Washington Language Proficiency Test – II (WLPT-II)

Form C
Technical Report
2007 – 2008 School Year

Dr. Terry Bergeson
State Superintendent of
Public Instruction

Prepared by
Pearson

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

August 31, 2008

TABLE OF CONTENTS

OVERVIEW OF THE REPORT

The Washington Language Proficiency Test - II (WLPT-II) Technical Report for the 2007 – 2008 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 2008 WLPT-II. Results of the calibration, equating, and scaling are also presented. More detailed and comprehensive descriptions of the 2008 WLPT-II equating are available in the separate technical document, *Washington Language Proficiency Test – II Equating Study Report (2007 – 2008 School Year)*.

The **Summary of Operational Test Results** section presents the raw score to scale score tables by grade and the percentage of students at each performance level for the spring 2008 administration. This section also provides summary information for the May 2008 assessment (Wave 2) that used Form B 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.

1.2. Rationale and Purpose

In addition to the NCLB mandated assessment of K–12 ELLs, the legislation further requires that the assessment align to the State ELD standards. 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/default.aspx.

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 (FORM C)

2.1. Overview

The 2008 WLPT-II operational test (Form C) 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.

The test was developed using Stanford English Language Proficiency Test (SELP) Form C items, as well as augmented items developed by Washington State teachers. A data review committee reviewed these augmented items after the 2008 administration. Based on the performance of each augmented item with the Washington ELL students, the committee determined which augmented items should be included for operational scoring.

2.2. Test Specifications by Modality and Grade Span for WLPT-II Form C

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. Before the 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 2008 administration and decided to drop 1 CR Writing item (2-points) from Primary, 1 MC Reading item from Elementary, 1 MC Reading item from Middle Grades, and 3 MC Reading items from High School.

The final WLPT-II Form C has a total of 83 items for Primary grades, 82 for Elementary, 91 for Middle Grades, and 91 for High School. Speaking has 17 CR items in each grade span. There are 20 MC Listening items for each grade span, while Reading has 23 to 28 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.
- Three SR items in which students must copy printed text – a letter, a word, and a sentence – and three dictation SR items (Primary only).
- Two ERs, responding to graphics-based prompts.

For Elementary through High School, the number of Writing Conventions MC items in Form C 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, 5 SR items, and 2 ER prompts.

The test design for the 2008 WLPT-II Form C is shown in Table 2.1, excluding the dropped items. 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 2.1: 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	5	2	83
Elementary: 3-5	17	20	23	5	20	0	2	82
Middle Grades: 6-8	17	20	28	5	24	0	2	91
High School: 9-12	17	20	28	5	24	0	2	91

Table 2.2 provides the maximum number of points by modality and grade span, excluding the dropped items. The percentage of total points for Comprehension ranged from 39 percent to 41 percent.

Table 2.2: 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	8	8	113
Elementary: 3-5	38	20	23	5	20	0	8	109
Middle Grades: 6-8	38	20	28	5	24	0	8	118
High School: 9-12	38	20	28	5	24	0	8	118

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 2008 WLPT-II (Form C) was developed from SELP Form C.

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 use in forthcoming forms (including Forms A and B), 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 proposed final version of Form C 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.

2.7. Data Review

In April 2008, a data review committee consisting of Washington ESL professionals reviewed each augmented item on Form C and the associated item statistics. The committee decided not to use 1 Writing item worth 2-points from Primary, 1 Reading item from Elementary, 1 Reading item from Middle Grades, and 3 Writing items from High School. These items were excluded from the equating study, reported results, and all subsequent statistical analyses.

The item statistics used at the 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.

A description of the DIF method used follows, as well as a summary of the 2008 DIF results for Form C items used in the Data Review.

2.7. Differential Item Functioning

This section provides information about Differential Item Functioning (DIF) analyses for the 2008 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.

2.7.1. 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 2.3 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 2.3: 2 × *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).

2.7.2. 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.

2.7.3. 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 2.4 below.

Table 2.4: DIF Classification for OE Items

Category	Description	Criterion[1]
AA	No DIF	Non-significant Mantel χ^2 or Significant Mantel χ^2 and |*SMD*/SD| ≤ .17
BB	Weak DIF	Significant Mantel χ^2 and .17 < |*SMD*/SD| ≤ .25
CC	Strong DIF	Significant Mantel χ^2 and .25 < |*SMD*/SD|

[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:

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

2.7.4. 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 $D = -2.35 \cdot \beta_{M-H}$.

2.7.5. 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 2.5 below.

Table 2.5: DIF Classification for MC Items

Category	Description	Criterion
A	No DIF	Non-significant $M\text{-}H\ \chi^2$ or $\lvert D\rvert < 1.0$
B	Weak DIF	Significant $M\text{-}H\ \chi^2$ and $\lvert D\rvert < 1.5$ or Non-significant $M\text{-}H\ \chi^2$ and $\lvert D\rvert \geq 1.0$
C	Strong DIF	Significant $M\text{-}H\ \chi^2$ and $\lvert D\rvert \geq 1.5$

2.7.6. 2008 Form C DIF results

Table 2.6 summarizes the DIF results for augmented Form C items used at the 2008 Data Review. Of the 14 Primary items reviewed, 0 showed weak DIF (MC category B or CR category BB) and 0 showed strong DIF (MC category C or CR category CC). Two Writing CR items in Elementary showed weak DIF, favoring Females over Males. For the Middle School, two Reading MC items showed weak DIF and two Writing CR items showed strong DIF, all in favor of Females. Finally, for High School, there were two Writing items that showed weak DIF, also in favor of Females.

Table 2.6: DIF Results for Form C Items Used at 2008 Data Review

Grade Level	N Items Reviewed	N Items Showing Weak DIF (Categories B or BB)	N Items Showing Strong DIF (Categories C or CC)
Primary: Grades K-2	14	0	0
Elementary: Grades 3-5	15	2 Writing CR items	0
Middle Grades: Grades 6-8	18	2 Reading MC items	2 Writing CR items
High School: Grades 9-12	11	2 Writing CR items	0

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 usually discussed by the group to maintain consistency in scoring.

3.1. Rater Training and Intra-Rater Agreement

All PSC readers had a minimum of a Bachelor's degree and successfully completed generalized workshops in performance assessment scoring before ever being considered a potential reader for a specific project. In addition to the general reader training, all readers assigned to score the WLPT-II test were required to qualify on project-specific training with rubrics, anchor papers, and training papers.

The accuracy of scoring was monitored by room directors and team leaders who are experienced and proficient readers. Team leaders successfully completed a two-day general team-leading training workshop, and were seasoned PSC readers who have extensive experience in all facets of scoring. They carefully monitored the scoring and accuracy of their teams of readers.

The team leader monitored readers' scoring through the PSC read-behind system. In this case, unlike blind check scoring, the team leader received the response directly from the first reader. The score was assigned, but it was not entered into the system. This feature allowed the team leader to review and return items to the reader as needed to ensure the accuracy of scoring. The targeted agreement rates for scoring student responses were 70% perfect agreement, and less than 5% non-adjacent agreement. If a reader failed to achieve this agreement rate, he or she was retrained. Readers who failed to meet minimum scoring standards following retraining were removed from the project.

Beginning with the third day of scoring, in addition to regular student responses, readers scored a set of validity responses each day. Validity sets are responses that have been pre-scored by scoring experts, and consisted of five student papers. Readers scored ten validity responses each day. Each reader completed a blind scoring, which was compared with the known scores. A daily validity report was prepared indicating the number and percent in perfect agreement, *within* $\pm$ 1 score point agreement, and *beyond* $\pm$ 1 score point agreement. Any score greater than one point discrepant triggered retraining of the scorer. The targeted agreement for calibration responses was 80 percent perfect agreement, plus 20 percent adjacent agreement. The table below summarizes the overall results of the readers' daily intra-rater agreement for WLPT-II scoring. The summary indicates that the agreement rates met the target.

Table 3.1: 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	93	7
Elementary: Grades 3–5	98	2
Middle Grades: Grades 6–8	95	5
High School: Grades 9–12	98	2

3.2. Inter-Rater Agreement

During the scoring process, a check score (also called a blind read) monitoring process was followed. Ten percent of the student papers were read by two scorers, a reader (first read) and a team leader (check score). 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 reader's score was final.

Data from the check score procedure were analyzed under the two previously stated definitions of inter-rater agreement. The targeted agreement rate for calibration responses was 70% perfect agreement with no more than 5% of greater than ±1 score point discrepancy. Table 3.2 provides the rater agreement statistics for the open-ended Writing items on the 2008 WLPT-II. The statistics indicate that the degree of the inter-rater agreement was on target.

Table 3.2: Inter-Rater Agreement Statistics by Grade Span

	Inter-Rater Agreement	
Grade Span	% Perfect	± 1 Adjacent
Primary: Grades K–2	89.1	10.7
Elementary: Grades 3–5	72.0	27.1
Middle Grades: Grades 6–8	70.9	28.6
High School: Grades 9–12	71.5	27.9

3.3. Research File

After 100% of PSC scoring was completed, SCOPS merged all scoring files to create a Scored File. This file was verified by Pearson's Quality Assurance (QA) based on the description values in the file layout. Once verified, a Research File for the 2008 WLPT-II test was created and verified by QA again. After the verification and approval by QA, 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 2008 WLPT-II Form C 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 3.3) of this report provides the results on inter-rater agreement for WLPT-II.

4.6. Reliability of the Four Modalities

Table 4.1 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 4.1 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)

For the Listening modality of WLPT-II, the reliability ranged from 0.58 to 0.76 across grades with a median of 0.68, whereas for the Reading modality it ranged between 0.70 and 0.83 with a median of 0.79. For the Speaking modality the reliability ranged from 0.88 to 0.95 with a median of .92, and for the Writing modality, it ranged from 0.77 to 0.85 with a median of .82. Generally speaking, the Speaking modality showed higher reliability estimates than the other modalities for all grades. The reliability of the Comprehension score ranged from .80 to .86 with a median of .83. The reliability of the overall test was consistently high over all grades, ranging from 0.91 to 0.95, with a median of 0.93.

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.

Table 4.1: Descriptive Statistics and Reliability by Grade and Modality

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Reliability	SEM
	Composite[c]	83	113	104	12795	50.77	15.69	0.92	4.41
	Listening	20	20	20	12795	14.18	3.52	0.77	1.67
	Reading	24	24	23	12795	4.10	3.81	0.81	1.66
	Speaking	17	38	38	12795	23.83	9.41	0.93	2.52
K	Writing	22	31	30	12795	8.65	4.17	0.78	1.97
	Comprehension[d]	44	44	43	12795	18.28	5.69	0.82	2.43
	Social[e]	37	58	58	12795	38.01	11.70	0.92	3.33
	Academic[f]	46	55	52	12795	12.75	7.19	0.91	2.13
	Productive[g]	24	54	52	12795	28.87	10.51	0.92	3.00
	Composite[c]	83	113	112	13069	73.07	16.02	0.93	4.34
	Listening	20	20	20	13069	16.51	2.35	0.65	1.40
	Reading	24	24	24	13069	10.07	4.95	0.82	2.07
	Speaking	17	38	38	13069	29.25	7.55	0.91	2.30
1	Writing	22	31	31	13069	17.24	5.67	0.85	2.21
	Comprehension[d]	44	44	44	13069	26.58	6.11	0.82	2.56
	Social[e]	37	58	58	13069	45.76	8.93	0.89	2.92
	Academic[f]	46	55	54	13069	27.32	9.85	0.94	2.46
	Productive[g]	24	54	54	13069	38.26	9.08	0.90	2.85
	Composite[c]	83	113	112	9795	87.39	14.60	0.93	3.95
	Listening	20	20	20	9795	17.22	1.98	0.59	1.27
	Reading	24	24	24	9795	15.37	4.87	0.83	2.01
	Speaking	17	38	38	9795	31.97	6.20	0.89	2.02
2	Writing	22	31	31	9795	22.83	5.08	0.83	2.07
	Comprehension[d]	44	44	44	9795	32.59	5.95	0.83	2.44
	Social[e]	37	58	58	9795	49.19	7.38	0.88	2.56
	Academic[f]	46	55	55	9795	38.19	9.32	0.94	2.25
	Productive[g]	24	54	54	9795	43.30	7.75	0.89	2.62
	Composite[c]	82	109	108	7961	72.84	14.68	0.92	4.17
	Listening	20	20	20	7961	12.93	3.30	0.70	1.81
	Reading	23	23	23	7961	11.30	3.75	0.71	2.01
	Speaking	17	38	38	7961	31.93	6.09	0.90	1.95
3	Writing	22	28	28	7961	16.68	5.11	0.81	2.22
	Comprehension[d]	43	43	42	7961	24.23	6.20	0.81	2.72
	Social[e]	37	58	58	7961	44.87	8.19	0.88	2.80
	Academic[f]	45	51	51	7961	27.98	8.06	0.86	3.02
	Productive[g]	19	46	46	7961	36.68	7.02	0.90	2.27
	Composite[c]	82	109	107	6625	79.10	15.08	0.93	4.00
	Listening	20	20	20	6625	14.18	3.16	0.70	1.72
	Reading	23	23	23	6625	13.05	4.12	0.76	2.01
	Speaking	17	38	38	6625	32.89	5.92	0.90	1.83
4	Writing	22	28	28	6625	18.98	5.11	0.83	2.13
	Comprehension[d]	43	43	43	6625	27.24	6.49	0.83	2.66
	Social[e]	37	58	58	6625	47.07	8.05	0.89	2.64
	Academic[f]	45	51	51	6625	32.03	8.48	0.88	2.95
	Productive[g]	19	46	46	6625	38.30	6.92	0.90	2.17

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.

