668	728	0.1124	1.7708	3.4292
9	613	672	731	0.2506	1.8814	3.5121
10	616	675	732	0.3335	1.9643	3.5398
11	617	675	735	0.3611	1.9643	3.6227
12	617	678	740	0.3611	2.0472	3.7609

Note. I – Intermediate, A – Advanced, T – Transitional

Table D2: Applied 2009 WLPT-II (FORM A) Overall Performance Level Cut Scores

	Raw Score			Scale Score			Theta		
Grade	I	A	T	I	A	T	I	A	T
K	28	59	76	509	566	594	-2.6465	-1.0651	-0.2699
1	37	71	92	527	586	627	-2.1188	-0.5132	0.6438
2	46	81	100	544	603	650	-1.6637	-0.0121	1.2678
3	29	63	88	560	619	670	-1.2186	0.4202	1.8334
4	35	71	93	572	633	686	-0.9029	0.815	2.238
5	42	76	97	584	644	701	-0.5614	1.0799	2.6338
6	39	74	101	595	654	713	-0.2573	1.3693	3.0023
7	43	79	104	602	662	723	-0.0578	1.6131	3.2776
8	47	82	105	609	668	728	0.1332	1.7662	3.3783
9	39	76	104	613	673	732	0.229	1.9013	3.5033
10	41	77	104	616	675	732	0.3292	1.9461	3.5033
11	42	77	105	618	675	735	0.3783	1.9461	3.5881
12	42	79	107	618	678	740	0.3783	2.0369	3.7704

Note. I – Intermediate, A – Advanced, T – Transitional

APPENDIX E: WLPT-II SUMMARY STATISTICS FOR THE MAY ADMINISTRATION

Table E1: Descriptive Statistics of the WLPT-II Form C Scale Score (SS) by Grade and Modality

Grade	Modality	N Items	Max SS[a]	Max SS[b]	Min SS[c]	*N*	Mean	SD
K	Composite [d]	83	810	656	377	203	562.80	33.46
	Listening	20	718	718	314	203	582.72	51.95
	Reading	24	776	667	424	203	532.71	59.81
	Speaking	17	737	737	371	203	587.61	55.11
	Writing	22	776	653	418	203	540.95	42.33
	Comprehension [e]	44	783	666	313	203	558.12	39.95
	Social [f]	37	754	703	308	203	583.13	47.61
	Academic [g]	46	801	654	410	203	540.73	43.43
	Productive [h]	24	772	675	424	203	571.27	38.94
1	Composite [d]	83	810	718	416	191	578.98	65.24
	Listening	20	718	718	314	191	568.68	115.34
	Reading	24	776	776	424	191	566.31	78.47
	Speaking	17	737	737	371	191	580.90	104.99
	Writing	22	776	723	450	191	592.12	48.81
	Comprehension [e]	44	783	731	313	191	556.32	109.23
	Social [f]	37	754	754	308	191	566.57	118.16
	Academic [g]	46	801	724	459	191	585.48	49.85
	Productive [h]	24	772	772	460	191	587.59	57.42
2	Composite [d]	83	810	759	463	163	619.81	42.26
	Listening	20	718	718	314	163	600.12	82.43
	Reading	24	776	776	424	163	615.23	70.11
	Speaking	17	737	737	371	163	631.20	56.47
	Writing	22	776	776	519	163	626.71	48.85
	Comprehension [e]	44	783	783	313	163	606.22	85.22
	Social [f]	37	754	703	308	163	617.31	48.41
	Academic [g]	46	801	801	515	163	622.85	48.01
	Productive [h]	24	772	772	474	163	624.20	44.47
3	Composite [d]	82	857	703	560	116	644.87	33.27
	Listening	20	792	792	414	116	633.77	72.09
	Reading	23	826	774	430	116	637.22	68.86
	Speaking	17	765	765	511	116	658.97	51.84
	Writing	22	817	817	533	116	645.78	44.66
	Comprehension [e]	43	838	732	395	116	632.97	74.26
	Social [f]	37	807	711	537	116	645.72	34.04
	Academic [g]	45	847	732	504	116	643.24	43.92
	Productive [h]	19	799	799	515	116	652.88	42.25
4	Composite [d]	82	857	725	572	101	662.28	34.43
	Listening	20	792	739	414	101	643.96	70.62
	Reading	23	826	774	430	101	651.08	71.38
	Speaking	17	765	765	610	101	681.96	44.04
	Writing	22	817	817	556	101	667.39	48.53
	Comprehension [e]	43	838	744	395	101	645.67	74.56
	Social [f]	37	807	728	581	101	662.70	31.94
	Academic [g]	45	847	769	527	101	661.24	47.58
	Productive [h]	19	799	799	618	101	680.64	40.48

[a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Table E1: Descriptive Statistics of the WLPT-II Form C Scale Score (SS) by Grade and Modality (Continued)