Table 4.1: Descriptive Statistics and Reliability by Grade and Modality (Continued)

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Reliability	SEM
	Composite[c]	82	109	108	5789	83.04	15.36	0.94	3.89
	Listening	20	20	20	5789	14.89	3.10	0.72	1.65
	Reading	23	23	23	5789	14.35	4.27	0.78	1.98
	Speaking	17	38	38	5789	33.27	5.93	0.91	1.78
5	Writing	22	28	28	5789	20.52	5.11	0.84	2.05
	Comprehension[d]	43	43	42	5789	29.24	6.58	0.84	2.60
	Social[e]	37	58	58	5789	48.17	8.08	0.90	2.55
	Academic[f]	45	51	51	5789	34.87	8.67	0.89	2.87
	Productive[g]	19	46	46	5789	39.15	7.01	0.91	2.12
	Composite[c]	91	118	115	4719	86.69	16.03	0.94	4.01
	Listening	20	20	20	4719	13.79	3.11	0.67	1.78
	Reading	28	28	28	4719	16.57	4.88	0.80	2.17
	Speaking	17	38	38	4719	33.73	6.11	0.93	1.65
6	Writing	26	32	31	4719	22.59	5.06	0.83	2.07
	Comprehension[d]	48	48	48	4719	30.36	7.19	0.85	2.82
	Social[e]	37	58	58	4719	47.52	8.19	0.90	2.57
	Academic[f]	54	60	58	4719	39.17	9.21	0.89	3.02
	Productive[g]	19	46	46	4719	38.81	6.89	0.93	1.88
	Composite[c]	91	118	116	3688	85.98	17.79	0.95	4.09
	Listening	20	20	20	3688	13.60	3.30	0.71	1.78
	Reading	28	28	28	3688	16.83	5.08	0.82	2.16
	Speaking	17	38	38	3688	32.89	6.90	0.94	1.76
7	Writing	26	32	32	3688	22.66	5.39	0.85	2.08
	Comprehension[d]	48	48	48	3688	30.42	7.62	0.86	2.82
	Social[e]	37	58	58	3688	46.49	9.26	0.92	2.67
	Academic[f]	54	60	59	3688	39.49	9.77	0.90	3.02
	Productive[g]	19	46	46	3688	38.10	7.83	0.94	1.97
	Composite[c]	91	118	115	3609	88.31	17.54	0.95	4.00
	Listening	20	20	20	3609	13.87	3.31	0.72	1.76
	Reading	28	28	28	3609	17.90	5.16	0.83	2.12
	Speaking	17	38	38	3609	33.10	6.63	0.93	1.73
8	Writing	26	32	32	3609	23.44	5.18	0.85	2.03
	Comprehension[d]	48	48	47	3609	31.77	7.74	0.87	2.77
	Social[e]	37	58	58	3609	46.97	9.05	0.92	2.62
	Academic[f]	54	60	59	3609	41.34	9.64	0.91	2.96
	Productive[g]	19	46	46	3609	38.53	7.53	0.93	1.94

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

Table 4.1: Descriptive Statistics and Reliability by Grade and Modality (Continued)

Grade	Modality	N Items	Max Points[a]	Max Points[b]	*N*	Mean	SD	Reliability	SEM
9	Composite[c]	91	118	113	4093	82.07	20.08	0.95	4.31
	Listening	20	20	20	4093	12.89	3.81	0.77	1.84
	Reading	28	28	28	4093	16.90	4.84	0.80	2.15
	Speaking	17	38	38	4093	31.34	8.55	0.95	1.84
	Writing	26	32	32	4093	20.94	5.64	0.84	2.26
	Comprehension[d]	48	48	48	4093	29.79	7.92	0.87	2.84
	Social[e]	37	58	58	4093	44.23	11.51	0.94	2.84
	Academic[f]	54	60	57	4093	37.84	9.80	0.90	3.13
	Productive[g]	19	46	46	4093	36.46	9.79	0.95	2.10
10	Composite[c]	91	118	116	3345	87.22	17.73	0.95	4.09
	Listening	20	20	20	3345	13.63	3.53	0.74	1.80
	Reading	28	28	28	3345	18.45	4.71	0.80	2.10
	Speaking	17	38	38	3345	32.76	7.13	0.94	1.72
	Writing	26	32	32	3345	22.38	5.13	0.82	2.17
	Comprehension[d]	48	48	47	3345	32.08	7.48	0.86	2.78
	Social[e]	37	58	58	3345	46.39	9.78	0.93	2.66
	Academic[f]	54	60	60	3345	40.83	9.17	0.89	3.03
	Productive[g]	19	46	46	3345	38.34	8.20	0.94	1.99
11	Composite[c]	91	118	113	2653	89.57	15.33	0.93	3.99
	Listening	20	20	20	2653	13.94	3.27	0.70	1.79
	Reading	28	28	28	2653	19.30	4.36	0.77	2.07
	Speaking	17	38	38	2653	33.39	6.01	0.92	1.66
	Writing	26	32	32	2653	22.94	4.75	0.80	2.13
	Comprehension[d]	48	48	47	2653	33.25	6.79	0.84	2.76
	Social[e]	37	58	58	2653	47.34	8.32	0.90	2.57
	Academic[f]	54	60	59	2653	42.24	8.41	0.87	2.98
	Productive[g]	19	46	46	2653	39.25	6.85	0.92	1.93
12	Composite[c]	91	118	115	1854	89.26	15.52	0.93	4.00
	Listening	20	20	20	1854	13.79	3.31	0.70	1.81
	Reading	28	28	28	1854	19.31	4.60	0.80	2.06
	Speaking	17	38	38	1854	33.22	5.83	0.92	1.67
	Writing	26	32	32	1854	22.94	4.81	0.80	2.15
	Comprehension[d]	48	48	47	1854	33.10	7.08	0.85	2.76
	Social[e]	37	58	58	1854	47.01	8.14	0.90	2.58
	Academic[f]	54	60	59	1854	42.25	8.73	0.88	2.99
	Productive[g]	19	46	46	1854	39.12	6.67	0.91	1.97

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.

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 2008 WLPT-II Form C items (original Form C 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 2008 WLPT-II Form C 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 2008 WLPT-II test is also depicted by the point-biserial correlation and point-polyserial correlation coefficients (item-total correlations), which are contained in Tables C1 – C13 in Appendix C.

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 5.1 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 5.1: Intercorrelations Among Modalities by Grade

Grade	Modality	Listening	Reading	Speaking	Writing
K	Listening	1.00	--	--	--
	Reading	0.20	1.00	--	--
	Speaking	0.54	0.16	1.00	--
	Writing	0.44	0.62	0.36	1.00
1	Listening	1.00	--	--	--
	Reading	0.31	1.00	--	--
	Speaking	0.49	0.29	1.00	--
	Writing	0.44	0.72	0.45	1.00
2	Listening	1.00	--	--	--
	Reading	0.40	1.00	--	--
	Speaking	0.49	0.41	1.00	--
	Writing	0.45	0.75	0.49	1.00
3	Listening	1.00	--	--	--
	Reading	0.55	1.00	--	--
	Speaking	0.48	0.41	1.00	--
	Writing	0.60	0.65	0.50	1.00
4	Listening	1.00	--	--	--
	Reading	0.58	1.00	--	--
	Speaking	0.53	0.45	1.00	--
	Writing	0.63	0.68	0.55	1.00
5	Listening	1.00	--	--	--
	Reading	0.59	1.00	--	--
	Speaking	0.56	0.48	1.00	--
	Writing	0.63	0.71	0.59	1.00
6	Listening	1.00	--	--	--
	Reading	0.60	1.00	--	--
	Speaking	0.53	0.49	1.00	--
	Writing	0.64	0.72	0.61	1.00

Table 5.1: Intercorrelations Among Modalities by Grade (continued)

Grade	Modality	Listening	Reading	Speaking	Writing
7	Listening	1.00	--	--	--
	Reading	0.63	1.00	--	--
	Speaking	0.60	0.54	1.00	--
	Writing	0.69	0.74	0.69	1.00
8	Listening	1.00	--	--	--
	Reading	0.65	1.00	--	--
	Speaking	0.62	0.56	1.00	--
	Writing	0.70	0.74	0.70	1.00
9	Listening	1.00	--	--	--
	Reading	0.67	1.00	--	--
	Speaking	0.69	0.59	1.00	--
	Writing	0.74	0.75	0.72	1.00
10	Listening	1.00	--	--	--
	Reading	0.64	1.00	--	--
	Speaking	0.64	0.56	1.00	--
	Writing	0.70	0.74	0.69	1.00
11	Listening	1.00	--	--	--
	Reading	0.58	1.00	--	--
	Speaking	0.57	0.47	1.00	--
	Writing	0.64	0.71	0.60	1.00
12	Listening	1.00	--	--	--
	Reading	0.59	1.00	--	--
	Speaking	0.55	0.50	1.00	--
	Writing	0.63	0.72	0.60	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 2008 WLPT-II. After eliminating anomalies and other exclusion criteria used in the equating process, approximately 96 percent of the total testing population from 2008 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 5.2 summarizes this comparison for each grade span.

Table 5.2: Principal Component Eigenvalues by Grade Span

Grade Span	Component Number	Eigenvalue	Eigenvalue Ratio
Primary: Grades K-2	1	32.34	4.16
	2	7.77	
Elementary: Grades 3-5	1	23.11	4.59
	2	5.03	
Middle Grades: Grades 6-8	1	29.24	6.50
	2	4.50	
High School: Grades 9-12	1	29.35	6.70
	2	4.38	

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 4.16 to 6.70, increasing as a function of grade span. 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 2008 WLPT-II Form C are presented by grade level in Tables C1 – C13 in Appendix C by grade. The following item information and statistics are presented for each item by grade:

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

6.2. Composite-Level Statistics by Ethnicity and Home Language

Tables 6.1 and 6.2 contain summary statistics on the total test (composite) score by native language and by ethnicity 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 6.1: 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	83	113	95	5	463	55.27	14.41
	Asian			104	0	2148	57.88	16.38
	Caucasian			104	0	1552	51.21	15.68
	Hispanic			92	0	8091	48.46	14.94
	Other			91	6	397	53.36	14.43
	Unidentified			95	0	144	48.06	17.60
1	Black/African	83	113	109	12	383	72.34	16.74
	Asian			112	0	2050	79.35	16.07
	Caucasian			110	2	1569	75.71	16.02
	Hispanic			112	0	8476	71.10	15.50
	Other			112	33	444	73.93	15.15
	Unidentified			106	11	147	70.55	18.39
2	Black/African	83	113	108	4	315	83.50	18.28
	Asian			111	17	1293	91.54	13.50
	Caucasian			112	3	1164	89.85	14.61
	Hispanic			112	0	6641	86.44	14.36
	Other			109	43	293	86.60	12.46
	Unidentified			110	15	89	81.97	21.75
3	Black/African	82	109	95	13	267	69.76	15.91
	Asian			104	0	1008	74.81	15.74
	Caucasian			107	0	950	74.36	15.99
	Hispanic			108	0	5472	72.50	14.00
	Other			102	11	228	70.18	16.16
	Unidentified			99	18	36	69.31	19.29
4	Black/African	82	109	103	13	235	73.79	16.99
	Asian			106	0	867	80.78	16.08
	Caucasian			106	8	730	79.79	16.25
	Hispanic			107	0	4604	79.13	14.26
	Other			103	11	157	76.36	16.32
	Unidentified			107	1	32	66.16	30.42
5	Black/African	82	109	105	17	220	77.42	18.86
	Asian			108	14	798	84.18	15.91
	Caucasian			106	10	605	83.93	16.47
	Hispanic			106	0	4001	83.16	14.67
	Other			103	23	125	79.38	15.71
	Unidentified			104	24	40	77.85	22.14
6	Black/African	91	118	111	9	185	80.41	20.60
	Asian			114	1	598	87.33	17.25
	Caucasian			115	8	544	88.17	15.81
	Hispanic			113	2	3222	86.79	15.48
	Other			110	37	119	84.95	14.49
	Unidentified			108	27	51	84.35	17.67

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Table 6.1: Descriptive Statistics by Grade and Ethnicity (continued)

Grade	Ethnicity	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
7	Black/African	91	118	110	23	153	78.63	19.26
	Asian			113	0	529	84.73	18.86
	Caucasian			116	0	362	85.80	18.86
	Hispanic			113	0	2505	86.75	17.20
	Other			108	24	108	86.53	16.60
	Unidentified			109	31	31	81.39	20.53
8	Black/African	91	118	111	27	155	83.26	18.19
	Asian			114	0	546	88.23	17.22
	Caucasian			111	0	347	89.45	18.79
	Hispanic			115	0	2460	88.50	17.37
	Other			109	35	71	88.65	15.30
	Unidentified			110	20	30	86.27	20.74
9	Black/African	91	118	111	0	317	70.24	23.83
	Asian			113	3	708	83.83	18.33
	Caucasian			113	7	398	86.22	17.69
	Hispanic			112	3	2534	82.46	19.85
	Other			110	50	86	86.97	13.52
	Unidentified			105	0	50	71.16	25.95
10	Black/African	91	118	109	31	206	79.03	17.59
	Asian			116	12	563	88.65	16.94
	Caucasian			112	16	417	88.65	16.78
	Hispanic			114	0	2061	87.41	17.95
	Other			111	51	63	90.05	13.29
	Unidentified			108	22	35	79.66	21.37
11	Black/African	91	118	109	39	187	83.05	14.81
	Asian			113	0	474	91.03	14.73
	Caucasian			113	17	356	88.72	17.47
	Hispanic			112	20	1578	90.05	14.93
	Other			106	55	43	93.23	9.72
	Unidentified			108	30	15	84.40	20.15
12	Black/African	91	118	109	3	152	83.18	17.44
	Asian			114	4	370	89.18	16.31
	Caucasian			115	6	251	88.98	15.52
	Hispanic			114	3	1024	90.37	14.56
	Other			110	29	37	86.35	19.03
	Unidentified			110	58	20	89.30	15.85

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Table 6.2: Descriptive Statistics by Grade and Language

Grade	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
K	Spanish	83	113	92	0	8177	48.62	14.91
	Russian			101	0	709	48.88	15.76
	Vietnamese			104	14	655	53.86	14.34
	Ukrainian			85	1	358	47.59	14.55
	Korean			98	0	219	60.89	16.30
	Tagalog			98	12	134	56.55	13.68
	Other			104	0	2484	56.88	16.45
	Unidentified			93	0	59	48.15	18.90
1	Spanish	83	113	112	0	8606	71.22	15.50
	Russian			109	8	718	75.14	15.85
	Vietnamese			112	0	636	78.39	16.27
	Ukrainian			110	22	365	73.06	15.88
	Korean			111	22	201	82.77	16.39
	Tagalog			112	28	196	79.68	15.35
	Other			111	2	2289	76.62	16.39
	Unidentified			106	11	58	68.31	20.30
2	Spanish	83	113	112	0	6687	86.51	14.29
	Russian			111	3	564	89.81	14.30
	Vietnamese			111	26	421	91.29	14.55
	Ukrainian			110	13	310	89.49	15.03
	Korean			111	39	128	93.31	13.47
	Tagalog			110	57	119	93.84	10.32
	Other			112	4	1522	88.15	15.34
	Unidentified			106	22	44	76.77	23.62
3	Spanish	82	109	108	0	5495	72.56	13.95
	Russian			104	2	434	74.61	16.00
	Vietnamese			100	21	228	76.82	14.11
	Ukrainian			98	11	253	74.52	15.40
	Korean			100	0	140	71.72	21.16
	Tagalog			98	46	115	77.20	9.38
	Other			107	3	1273	72.27	16.29
	Unidentified			99	18	23	66.39	21.54
4	Spanish	82	109	107	0	4607	79.19	14.23
	Russian			103	8	332	79.74	16.06
	Vietnamese			106	0	213	81.86	15.19
	Ukrainian			103	24	213	80.75	16.11
	Korean			106	27	108	79.16	18.69
	Tagalog			105	56	90	82.86	11.65
	Other			106	11	1040	77.54	17.11
	Unidentified			107	1	22	66.18	31.78
5	Spanish	82	109	106	0	3987	83.34	14.29
	Russian			106	10	279	84.07	17.67
	Vietnamese			105	16	161	82.60	18.07
	Ukrainian			105	26	185	84.26	15.63
	Korean			108	36	128	87.73	13.92
	Tagalog			104	42	101	85.59	12.46
	Other			106	8	930	80.46	18.17
	Unidentified			104	28	18	77.78	23.48

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Table 6.2: Descriptive Statistics by Grade and Language (continued)

Grade	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
6	Spanish	91	118	113	2	3200	86.84	15.47
	Russian			111	8	254	88.41	16.08
	Vietnamese			108	1	134	85.73	18.74
	Ukrainian			113	27	132	86.12	16.74
	Korean			110	38	100	87.58	15.65
	Tagalog			110	47	67	90.34	12.23
	Other			115	7	805	85.50	17.81
	Unidentified			103	45	27	83.52	14.76
7	Spanish	91	118	113	0	2495	86.71	17.31
	Russian			116	14	183	86.17	18.10
	Vietnamese			108	21	98	81.34	20.54
	Ukrainian			111	30	97	87.98	16.89
	Korean			113	19	86	83.36	20.25
	Tagalog			110	0	66	87.42	19.25
	Other			112	0	646	83.93	18.25
	Unidentified			107	31	17	77.47	23.72
8	Spanish	91	118	115	0	2434	88.74	17.01
	Russian			111	0	168	87.85	21.56
	Vietnamese			110	24	109	85.76	19.09
	Ukrainian			110	32	100	91.13	14.46
	Korean			114	43	107	88.93	16.59
	Tagalog			111	65	62	93.74	11.49
	Other			114	0	610	86.23	18.72
	Unidentified			110	20	19	82.47	24.40
9	Spanish	91	118	112	3	2529	82.61	19.73
	Russian			113	7	192	86.52	18.25
	Vietnamese			108	3	146	78.37	21.76
	Ukrainian			108	7	122	85.65	17.18
	Korean			113	34	114	87.54	15.96
	Tagalog			110	50	110	91.39	11.31
	Other			111	0	853	78.23	21.34
	Unidentified			103	0	27	64.37	30.70
10	Spanish	91	118	114	0	2058	87.39	18.01
	Russian			112	16	197	87.22	17.61
	Vietnamese			112	28	99	87.21	18.58
	Ukrainian			111	35	123	90.15	16.35
	Korean			114	48	89	91.49	14.98
	Tagalog			112	56	63	94.83	9.68
	Other			116	12	696	85.33	17.32
	Unidentified			105	22	20	75.20	23.83
11	Spanish	91	118	112	17	1582	90.03	14.94
	Russian			109	22	163	88.05	17.61
	Vietnamese			111	44	92	89.96	14.98
	Ukrainian			109	34	98	88.61	18.14
	Korean			111	61	75	94.93	10.20
	Tagalog			111	71	50	93.68	8.92
	Other			113	0	586	88.01	15.81
	Unidentified			100	30	7	74.29	22.82

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

Table 6.2: Descriptive Statistics by Grade and Language (continued)

Grade	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
12	Spanish	91	118	114	3	1027	90.33	14.53
	Russian			115	6	115	86.17	17.09
	Vietnamese			111	47	83	85.86	17.59
	Ukrainian			110	32	64	89.70	16.32
	Korean			106	73	34	93.71	7.93
	Tagalog			108	78	34	94.62	7.73
	Other			115	3	482	87.55	17.10
	Unidentified			110	58	15	89.27	15.07

[a] Maximum points possible

[b] Maximum points observed

[c] Minimum points observed

6.3. Modality-Level Descriptive Statistics

Table 4.1 showed the classical statistics of central tendency, variability, and score precision for the four modality scores, as well as the overall, composite score.

Tables 6.3 – 6.4 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 6.3: 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	83	113	109	4	1161	68.56	20.04
		Asian			112	0	5491	73.82	20.71
		Caucasian			112	0	4285	70.68	22.08
		Hispanic			112	0	23208	67.59	21.41
		Other			112	6	1134	70.00	19.41
		Unidentified			110	0	380	64.70	23.36
	Listening	Black/African	20	20	20	0	1161	15.62	3.11
		Asian			20	0	5491	16.39	2.67
		Caucasian			20	0	4285	16.15	3.05
		Hispanic			20	0	23208	15.69	3.09
		Other			20	0	1134	16.36	2.73
		Unidentified			20	0	380	15.18	3.83
	Reading	Black/African	24	24	24	0	1161	8.74	6.13
		Asian			24	0	5491	10.56	6.79
		Caucasian			24	0	4285	9.40	6.67
		Hispanic			24	0	23208	9.15	6.22
		Other			24	0	1134	9.42	6.09
		Unidentified			24	0	380	8.09	6.52
	Speaking	Black/African	17	38	38	0	1161	28.91	7.88
		Asian			38	0	5491	29.39	7.73
		Caucasian			38	0	4285	28.81	8.44
		Hispanic			38	0	23208	27.54	8.87
		Other			38	0	1134	28.72	7.58
		Unidentified			38	0	380	27.10	9.67
	Writing	Black/African	22	31	31	0	1161	15.29	7.41
		Asian			31	0	5491	17.48	7.66
		Caucasian			31	0	4285	16.32	7.85
		Hispanic			31	0	23208	15.21	7.48
		Other			31	0	1134	15.49	7.41
		Unidentified			29	0	380	14.34	7.80
	Comprehension[e]	Black/African	44	44	44	0	1161	24.36	7.99
		Asian			44	0	5491	26.96	8.35
		Caucasian			44	0	4285	25.55	8.56
		Hispanic			44	0	23208	24.85	8.12
		Other			44	0	1134	25.78	7.64
		Unidentified			43	0	380	23.26	8.86

[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 6.3: Descriptive Statistics by Grade Span and Ethnicity for Modalities (continued)

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	82	109	105	13	722	73.41	17.46
		Asian			108	0	2673	79.54	16.37
		Caucasian			107	0	2285	78.63	16.67
		Hispanic			108	0	14077	77.70	14.95
		Other			103	11	510	74.34	16.54
		Unidentified			107	1	108	71.54	24.37
	Listening	Black/African	20	20	20	0	722	12.84	3.74
		Asian			20	0	2673	14.21	3.40
		Caucasian			20	0	2285	14.00	3.44
		Hispanic			20	0	14077	13.91	3.19
		Other			20	1	510	13.04	3.59
		Unidentified			20	1	108	13.03	4.56
	Reading	Black/African	23	23	23	0	722	11.61	4.57
		Asian			23	0	2673	13.59	4.36
		Caucasian			23	0	2285	12.95	4.39
		Hispanic			23	0	14077	12.63	4.09
		Other			22	0	510	12.07	4.39
		Unidentified			22	0	108	11.62	5.54
	Speaking	Black/African	17	38	38	0	722	31.69	6.31
		Asian			38	0	2673	32.31	6.40
		Caucasian			38	0	2285	32.64	6.27
		Hispanic			38	0	14077	32.77	5.84
		Other			38	6	510	31.90	6.05
		Unidentified			38	0	108	30.51	9.34
	Writing	Black/African	22	28	28	0	722	17.27	5.88
		Asian			28	0	2673	19.44	5.29
		Caucasian			28	0	2285	19.04	5.61
		Hispanic			28	0	14077	18.38	5.22
		Other			28	0	510	17.33	5.63
		Unidentified			27	0	108	16.38	7.55
	Comprehension[e]	Black/African	43	43	41	1	722	24.45	7.58
		Asian			43	0	2673	27.80	7.05
		Caucasian			42	0	2285	26.95	7.09
		Hispanic			43	0	14077	26.54	6.47
		Other			40	3	510	25.11	7.26
		Unidentified			42	1	108	24.65	9.49

[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 6.3: Descriptive Statistics by Grade Span and Ethnicity for Modalities (continued)