Grade	Modality	N Items	Max SS[a]	Max SS[b]	Min SS[c]	*N*	Mean	SD
	Composite [d]	82	857	725	544	79	662.95	39.77
	Listening	20	792	739	414	79	640.65	90.58
	Reading	23	826	746	430	79	646.51	84.31
	Speaking	17	765	765	567	79	680.72	53.43
5	Writing	22	817	765	546	79	674.62	54.66
	Comprehension [e]	43	838	744	395	79	640.11	93.16
	Social [f]	37	807	755	548	79	660.53	36.99
	Academic [g]	45	847	742	517	79	663.33	52.67
	Productive [h]	19	799	799	580	79	684.23	50.25
	Composite [d]	91	900	755	602	88	684.13	35.70
	Listening	20	829	747	443	88	655.76	92.56
	Reading	28	860	748	445	88	668.77	71.50
	Speaking	17	795	795	587	88	721.49	53.04
6	Writing	26	875	817	563	88	689.05	52.91
	Comprehension [e]	48	873	748	418	88	666.03	70.29
	Social [f]	37	841	763	596	88	689.26	35.31
	Academic [g]	54	894	760	538	88	680.26	48.82
	Productive [h]	19	868	868	593	88	720.94	54.63
	Composite [d]	91	900	761	614	68	690.75	31.40
	Listening	20	829	776	443	68	652.46	98.56
	Reading	28	860	779	621	68	691.13	38.79
	Speaking	17	795	795	587	68	717.06	59.33
7	Writing	26	875	817	619	68	711.75	43.87
	Comprehension [e]	48	873	766	589	68	680.99	39.23
	Social [f]	37	841	790	599	68	686.97	42.36
	Academic [g]	54	894	760	634	68	697.75	33.01
	Productive [h]	19	868	868	590	68	731.18	67.96
	Composite [d]	91	900	774	589	66	685.52	47.32
	Listening	20	829	776	443	66	651.89	105.77
	Reading	28	860	779	445	66	656.42	98.88
	Speaking	17	795	795	616	66	735.20	54.42
8	Writing	26	875	817	523	66	686.50	69.35
	Comprehension [e]	48	873	766	418	66	650.62	105.11
	Social [f]	37	841	790	616	66	695.38	42.40
	Academic [g]	54	894	794	498	66	674.86	68.61
	Productive [h]	19	868	868	618	66	745.86	65.32

[a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Table E1: Descriptive Statistics of the WLPT-II Form C Scale Score (SS) by Grade and Modality (Continued)

Grade	Modality	N Items	Max SS[a]	Max SS[b]	Min SS[c]	*N*	Mean	SD
9	Composite [d]	91	900	778	515	110	698.07	52.82
	Listening	20	848	848	487	110	694.35	68.70
	Reading	28	866	786	432	110	678.86	78.25
	Speaking	17	806	806	451	110	724.53	88.39
	Writing	26	873	791	567	110	707.48	52.43
	Comprehension [e]	48	883	777	424	110	682.44	79.11
	Social [f]	37	858	858	440	110	701.24	72.16
	Academic [g]	54	895	770	529	110	695.04	52.91
	Productive [h]	19	851	851	541	110	738.02	76.93
10	Composite [d]	91	900	793	607	85	705.06	38.06
	Listening	20	848	795	487	85	701.28	55.09
	Reading	28	866	786	432	85	692.29	62.25
	Speaking	17	806	806	544	85	732.89	65.18
	Writing	26	873	791	567	85	711.98	48.18
	Comprehension [e]	48	883	789	424	85	694.20	59.60
	Social [f]	37	858	806	590	85	709.79	42.35
	Academic [g]	54	895	788	551	85	702.52	45.37
	Productive [h]	19	851	851	541	85	745.36	73.61
11	Composite [d]	91	900	793	494	65	694.83	65.74
	Listening	20	848	795	487	65	687.48	73.06
	Reading	28	866	814	432	65	691.11	94.83
	Speaking	17	806	806	451	65	707.54	107.21
	Writing	26	873	791	553	65	709.29	57.34
	Comprehension [e]	48	883	805	424	65	685.09	90.82
	Social [f]	37	858	806	440	65	684.23	89.26
	Academic [g]	54	895	788	515	65	702.69	62.50
	Productive [h]	19	851	851	541	65	730.15	90.30
12	Composite [d]	91	900	785	515	42	703.07	64.10
	Listening	20	848	795	487	42	694.00	74.99
	Reading	28	866	866	432	42	694.38	94.43
	Speaking	17	806	806	451	42	724.10	104.25
	Writing	26	873	791	579	42	714.29	48.82
	Comprehension [e]	48	883	789	424	42	689.36	91.52
	Social [f]	37	858	806	440	42	695.93	90.16
	Academic [g]	54	895	817	541	42	707.12	58.00
	Productive [h]	19	851	851	560	42	743.50	80.50

[a] Maximum Scale Score possible

[b] Maximum Scale Score observed

[c] Minimum Scale Score observed

[d] Composite score is based on Listening, Reading, Speaking, and Writing subtest items

[e] Comprehension score is based on Listening and Reading subtest items

[f] Social score is based on Listening and Speaking subtest items

[g] Academic score is based on Writing and Reading subtest items

[h] Productive score is based on Writing CR and Speaking subtest items. Sample size (*N*) is different for Productive as students who did not take any Writing CR items did not receive a Productive score.

Table E2: Percentage of Students in Each Proficiency Level by Grade for Form C

Grade	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional
K	3	51	31	15
1	18	24	35	24
2	5	26	44	26
3	0	17	61	22
4	0	18	57	25
5	5	16	67	11
6	0	17	61	22
7	0	15	69	16
8	8	23	47	23
9	7	12	54	27
10	1	24	48	27
11	12	8	48	32
12	10	5	62	24

Note. The percentages within a grade may not sum to 100 due to rounding error.