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	91	118	111	9	493	80.76	19.50
		Asian			114	0	1673	86.80	17.81
		Caucasian			116	0	1253	87.84	17.62
		Hispanic			115	0	8187	87.29	16.61
		Other			110	24	298	86.40	15.49
		Unidentified			110	20	112	84.04	19.24
	Listening	Black/African	20	20	20	0	493	12.44	3.90
		Asian			20	0	1673	13.80	3.52
		Caucasian			20	0	1253	13.83	3.40
		Hispanic			20	0	8187	13.82	3.07
		Other			20	0	298	13.70	3.29
		Unidentified			19	1	112	13.08	3.54
	Reading	Black/African	28	28	28	0	493	15.17	5.91
		Asian			28	0	1673	17.65	5.21
		Caucasian			28	0	1253	17.32	5.21
		Hispanic			28	0	8187	17.03	4.92
		Other			26	0	298	16.47	4.75
		Unidentified			26	0	112	16.52	5.77
	Speaking	Black/African	17	38	38	5	493	32.14	6.15
		Asian			38	0	1673	32.23	6.38
		Caucasian			38	0	1253	33.50	6.34
		Hispanic			38	0	8187	33.55	6.59
		Other			38	0	298	33.40	5.86
		Unidentified			38	0	112	32.31	7.50
	Writing	Black/African	26	32	32	1	493	21.01	6.16
		Asian			31	0	1673	23.13	5.39
		Caucasian			32	0	1253	23.19	5.54
		Hispanic			32	0	8187	22.89	5.03
		Other			31	2	298	22.83	5.19
		Unidentified			30	2	112	22.13	5.77
	Comprehension[e]	Black/African	48	48	47	1	493	27.61	9.14
		Asian			48	0	1673	31.45	8.04
		Caucasian			48	0	1253	31.15	7.80
		Hispanic			47	0	8187	30.85	7.20
		Other			44	0	298	30.18	7.12
		Unidentified			44	4	112	29.60	8.47

[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 6.3: Descriptive Statistics by Grade Span and Ethnicity for Modalities (continued)

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	91	118	111	0	862	77.40	20.37
		Asian			116	0	2115	87.66	17.08
		Caucasian			115	6	1422	88.05	17.02
		Hispanic			114	0	7197	86.67	17.91
		Other			111	29	229	88.89	14.04
		Unidentified			110	0	120	78.32	23.30
	Listening	Black/African	20	20	20	0	862	11.48	3.96
		Asian			20	0	2115	13.61	3.49
		Caucasian			20	0	1422	13.82	3.39
		Hispanic			20	0	7197	13.60	3.50
		Other			20	3	229	14.01	3.15
		Unidentified			20	0	120	12.22	4.20
	Reading	Black/African	28	28	27	0	862	15.82	5.33
		Asian			28	0	2115	19.07	4.70
		Caucasian			28	0	1422	18.62	4.70
		Hispanic			28	0	7197	18.24	4.64
		Other			27	4	229	18.08	4.16
		Unidentified			28	0	120	16.73	5.44
	Speaking	Black/African	17	38	38	0	862	30.58	7.42
		Asian			38	0	2115	31.94	6.56
		Caucasian			38	0	1422	33.05	6.52
		Hispanic			38	0	7197	32.76	7.58
		Other			38	11	229	34.18	4.79
		Unidentified			38	0	120	29.54	10.37
	Writing	Black/African	26	32	31	0	862	19.52	6.28
		Asian			32	0	2115	23.05	5.03
		Caucasian			32	0	1422	22.56	5.18
		Hispanic			32	0	7197	22.06	5.08
		Other			31	6	229	22.61	4.79
		Unidentified			29	0	120	19.83	6.35
	Comprehension[e]	Black/African	48	48	45	0	862	27.30	8.56
		Asian			46	0	2115	32.68	7.46
		Caucasian			47	0	1422	32.44	7.40
		Hispanic			48	0	7197	31.85	7.32
		Other			43	11	229	32.10	6.51
		Unidentified			44	0	120	28.95	8.79

[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 6.4: 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	83	113	112	0	23470	67.70	21.35
		Russian			111	0	1991	69.94	22.72
		Vietnamese			112	0	1712	72.18	21.48
		Ukrainian			110	1	1033	69.17	22.82
		Korean			111	0	548	76.49	20.60
		Tagalog			112	12	449	76.53	19.73
		Other			112	0	6295	71.62	20.55
		Unidentified			106	0	161	63.24	23.87
	Listening	Spanish	20	20	20	0	23470	15.71	3.08
		Russian			20	0	1991	16.06	3.14
		Vietnamese			20	0	1712	16.29	2.70
		Ukrainian			20	1	1033	15.88	3.11
		Korean			20	0	548	16.40	2.39
		Tagalog			20	0	449	16.38	2.56
		Other			20	0	6295	16.22	2.92
		Unidentified			19	0	161	14.99	4.05
	Reading	Spanish	24	24	24	0	23470	9.16	6.22
		Russian			24	0	1991	9.27	6.70
		Vietnamese			24	0	1712	10.20	6.92
		Ukrainian			24	0	1033	9.26	6.60
		Korean			24	0	548	11.73	6.90
		Tagalog			24	0	449	11.56	6.55
		Other			24	0	6295	9.70	6.53
		Unidentified			23	0	161	8.70	6.34
	Speaking	Spanish	17	38	38	0	23470	27.61	8.84
		Russian			38	0	1991	28.28	8.65
		Vietnamese			38	0	1712	28.70	8.16
		Ukrainian			38	0	1033	27.75	8.89
		Korean			38	0	548	29.55	7.74
		Tagalog			38	0	449	30.02	7.21
		Other			38	0	6295	29.31	7.80
		Unidentified			38	0	161	25.22	10.51
	Writing	Spanish	22	31	31	0	23470	15.21	7.48
		Russian			31	0	1991	16.34	8.02
		Vietnamese			31	0	1712	16.99	7.78
		Ukrainian			31	0	1033	16.27	7.97
		Korean			31	0	548	18.80	7.60
		Tagalog			31	2	449	18.58	7.63
		Other			31	0	6295	16.39	7.58
		Unidentified			29	0	161	14.33	7.77
	Comprehension[e]	Spanish	44	44	44	0	23470	24.88	8.11
		Russian			44	0	1991	25.33	8.70
		Vietnamese			44	0	1712	26.49	8.51
		Ukrainian			43	1	1033	25.14	8.57
		Korean			44	0	548	28.14	8.36
		Tagalog			43	3	449	27.94	7.98
		Other			44	0	6295	25.92	8.27
		Unidentified			41	0	161	23.69	8.80

[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 6.4: Descriptive Statistics by Grade Span and Language (continued)

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	82	109	108	0	14089	77.78	14.83
		Russian			106	2	1045	78.77	16.91
		Vietnamese			106	0	602	80.15	15.82
		Ukrainian			105	11	651	79.33	16.19
		Korean			108	0	376	79.31	19.42
		Tagalog			105	42	306	81.63	11.68
		Other			107	3	3243	76.31	17.45
		Unidentified			107	1	63	69.57	26.18
	Listening	Spanish	20	20	20	0	14089	13.93	3.18
		Russian			20	0	1045	13.91	3.52
		Vietnamese			20	0	602	14.46	3.23
		Ukrainian			20	1	651	14.21	3.30
		Korean			20	0	376	14.44	3.73
		Tagalog			20	2	306	14.41	2.57
		Other			20	0	3243	13.51	3.66
		Unidentified			20	1	63	12.95	4.81
	Reading	Spanish	23	23	23	0	14089	12.64	4.08
		Russian			22	0	1045	13.02	4.42
		Vietnamese			23	0	602	13.88	4.25
		Ukrainian			23	1	651	13.14	4.36
		Korean			23	0	376	14.05	4.77
		Tagalog			22	5	306	13.88	3.82
		Other			23	0	3243	12.53	4.52
		Unidentified			22	0	63	11.40	5.72
	Speaking	Spanish	17	38	38	0	14089	32.81	5.78
		Russian			38	0	1045	32.65	6.31
		Vietnamese			38	0	602	32.53	6.46
		Ukrainian			38	0	651	32.62	6.22
		Korean			38	0	376	30.77	7.55
		Tagalog			38	13	306	33.36	4.17
		Other			38	0	3243	32.05	6.52
		Unidentified			38	0	63	29.21	10.32
	Writing	Spanish	22	28	28	0	14089	18.41	5.20
		Russian			28	0	1045	19.19	5.55
		Vietnamese			28	0	602	19.29	5.02
		Ukrainian			28	2	651	19.36	5.57
		Korean			28	0	376	20.05	5.90
		Tagalog			28	7	306	19.98	4.47
		Other			28	0	3243	18.22	5.77
		Unidentified			27	0	63	16.02	7.96
	Comprehension[e]	Spanish	43	43	43	0	14089	26.56	6.44
		Russian			42	2	1045	26.93	7.21
		Vietnamese			41	0	602	28.33	6.75
		Ukrainian			42	2	651	27.34	6.95
		Korean			43	0	376	28.49	7.92
		Tagalog			42	13	306	28.29	5.42
		Other			42	0	3243	26.04	7.48
		Unidentified			42	1	63	24.35	9.92

[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 6.4: Descriptive Statistics by Grade Span and Language (continued)

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
	Composite[d]	Spanish	91	118	115	0	8129	87.37	16.54
		Russian			116	0	605	87.58	18.34
		Vietnamese			110	1	341	84.48	19.43
		Ukrainian			113	27	329	88.19	16.22
		Korean			114	19	293	86.83	17.54
		Tagalog			111	0	195	90.44	14.94
		Other			115	0	2061	85.22	18.24
		Unidentified			110	20	63	81.57	20.39
	Listening	Spanish	20	20	20	0	8129	13.84	3.06
		Russian			20	0	605	13.91	3.43
		Vietnamese			20	0	341	13.43	3.71
		Ukrainian			20	3	329	13.90	3.18
		Korean			20	5	293	13.78	3.47
		Tagalog			20	0	195	14.55	3.06
		Other			20	0	2061	13.38	3.65
		Unidentified			18	1	63	12.41	3.67
	Reading	Spanish	28	28	28	0	8129	17.05	4.90
		Russian			28	0	605	17.24	5.26
		Vietnamese			27	0	341	17.07	5.26
		Ukrainian			28	3	329	17.44	4.92
		Korean			28	0	293	18.69	5.24
		Tagalog			27	0	195	18.27	4.76
Middle Grades (Grades 6-8)		Other			28	0	2061	16.61	5.50
		Unidentified			26	3	63	16.57	5.66
	Speaking	Spanish	17	38	38	0	8129	33.59	6.56
		Russian			38	0	605	33.38	6.71
		Vietnamese			38	0	341	31.20	7.59
		Ukrainian			38	0	329	33.46	6.01
		Korean			38	0	293	31.51	6.00
		Tagalog			38	0	195	33.47	5.09
		Other			38	0	2061	32.69	6.20
		Unidentified			38	0	63	31.21	8.28
	Writing	Spanish	26	32	32	0	8129	22.90	5.02
		Russian			32	0	605	23.04	5.73
		Vietnamese			31	0	341	22.77	5.58
		Ukrainian			31	4	329	23.40	5.22
		Korean			31	0	293	22.86	5.28
		Tagalog			31	0	195	24.14	4.63
		Other			32	0	2061	22.54	5.71
		Unidentified			30	2	63	21.38	5.95
	Comprehension[e]	Spanish	48	48	47	0	8129	30.88	7.17
		Russian			47	0	605	31.15	7.89
		Vietnamese			44	0	341	30.50	8.30
		Ukrainian			46	10	329	31.34	7.16
		Korean			48	9	293	32.47	8.05
		Tagalog			44	0	195	32.82	7.18
		Other			48	0	2061	29.99	8.43
		Unidentified			42	4	63	28.98	8.60

[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 6.4: Descriptive Statistics by Grade Span and Language (continued)

Grade Span	Modality	Language	N Items	Max Points[a]	Max Points[b]	Min Points[c]	*N*	Mean	SD
	Composite[d]	Spanish	91	118	114	0	7196	86.71	17.86
		Russian			115	6	667	87.04	17.68
		Vietnamese			112	3	420	84.47	19.38
		Ukrainian			111	7	407	88.36	17.08
		Korean			114	34	312	91.12	14.02
		Tagalog			112	50	257	93.11	10.12
		Other			116	0	2617	84.02	18.84
		Unidentified			110	0	69	73.93	26.47
	Listening	Spanish	20	20	20	0	7196	13.61	3.49
		Russian			20	0	667	13.74	3.41
		Vietnamese			20	0	420	12.80	3.78
		Ukrainian			20	0	407	13.93	3.45
		Korean			20	5	312	14.61	2.93
		Tagalog			20	6	257	14.65	2.42
		Other			20	0	2617	12.83	3.81
		Unidentified			20	0	69	12.01	4.54
	Reading	Spanish	28	28	28	0	7196	18.25	4.64
		Russian			28	1	667	18.41	4.71
		Vietnamese			28	3	420	18.79	5.08
		Ukrainian			28	0	407	18.86	4.67
		Korean			28	6	312	20.55	3.82
		Tagalog			27	8	257	19.28	3.52
High School (Grades 9-12)		Other			28	0	2617	17.69	5.15
		Unidentified			25	0	69	15.62	5.46
	Speaking	Spanish	17	38	38	0	7196	32.78	7.54
		Russian			38	0	667	32.65	7.06
		Vietnamese			38	0	420	30.18	7.65
		Ukrainian			38	5	407	32.95	6.49
		Korean			38	11	312	32.11	5.42
		Tagalog			38	8	257	34.40	3.79
		Other			38	0	2617	31.92	6.84
		Unidentified			38	0	69	27.25	12.34
	Writing	Spanish	26	32	32	0	7196	22.06	5.07
		Russian			32	0	667	22.24	5.32
		Vietnamese			31	0	420	22.70	5.32
		Ukrainian			31	2	407	22.62	5.13
		Korean			32	7	312	23.85	4.52
		Tagalog			31	10	257	24.77	3.59
		Other			32	0	2617	21.59	5.73
		Unidentified			29	0	69	19.04	6.78
	Comprehension[e]	Spanish	48	48	48	0	7196	31.86	7.31
		Russian			47	1	667	32.15	7.45
		Vietnamese			46	3	420	31.58	8.15
		Ukrainian			47	0	407	32.79	7.40
		Korean			46	16	312	35.16	6.04
		Tagalog			44	15	257	33.93	5.05
		Other			46	0	2617	30.52	8.25
		Unidentified			44	0	69	27.64	9.17

[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

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).

7.1. 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 7.1, 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 7.2 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 7.1: Sample Item Characteristic Curve

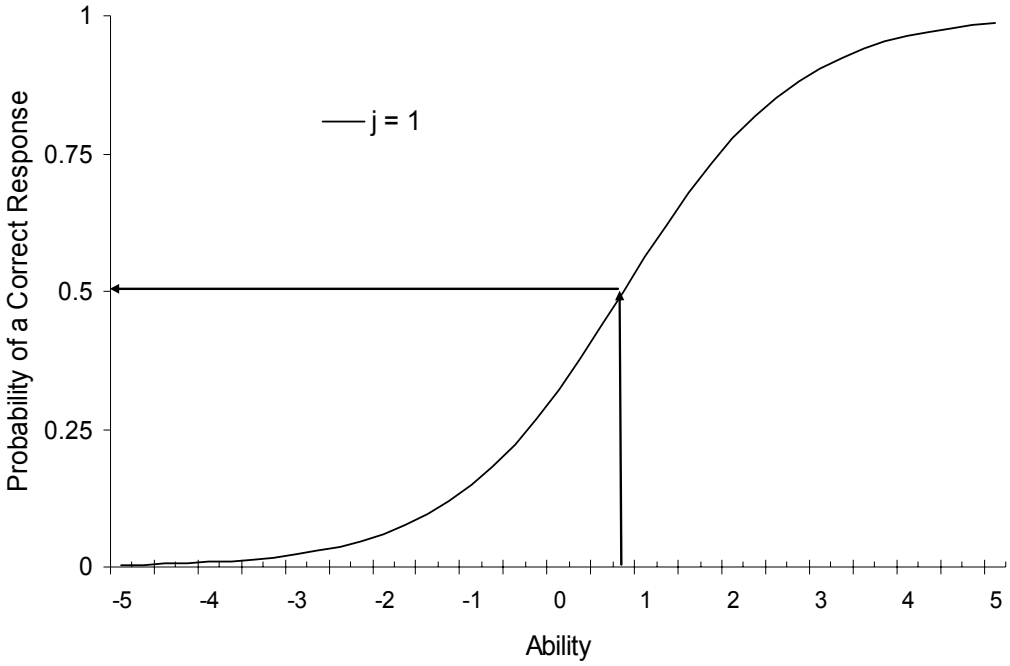
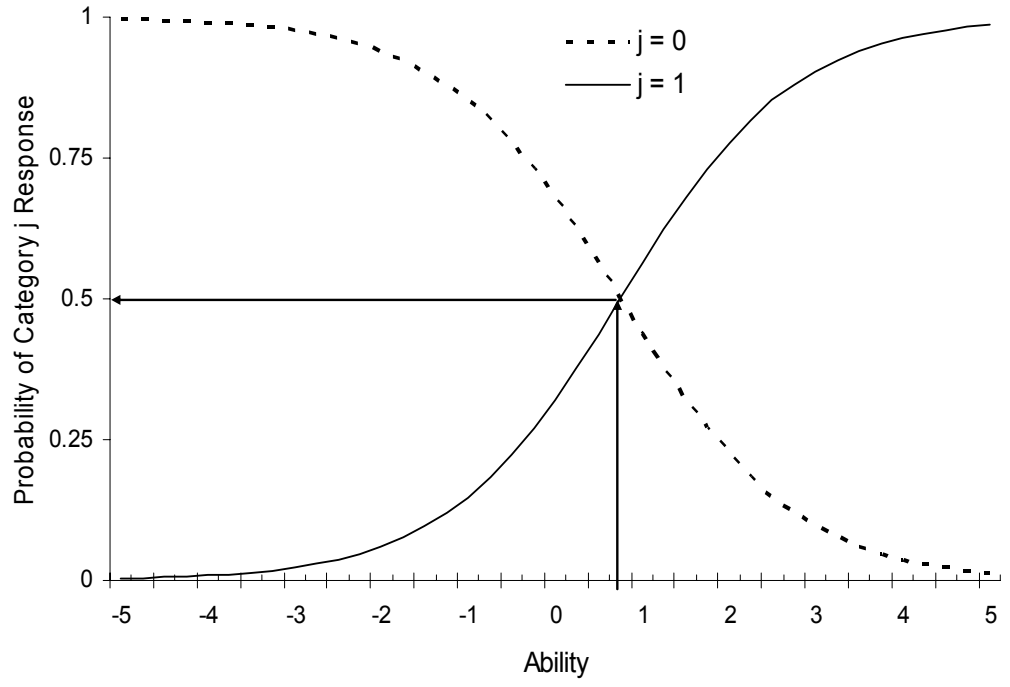


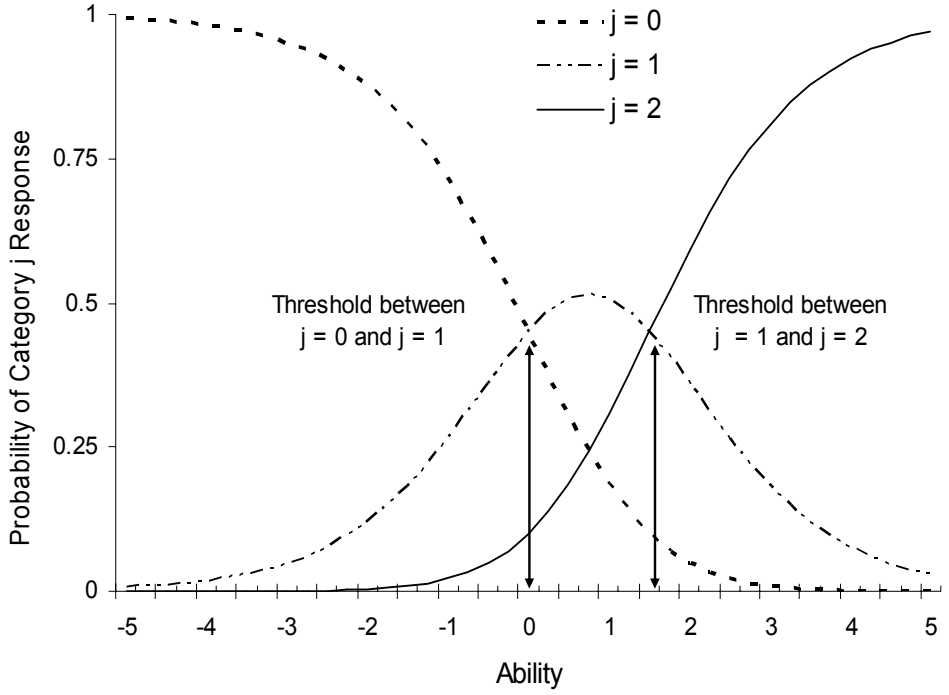
Figure 7.2: Category Response Curves for a Single-Point Item



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 7.3.

The thick dotted curve (j = 0) in Figure 7.3 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 7.3: Category Response Curves for a Two-Point Item



Although the actual computations are more complex, the points at which lines cross in Figure 7.3 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.

7.2. Calibration, Equating, and Scaling of the WLPT-II

The WLPT-II Form C equating study for the 2007 – 2008 school year 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.

7.2.1. 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).*

7.2.2. 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 7.1 summarizes the INFIT and OUTFIT values at each grade span.

Table 7.1: 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	83	0.97	0.15	0.96	0.32	95	73
Elementary: 3-5	82	0.98	0.12	0.99	0.20	98	90
Middle Grades: 6-8	91	0.97	0.13	0.95	0.26	97	78
High School: 9-12	91	0.97	0.14	0.98	0.27	98	71

7.2.3. 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.

A more detailed and comprehensive description of the WLPT-II Form C equating study for the 2007 – 2008 school year, along with the results and WINSTEPS outputs, can be found in the separate report, *Washington Language Proficiency Test – II Form C Equating Study Report (2007 – 2008 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 summaries of the WLPT-II spring and May administrations.

8.1. Spring Administration of the WLPT-II

Table 8.1 presents the scale score summary by grade for each modality (including Comprehension) as well as the overall (composite) test. 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 8.2 contains the percentage of students in each of the proficiency levels by grade. The original adopted WLPT-II overall proficiency cut-scores and the equated 2008 cut-scores can be found in Tables D1 and D2, respectively, in Appendix D.

Note that both Tables 8.1 and 8.2 are based on the full 100% dataset of the WLPT-II Form C 2008 operation administration.

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 B WLPT. For a full discussion and analysis of the Form B WLPT, including the validity, reliability, equating, etc., please refer to the 2007 technical and the equating study reports.

The scale score summaries (as in Table 8.1) 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 8.2).

Table 8.1: Descriptive Statistics of the WLPT-II 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	673	300	12795	550.36	32.34
	Listening	20	718	718	314	12795	567.34	54.51
	Reading	24	776	724	424	12795	515.90	57.44
	Speaking	17	737	737	371	12795	573.75	59.93
	Writing	22	776	723	362	12795	527.38	38.38
	Comprehension [e]	44	783	731	313	12795	544.26	38.26
	Social [f]	37	754	754	308	12795	570.07	48.20
	Academic [g]	46	801	708	356	12795	526.73	40.03
	Productive [h]	24	772	692	345	12690	559.60	39.55
1	Composite [d]	83	810	759	300	13069	593.62	32.80
	Listening	20	718	718	314	13069	606.00	44.85
	Reading	24	776	776	424	13069	585.25	45.35
	Speaking	17	737	737	371	13069	608.95	57.62
	Writing	22	776	776	362	13069	590.39	40.34
	Comprehension [e]	44	783	783	313	13069	591.97	35.34
	Social [f]	37	754	754	308	13069	604.07	43.74
	Academic [g]	46	801	750	356	13069	588.61	37.58
	Productive [h]	24	772	772	345	13026	597.55	38.69
2	Composite [d]	83	810	759	300	9795	625.57	34.37
	Listening	20	718	718	314	9795	620.53	42.58
	Reading	24	776	776	424	9795	630.55	45.13
	Speaking	17	737	737	371	9795	631.81	57.06
	Writing	22	776	776	362	9795	632.64	43.33
	Comprehension [e]	44	783	783	313	9795	626.63	37.38
	Social [f]	37	754	754	308	9795	622.99	42.45
	Academic [g]	46	801	801	356	9795	630.32	38.70
	Productive [h]	24	772	772	397	9782	623.85	41.54
3	Composite [d]	82	857	807	368	7961	643.90	31.61
	Listening	20	792	792	414	7961	641.68	39.44
	Reading	23	826	826	430	7961	644.09	39.30
	Speaking	17	765	765	404	7961	657.25	53.39
	Writing	22	817	817	437	7961	643.74	41.11
	Comprehension [e]	43	838	787	395	7961	643.07	33.70
	Social [f]	37	807	807	383	7961	645.22	35.19
	Academic [g]	45	847	847	408	7961	643.50	36.23
	Productive [h]	19	799	799	401	7845	649.44	40.15
4	Composite [d]	82	857	781	368	6625	658.76	34.74
	Listening	20	792	792	414	6625	657.25	40.64
	Reading	23	826	826	430	6625	660.72	42.60
	Speaking	17	765	765	404	6625	668.72	56.41
	Writing	22	817	817	437	6625	663.51	44.95
	Comprehension [e]	43	838	838	395	6625	658.83	36.03
	Social [f]	37	807	807	383	6625	657.59	38.65
	Academic [g]	45	847	847	408	6625	661.11	39.00
	Productive [h]	19	799	799	401	6529	664.30	45.90

[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 8.1: Descriptive Statistics of the WLPT-II 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	807	368	5789	669.39	36.46
	Listening	20	792	792	414	5789	666.89	41.97
	Reading	23	826	826	430	5789	673.67	44.86
	Speaking	17	765	765	404	5789	674.98	57.71
5	Writing	22	817	817	437	5789	678.87	48.67
	Comprehension [e]	43	838	787	395	5789	669.87	37.02
	Social [f]	37	807	807	383	5789	665.05	40.27
	Academic [g]	45	847	847	408	5789	674.42	41.09
	Productive [h]	19	799	799	401	5715	674.63	51.42
	Composite [d]	91	900	805	442	4719	689.01	33.78
	Listening	20	829	829	443	4719	690.79	39.16
	Reading	28	860	860	445	4719	686.56	41.35
	Speaking	17	795	795	437	4719	714.17	60.79
6	Writing	26	875	817	442	4719	688.50	42.05
	Comprehension [e]	48	873	873	418	4719	687.69	34.73
	Social [f]	37	841	841	414	4719	695.09	38.75
	Academic [g]	54	894	812	417	4719	686.55	37.70
	Productive [h]	19	868	868	434	4652	703.10	47.46
	Composite [d]	91	900	822	390	3688	687.86	38.62
	Listening	20	829	829	443	3688	688.67	42.21
	Reading	28	860	860	445	3688	688.59	43.18
	Speaking	17	795	795	437	3688	706.41	64.14
7	Writing	26	875	875	442	3688	689.85	46.34
	Comprehension [e]	48	873	873	418	3688	687.83	38.15
	Social [f]	37	841	841	414	3688	690.48	43.00
	Academic [g]	54	894	841	417	3688	688.00	41.36
	Productive [h]	19	868	868	434	3640	700.17	52.58
	Composite [d]	91	900	805	390	3609	693.70	39.28
	Listening	20	829	829	443	3609	692.09	42.81
	Reading	28	860	860	445	3609	697.62	44.95
	Speaking	17	795	795	437	3609	708.57	63.43
8	Writing	26	875	875	442	3609	697.10	45.97
	Comprehension [e]	48	873	822	418	3609	694.40	39.60
	Social [f]	37	841	841	414	3609	693.42	43.19
	Academic [g]	54	894	841	417	3609	696.01	42.13
	Productive [h]	19	868	868	434	3564	704.80	53.11

[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 8.1: Descriptive Statistics of the WLPT-II 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	793	399	4093	694.36	40.26
	Listening	20	848	848	487	4093	699.54	43.83
	Reading	28	866	866	432	4093	693.26	43.42
	Speaking	17	806	806	451	4093	712.34	70.08
	Writing	26	873	873	449	4093	695.23	44.90
	Comprehension [e]	48	883	883	424	4093	695.11	39.03
	Social [f]	37	858	858	440	4093	698.52	47.84
	Academic [g]	54	895	800	414	4093	693.49	40.71
	Productive [h]	19	851	851	449	3990	704.20	54.90
10	Composite [d]	91	900	829	399	3345	705.69	37.09
	Listening	20	848	848	487	3345	708.08	41.52
	Reading	28	866	866	432	3345	707.15	42.33
	Speaking	17	806	806	451	3345	724.36	65.90
	Writing	26	873	873	449	3345	707.18	41.90
	Comprehension [e]	48	883	832	424	3345	706.37	36.07
	Social [f]	37	858	858	440	3345	708.44	44.44
	Academic [g]	54	895	895	414	3345	706.14	38.06
	Productive [h]	19	851	851	449	3287	717.40	55.62
11	Composite [d]	91	900	793	399	2653	710.27	33.85
	Listening	20	848	848	487	2653	711.61	38.74
	Reading	28	866	866	432	2653	714.49	39.83
	Speaking	17	806	806	451	2653	725.39	60.02
	Writing	26	873	873	449	2653	711.51	40.10
	Comprehension [e]	48	883	832	424	2653	711.94	33.20
	Social [f]	37	858	858	440	2653	711.13	39.93
	Academic [g]	54	895	844	414	2653	711.85	35.85
	Productive [h]	19	851	851	526	2627	720.72	52.17
12	Composite [d]	91	900	813	494	1854	709.46	34.86
	Listening	20	848	848	487	1854	709.83	39.86
	Reading	28	866	866	432	1854	714.61	44.50
	Speaking	17	806	806	451	1854	721.91	59.87
	Writing	26	873	873	449	1854	711.51	42.40
	Comprehension [e]	48	883	832	424	1854	711.10	36.51
	Social [f]	37	858	858	440	1854	708.94	39.93
	Academic [g]	54	895	844	414	1854	711.81	39.00
	Productive [h]	19	851	851	449	1827	718.96	52.94

[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 8.2: Percentage of Students in Each Proficiency Level by Grade

Grade	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional
K	9	59	27	5
1	2	36	47	15
2	2	20	52	26
3	2	15	64	20
4	2	16	62	20
5	3	16	65	16
6	2	10	64	24
7	3	16	65	16
8	3	16	65	16
9	4	20	58	18
10	2	17	56	25
11	1	13	61	24
12	1	15	64	20

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

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 Table 9.1.

Table 9.1: An Example of Classification Accuracy Table: Proportions of Students Classified into Proficiency Levels by True Scores vs. Observed Scores

	Observed Score Status				
True 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 Table 9.2. It shows the proportions of student performance classified into each proficiency category by the actual test taken and by the hypothetical alternate test form.

Table 9.2: 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 Tables 9.1 and 9.2.

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 9.1 below. It represents a proportion (or percentage) of correct classifications across all the levels. Based on the example shown in Table 9.1, 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 Table 9.1, the overall false positive rate equals .07, which indicates that 7% 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 Table 9.1, 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 9.1: 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 Table 9.2, 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 Table 9.2, 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 9.2). As an example, the consistency at level *Intermediate* is computed from the data in Table 9.2. 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 9.2: Accuracy or Consistency Conditional on Level— Intermediate Equals the Ratio of A Over B

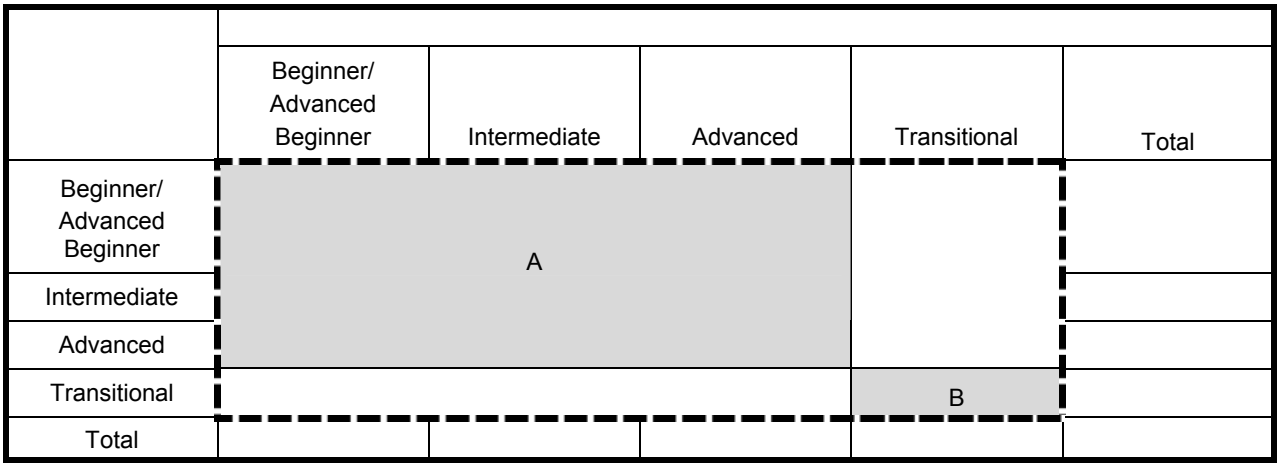
	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional	Total
Beginner/ Advanced Beginner					
Intermediate		A			B
Advanced					
Transitional					
Total					

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 Table 9.1, 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 9.3. 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 9.3: 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 Table 9.1, 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 Table 9.2 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

The procedure detailed in Livingston and Lewis (1995) for estimating decision accuracy and consistency is a multi-step process. The following is a summary of these steps as it was applied in the present report:

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 C 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.

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.

3. Estimate the conditional classification distribution for an alternative form of the test at each level of the proportional true score; i.e, estimate $P\left(X < x_j^* \middle| T_p\right)$, where x_j^* is the *j*-th cut score or cut point. For the results to be presented, scaled cut scores were used.
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.
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.
6. Estimate the joint classification distribution of scores on two alternate forms. Then form a two-way classification table using this joint distribution.
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 9.3 presents the overall classification accuracy results by grade. The overall classification accuracies ranged from 0.78 to 0.86. The overall false positive rates ranged from 0.07 to 0.12, while the false negative rates ranged from 0.06 to 0.11. In most cases, the overall false negative rate at a given grade level tended to be lower than the false positive rate. The Transitional false positive rates ranged from 0.02 to 0.10, while the transitional false negative rates ranged from 0.00 to a maximum of 0.09. The accuracy results for Transitional included three 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 9.3: Overall Accuracy Results by Grade

	Diagonals					Overall			Transitional	
Grade	B/AB	I	A	T*	Overall	False Positive	False Negative		False Positive	False Negative
K	0.06	0.54	0.22	0.04	0.86	0.08	0.07	K	0.02	0.01
1	0.01	0.33	0.40	0.11	0.86	0.07	0.07	1	0.03	0.02
2	0.01	0.18	0.43	0.19	0.81	0.09	0.10	2	0.06	0.07
3	0.00	0.14	0.54	0.14	0.82	0.08	0.09	3	0.06	0.06
4	0.01	0.15	0.54	0.10	0.80	0.09	0.10	4	0.07	0.08
5	0.02	0.15	0.69	*	0.86	0.12	0.02	5	*	*
6	0.01	0.11	0.53	0.13	0.78	0.12	0.11	6	0.10	0.09
7	0.02	0.15	0.69	*	0.86	0.12	0.02	7	*	*
8	0.02	0.16	0.68	*	0.86	0.12	0.02	8	*	*
9	0.02	0.19	0.51	0.10	0.82	0.09	0.09	9	0.06	0.06
10	0.01	0.14	0.47	0.18	0.81	0.09	0.10	10	0.07	0.08
11	0.00	0.12	0.52	0.18	0.82	0.09	0.09	11	0.07	0.07
12	0.00	0.13	0.62	0.06	0.82	0.12	0.06	12	0.10	0.04

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 9.4 presents the overall classification consistency results. Overall classification consistency ranged from 0.72 to 0.80 across grades.

Table 9.4: Overall Consistency Results by Grade

	Diagonals					
Grade	B/AB	I	A	T	Overall	Kappa
K	0.06	0.50	0.20	0.03	0.80	0.64
1	0.01	0.31	0.37	0.10	0.80	0.67
2	0.01	0.16	0.40	0.16	0.74	0.58
3	0.00	0.14	0.50	0.11	0.76	0.55
4	0.01	0.14	0.51	0.08	0.74	0.51
5	0.02	0.14	0.61	0.02	0.79	0.53
6	0.01	0.10	0.49	0.12	0.72	0.47
7	0.02	0.14	0.61	0.02	0.79	0.55
8	0.02	0.15	0.60	0.03	0.79	0.55
9	0.02	0.18	0.48	0.09	0.77	0.59
10	0.01	0.14	0.44	0.15	0.74	0.55
11	0.00	0.12	0.48	0.15	0.76	0.57
12	0.00	0.13	0.56	0.08	0.77	0.53

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

Table 9.5 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 Below proficiency level results ranged from the 60s to the 80s. On the other hand, the consistency results for the Intermediate and Advanced proficiency levels were largely in the .70s and .80, while the Below proficiency level results ranged from the 60s to the 70s. Conditional accuracy results for Transitional included three grades with values = 0.00, indicated by '*' in the table. This, similar to the results in Table 9.3, however, is not an actual indication of accuracy but an artifact of the method as explained in the footnote.

Table 9.5: Conditional Accuracy and Consistency Results by Grade

	Accuracy				Consistency			
Grade	B/AB	I	A	T*	B/AB	I	A	T
K	0.77	0.92	0.79	0.65	0.72	0.86	0.72	0.61
1	0.69	0.88	0.87	0.77	0.67	0.84	0.79	0.72
2	0.78	0.87	0.82	0.76	0.74	0.81	0.76	0.64
3	0.68	0.83	0.87	0.69	0.66	0.79	0.81	0.57
4	0.75	0.86	0.84	0.59	0.72	0.80	0.80	0.46
5	0.83	0.86	0.97	*	0.77	0.81	0.86	0.23
6	0.79	0.87	0.84	0.57	0.74	0.82	0.78	0.50
7	0.80	0.88	0.98	*	0.79	0.83	0.86	0.24
8	0.83	0.87	0.98	*	0.78	0.83	0.86	0.25
9	0.82	0.88	0.86	0.61	0.78	0.83	0.81	0.51
10	0.79	0.88	0.83	0.71	0.74	0.83	0.78	0.60
11	0.68	0.85	0.87	0.71	0.66	0.81	0.80	0.62
12	0.72	0.87	0.91	0.39	0.71	0.81	0.83	0.47

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 9.6 presents the cut point classification accuracy and classification consistency results. Accuracy ranged from 0.81 to 0.97 and consistency ranged from 0.76 to 0.96.

Table 9.6: Cut Point Accuracy and Consistency by Grade

Grade	Accuracy	Consistency
K	0.97	0.96
1	0.94	0.92
2	0.87	0.82
3	0.87	0.83
4	0.85	0.81
5	0.90	0.85
6	0.81	0.76
7	0.90	0.85
8	0.90	0.85
9	0.87	0.84
10	0.85	0.80
11	0.86	0.81
12	0.86	0.83

REFERENCES

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

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

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.

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.

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.

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.

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 C RAW SCORE TO SCALE SCORE CONVERSION TABLES

Table A1: Form C 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	-8.4352	2.0140	299	73
1	-7.0069	1.0274	350	37
2	-6.2589	0.7445	377	27
3	-5.8000	0.6214	394	22
4	-5.4603	0.5487	406	20
5	-5.1871	0.4992	416	18
6	-4.9566	0.4625	425	17
7	-4.7563	0.4336	432	16
8	-4.5786	0.4101	438	15
9	-4.4185	0.3904	444	14
10	-4.2728	0.3735	449	14
11	-4.1388	0.3589	454	13
12	-4.0146	0.3461	459	13
13	-3.8989	0.3347	463	12
14	-3.7901	0.3247	467	12
15	-3.6877	0.3157	471	11
16	-3.5905	0.3077	474	11
17	-3.4981	0.3005	477	11
18	-3.4098	0.2940	481	11
19	-3.3251	0.2881	484	10
20	-3.2437	0.2828	487	10
21	-3.1651	0.2779	489	10
22	-3.0892	0.2734	492	10
23	-3.0155	0.2693	495	10
24	-2.9440	0.2655	497	10
25	-2.8745	0.2620	500	9
26	-2.8066	0.2588	502	9
27	-2.7404	0.2557	505	9
28	-2.6758	0.2529	507	9
29	-2.6125	0.2503	509	9
30	-2.5505	0.2479	512	9
31	-2.4896	0.2455	514	9
32	-2.4299	0.2434	516	9
33	-2.3711	0.2414	518	9
34	-2.3133	0.2395	520	9
35	-2.2563	0.2377	522	9
36	-2.2002	0.2360	524	9
37	-2.1449	0.2345	527	8
38	-2.0903	0.2330	528	8
39	-2.0363	0.2316	530	8
40	-1.9830	0.2304	532	8

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
41	-1.9302	0.2291	534	8
42	-1.8779	0.2280	536	8
43	-1.8261	0.2270	538	8
44	-1.7749	0.2260	540	8
45	-1.7240	0.2251	542	8
46	-1.6735	0.2243	544	8
47	-1.6234	0.2235	545	8
48	-1.5737	0.2228	547	8
49	-1.5241	0.2222	549	8
50	-1.4749	0.2216	551	8
51	-1.4260	0.2211	552	8
52	-1.3772	0.2206	554	8
53	-1.3285	0.2202	556	8
54	-1.2801	0.2199	558	8
55	-1.2319	0.2197	559	8
56	-1.1836	0.2195	561	8
57	-1.1355	0.2193	563	8
58	-1.0874	0.2192	565	8
59	-1.0394	0.2192	566	8
60	-0.9913	0.2193	568	8
61	-0.9432	0.2194	570	8
62	-0.8951	0.2195	572	8
63	-0.8469	0.2198	573	8
64	-0.7984	0.2201	575	8
65	-0.7499	0.2204	577	8
66	-0.7013	0.2209	579	8
67	-0.6524	0.2214	580	8
68	-0.6032	0.2219	582	8
69	-0.5538	0.2226	584	8
70	-0.5041	0.2233	586	8
71	-0.4541	0.2241	588	8
72	-0.4037	0.2250	589	8
73	-0.3528	0.2260	591	8
74	-0.3015	0.2271	593	8
75	-0.2497	0.2282	595	8
76	-0.1974	0.2295	597	8
77	-0.1444	0.2308	599	8
78	-0.0907	0.2323	601	8
79	-0.0365	0.2338	603	8
80	0.0186	0.2355	605	9
81	0.0746	0.2373	607	9

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
82	0.1313	0.2392	609	9
83	0.1890	0.2413	611	9
84	0.2478	0.2435	613	9
85	0.3076	0.2459	615	9
86	0.3687	0.2484	617	9
87	0.4310	0.2511	620	9
88	0.4948	0.2540	622	9
89	0.5602	0.2572	624	9
90	0.6271	0.2605	627	9
91	0.6959	0.2642	629	10
92	0.7668	0.2682	632	10
93	0.8399	0.2725	634	10
94	0.9154	0.2772	637	10
95	0.9936	0.2823	640	10
96	1.0748	0.2880	643	10
97	1.1596	0.2942	646	11
98	1.2481	0.3011	650	11
99	1.3411	0.3089	652	11
100	1.4392	0.3176	656	11
101	1.5431	0.3275	660	12
102	1.6542	0.3389	664	12
103	1.7734	0.3521	668	13
104	1.9028	0.3676	673	13
105	2.0446	0.3861	678	14
106	2.2022	0.4086	684	15
107	2.3805	0.4369	690	16
108	2.5871	0.4737	698	17
109	2.8347	0.5240	706	19
110	3.1470	0.5984	718	22
111	3.5773	0.7245	733	26
112	4.2956	1.0126	759	37
113	5.7009	2.0063	810	73

Table A2: Form C 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	-8.0012	2.0298	314	73
1	-6.5223	1.0614	368	38
2	-5.6981	0.7973	398	29
3	-5.1522	0.6918	418	25
4	-4.7143	0.6368	433	23
5	-4.3303	0.6052	447	22
6	-3.9761	0.5866	460	21
7	-3.6387	0.5762	472	21
8	-3.3100	0.5714	484	21
9	-2.9843	0.5706	496	21
10	-2.6574	0.5734	508	21
11	-2.3254	0.5795	520	21
12	-1.9842	0.5893	532	21
13	-1.6290	0.6034	545	22
14	-1.2537	0.6227	559	23
15	-0.8503	0.6489	573	23
16	-0.4066	0.6854	589	25
17	0.0993	0.7412	608	27
18	0.7175	0.8419	630	30
19	1.6143	1.0945	662	40
20	3.1455	2.0465	718	74

Table A3: Form C 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	-6.4282	1.9792	371	72
1	-5.0920	0.9731	420	35
2	-4.4383	0.6871	443	25
3	-4.0519	0.5677	457	21
4	-3.7686	0.5015	468	18
5	-3.5390	0.4597	476	17
6	-3.3412	0.4313	483	16
7	-3.1643	0.4110	489	15
8	-3.0018	0.3958	495	14
9	-2.8500	0.3840	501	14
10	-2.7063	0.3745	506	14
11	-2.5690	0.3666	511	13
12	-2.4372	0.3598	516	13
13	-2.3099	0.3540	520	13
14	-2.1863	0.3490	525	13
15	-2.0660	0.3448	529	12
16	-1.9483	0.3414	533	12
17	-1.8328	0.3388	538	12
18	-1.7186	0.3371	542	12
19	-1.6054	0.3363	546	12
20	-1.4923	0.3365	550	12
21	-1.3787	0.3377	554	12
22	-1.2639	0.3401	558	12
23	-1.1471	0.3437	562	12
24	-1.0274	0.3485	567	13
25	-0.9038	0.3548	571	13
26	-0.7751	0.3628	576	13
27	-0.6401	0.3725	581	13
28	-0.4970	0.3844	586	14
29	-0.3437	0.3989	591	14
30	-0.1777	0.4166	598	15
31	0.0048	0.4384	604	16
32	0.2088	0.4658	611	17
33	0.4418	0.5013	620	18
34	0.7167	0.5497	630	20
35	1.0570	0.6215	642	22
36	1.5155	0.7437	659	27
37	2.2615	1.0260	686	37
38	3.6869	2.0128	737	73

Table A4: Form C 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	-4.9717	2.0207	424	73
1	-3.5232	1.0406	476	38
2	-2.7472	0.7635	505	28
3	-2.2585	0.6456	522	23
4	-1.8869	0.5782	536	21
5	-1.5789	0.5344	547	19
6	-1.3101	0.5039	557	18
7	-1.0678	0.4818	565	17
8	-0.8437	0.4656	573	17
9	-0.6327	0.4539	581	16
10	-0.4306	0.4456	588	16
11	-0.2346	0.4404	595	16
12	-0.0419	0.4380	602	16
13	0.1498	0.4383	609	16
14	0.3430	0.4413	616	16
15	0.5404	0.4475	623	16
16	0.7447	0.4574	631	17
17	0.9602	0.4718	639	17
18	1.1920	0.4923	647	18
19	1.4482	0.5216	656	19
20	1.7419	0.5647	667	20
21	2.0968	0.6318	680	23
22	2.5667	0.7504	697	27
23	3.3216	1.0298	724	37
24	4.7525	2.0146	776	73

Table A5: Form C 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	-6.6864	2.0508	362	74
1	-5.1447	1.0988	418	40
2	-4.2498	0.8323	450	30
3	-3.6657	0.7045	471	25
4	-3.2307	0.6188	487	22
5	-2.8859	0.5590	500	20
6	-2.5969	0.5186	510	19
7	-2.3424	0.4921	519	18
8	-2.1094	0.4743	528	17
9	-1.8909	0.4611	536	17
10	-1.6835	0.4500	543	16
11	-1.4857	0.4397	550	16
12	-1.2967	0.4299	557	16
13	-1.1157	0.4211	564	15
14	-0.9416	0.4139	570	15
15	-0.7725	0.4086	576	15
16	-0.6071	0.4055	582	15
17	-0.4430	0.4049	588	15
18	-0.2786	0.4067	594	15
19	-0.1115	0.4110	600	15
20	0.0600	0.4179	606	15
21	0.2384	0.4273	613	15
22	0.4260	0.4395	619	16
23	0.6259	0.4549	627	16
24	0.8415	0.4744	634	17
25	1.0780	0.4993	643	18
26	1.3432	0.5322	653	19
27	1.6500	0.5782	664	21
28	2.0228	0.6475	677	23
29	2.5152	0.7669	695	28
30	3.2978	1.0445	723	38
31	4.7533	2.0233	776	73

Table A6: Form C 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	-6.5179	2.0042	368	73
1	-5.1194	1.0080	419	36
2	-4.4104	0.7182	444	26
3	-3.9892	0.5908	460	21
4	-3.6858	0.5155	471	19
5	-3.4469	0.4646	479	17
6	-3.2485	0.4276	486	15
7	-3.0781	0.3992	493	14
8	-2.9279	0.3766	498	14
9	-2.7931	0.3582	503	13
10	-2.6704	0.3429	507	12
11	-2.5572	0.3300	511	12
12	-2.4520	0.3189	515	12
13	-2.3533	0.3093	519	11
14	-2.2603	0.3009	522	11
15	-2.1720	0.2935	525	11
16	-2.0878	0.2869	528	10
17	-2.0073	0.2809	531	10
18	-1.9299	0.2755	534	10
19	-1.8554	0.2706	537	10
20	-1.7833	0.2661	539	10
21	-1.7136	0.2620	542	9
22	-1.6459	0.2582	544	9
23	-1.5802	0.2546	547	9
24	-1.5162	0.2513	549	9
25	-1.4539	0.2483	551	9
26	-1.3929	0.2454	554	9
27	-1.3333	0.2428	556	9
28	-1.2751	0.2403	558	9
29	-1.2178	0.2380	560	9
30	-1.1616	0.2358	562	9
31	-1.1065	0.2339	564	8
32	-1.0523	0.2320	566	8
33	-0.9989	0.2303	568	8
34	-0.9461	0.2288	570	8
35	-0.8941	0.2274	572	8
36	-0.8427	0.2261	573	8
37	-0.7918	0.2250	575	8
38	-0.7415	0.2239	577	8
39	-0.6915	0.2230	579	8
40	-0.6419	0.2223	581	8

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
41	-0.5927	0.2216	582	8
42	-0.5437	0.2211	584	8
43	-0.4950	0.2206	586	8
44	-0.4464	0.2203	588	8
45	-0.3979	0.2201	590	8
46	-0.3495	0.2199	591	8
47	-0.3011	0.2199	593	8
48	-0.2527	0.2200	595	8
49	-0.2043	0.2201	597	8
50	-0.1558	0.2204	598	8
51	-0.1071	0.2207	600	8
52	-0.0584	0.2211	602	8
53	-0.0094	0.2216	604	8
54	0.0398	0.2221	605	8
55	0.0893	0.2227	607	8
56	0.1390	0.2233	609	8
57	0.1890	0.2241	611	8
58	0.2394	0.2248	613	8
59	0.2901	0.2257	614	8
60	0.3413	0.2265	616	8
61	0.3927	0.2274	619	8
62	0.4447	0.2284	620	8
63	0.4971	0.2294	622	8
64	0.5499	0.2304	624	8
65	0.6032	0.2315	626	8
66	0.6572	0.2326	628	8
67	0.7115	0.2337	630	8
68	0.7664	0.2349	632	8
69	0.8218	0.2361	634	9
70	0.8779	0.2373	636	9
71	0.9345	0.2386	638	9
72	0.9918	0.2400	640	9
73	1.0497	0.2414	642	9
74	1.1083	0.2428	644	9
75	1.1676	0.2444	646	9
76	1.2277	0.2460	648	9
77	1.2886	0.2476	651	9
78	1.3504	0.2494	653	9
79	1.4131	0.2513	655	9
80	1.4767	0.2533	657	9
81	1.5414	0.2554	660	9

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
82	1.6071	0.2576	662	9
83	1.6741	0.2600	665	9
84	1.7424	0.2626	667	10
85	1.8121	0.2655	669	10
86	1.8834	0.2685	672	10
87	1.9564	0.2718	675	10
88	2.0312	0.2754	677	10
89	2.1081	0.2793	680	10
90	2.1874	0.2837	683	10
91	2.2691	0.2884	686	10
92	2.3539	0.2937	689	11
93	2.4418	0.2996	692	11
94	2.5335	0.3061	696	11
95	2.6294	0.3135	699	11
96	2.7303	0.3218	703	12
97	2.8368	0.3313	707	12
98	2.9502	0.3423	711	12
99	3.0717	0.3551	715	13
100	3.2030	0.3701	720	13
101	3.3466	0.3882	725	14
102	3.5058	0.4104	731	15
103	3.6854	0.4382	737	16
104	3.8931	0.4746	745	17
105	4.1413	0.5245	754	19
106	4.4540	0.5985	765	22
107	4.8844	0.7245	781	26
108	5.6022	1.0122	807	37
109	7.0070	2.0061	857	73

Table A7: Form C 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	-5.2501	2.0445	414	74
1	-3.7273	1.0880	469	39
2	-2.8487	0.8285	501	30
3	-2.2570	0.7201	522	26
4	-1.7858	0.6568	539	24
5	-1.3837	0.6137	554	22
6	-1.0270	0.5823	567	21
7	-0.7019	0.5591	579	20
8	-0.3991	0.5424	589	20
9	-0.1114	0.5313	600	19
10	0.1672	0.5252	610	19
11	0.4420	0.5240	620	19
12	0.7181	0.5277	630	19
13	1.0010	0.5370	640	19
14	1.2971	0.5528	651	20
15	1.6158	0.5776	662	21
16	1.9701	0.6157	675	22
17	2.3850	0.6772	690	25
18	2.9140	0.7885	709	29
19	3.7271	1.0574	739	38
20	5.2010	2.0286	792	73

Table A8: Form C 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	-5.5339	1.9902	404	72
1	-4.1705	0.9885	453	36
2	-3.4923	0.7015	478	25
3	-3.0892	0.5796	492	21
4	-2.7948	0.5105	503	18
5	-2.5577	0.4658	511	17
6	-2.3557	0.4346	519	16
7	-2.1772	0.4115	525	15
8	-2.0152	0.3938	531	14
9	-1.8659	0.3796	536	14
10	-1.7264	0.3679	541	13
11	-1.5947	0.3581	546	13
12	-1.4695	0.3497	551	13
13	-1.3499	0.3424	555	12
14	-1.2348	0.3362	559	12
15	-1.1235	0.3310	563	12
16	-1.0155	0.3267	567	12
17	-0.9098	0.3233	571	12
18	-0.8061	0.3210	575	12
19	-0.7036	0.3196	578	12
20	-0.6016	0.3193	582	12
21	-0.4995	0.3201	586	12
22	-0.3965	0.3220	590	12
23	-0.2918	0.3253	593	12
24	-0.1846	0.3300	597	12
25	-0.0736	0.3363	601	12
26	0.0420	0.3443	605	12
27	0.1640	0.3544	610	13
28	0.2940	0.3670	615	13
29	0.4343	0.3826	620	14
30	0.5879	0.4018	625	15
31	0.7587	0.4258	631	15
32	0.9528	0.4560	638	16
33	1.1783	0.4952	647	18
34	1.4490	0.5480	656	20
35	1.7901	0.6245	669	23
36	2.2560	0.7512	686	27
37	3.0179	1.0360	713	37
38	4.4618	2.0200	765	73

Table A9: Form C 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	-4.8052	2.0433	430	74
1	-3.2819	1.0912	485	39
2	-2.3868	0.8444	518	31
3	-1.7581	0.7522	540	27
4	-1.2320	0.7016	559	25
5	-0.7679	0.6612	576	24
6	-0.3565	0.6214	591	22
7	0.0062	0.5834	604	21
8	0.3271	0.5503	616	20
9	0.6148	0.5235	626	19
10	0.8779	0.5031	636	18
11	1.1233	0.4884	645	18
12	1.3567	0.4788	653	17
13	1.5833	0.4741	661	17
14	1.8076	0.4740	669	17
15	2.0341	0.4788	678	17
16	2.2679	0.4891	686	18
17	2.5150	0.5063	695	18
18	2.7841	0.5329	705	19
19	3.0887	0.5736	716	21
20	3.4533	0.6388	729	23
21	3.9316	0.7557	746	27
22	4.6942	1.0334	774	37
23	6.1304	2.0162	826	73

Table A10: Form C 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	-4.6139	2.0170	437	73
1	-3.1774	1.0321	489	37
2	-2.4204	0.7502	516	27
3	-1.9527	0.6289	533	23
4	-1.6021	0.5601	546	20
5	-1.3140	0.5164	556	19
6	-1.0629	0.4873	565	18
7	-0.8355	0.4678	574	17
8	-0.6232	0.4546	581	16
9	-0.4208	0.4458	589	16
10	-0.2248	0.4402	596	16
11	-0.0325	0.4368	603	16
12	0.1572	0.4349	610	16
13	0.3461	0.4343	616	16
14	0.5349	0.4348	623	16
15	0.7245	0.4364	630	16
16	0.9161	0.4393	637	16
17	1.1110	0.4438	644	16
18	1.3106	0.4501	651	16
19	1.5170	0.4589	659	17
20	1.7328	0.4707	667	17
21	1.9616	0.4867	675	18
22	2.2085	0.5083	684	18
23	2.4817	0.5383	694	19
24	2.7937	0.5815	705	21
25	3.1689	0.6482	719	23
26	3.6606	0.7654	736	28
27	4.4391	1.0416	765	38
28	5.8890	2.0210	817	73

Table A11: Form C 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	-5.8997	2.0080	390	73
1	-4.4896	1.0154	442	37
2	-3.7664	0.7275	468	26
3	-3.3324	0.6009	483	22
4	-3.0177	0.5256	495	19
5	-2.7690	0.4744	504	17
6	-2.5622	0.4365	511	16
7	-2.3847	0.4071	518	15
8	-2.2288	0.3833	523	14
9	-2.0894	0.3637	528	13
10	-1.9633	0.3471	533	13
11	-1.8479	0.3328	537	12
12	-1.7412	0.3204	541	12
13	-1.6420	0.3096	545	11
14	-1.5492	0.2999	548	11
15	-1.4618	0.2913	551	11
16	-1.3793	0.2836	554	10
17	-1.3008	0.2767	557	10
18	-1.2260	0.2704	560	10
19	-1.1545	0.2647	562	10
20	-1.0858	0.2595	565	9
21	-1.0197	0.2547	567	9
22	-0.9559	0.2504	569	9
23	-0.8942	0.2464	572	9
24	-0.8344	0.2428	574	9
25	-0.7763	0.2394	576	9
26	-0.7198	0.2364	578	9
27	-0.6645	0.2335	580	8
28	-0.6106	0.2309	582	8
29	-0.5579	0.2285	584	8
30	-0.5061	0.2264	586	8
31	-0.4554	0.2244	587	8
32	-0.4054	0.2226	589	8
33	-0.3562	0.2209	591	8
34	-0.3077	0.2194	593	8
35	-0.2599	0.2180	595	8
36	-0.2127	0.2168	596	8
37	-0.1659	0.2157	598	8
38	-0.1196	0.2147	600	8
39	-0.0737	0.2138	602	8
40	-0.0281	0.2131	603	8

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
41	0.0171	0.2124	605	8
42	0.0621	0.2119	606	8
43	0.1069	0.2114	608	8
44	0.1515	0.2111	609	8
45	0.1960	0.2108	611	8
46	0.2404	0.2106	613	8
47	0.2848	0.2105	614	8
48	0.3290	0.2104	616	8
49	0.3733	0.2105	617	8
50	0.4176	0.2105	619	8
51	0.4620	0.2107	621	8
52	0.5064	0.2109	622	8
53	0.5510	0.2112	624	8
54	0.5956	0.2115	625	8
55	0.6405	0.2119	627	8
56	0.6855	0.2124	629	8
57	0.7306	0.2128	630	8
58	0.7761	0.2134	632	8
59	0.8218	0.2139	634	8
60	0.8675	0.2145	635	8
61	0.9138	0.2152	637	8
62	0.9602	0.2159	639	8
63	1.0069	0.2166	640	8
64	1.0540	0.2173	642	8
65	1.1015	0.2181	644	8
66	1.1492	0.2189	646	8
67	1.1973	0.2197	647	8
68	1.2457	0.2206	649	8
69	1.2946	0.2215	651	8
70	1.3438	0.2224	653	8
71	1.3936	0.2234	654	8
72	1.4437	0.2244	656	8
73	1.4942	0.2255	658	8
74	1.5453	0.2265	660	8
75	1.5969	0.2277	662	8
76	1.6491	0.2289	664	8
77	1.7017	0.2301	665	8
78	1.7549	0.2314	668	8
79	1.8088	0.2327	669	8
80	1.8632	0.2341	671	8
81	1.9185	0.2356	673	9

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
82	1.9743	0.2372	675	9
83	2.0309	0.2388	677	9
84	2.0884	0.2406	679	9
85	2.1467	0.2424	682	9
86	2.2059	0.2444	684	9
87	2.2662	0.2465	686	9
88	2.3274	0.2487	688	9
89	2.3899	0.2511	690	9
90	2.4536	0.2536	693	9
91	2.5186	0.2564	695	9
92	2.5851	0.2593	697	9
93	2.6532	0.2624	700	9
94	2.7229	0.2658	702	10
95	2.7946	0.2695	705	10
96	2.8682	0.2735	708	10
97	2.9443	0.2778	710	10
98	3.0227	0.2825	713	10
99	3.1040	0.2877	716	10
100	3.1884	0.2933	719	11
101	3.2762	0.2995	722	11
102	3.3679	0.3064	726	11
103	3.4642	0.3141	729	11
104	3.5654	0.3226	733	12
105	3.6727	0.3323	737	12
106	3.7866	0.3432	741	12
107	3.9087	0.3557	745	13
108	4.0401	0.3700	750	13
109	4.1833	0.3869	755	14
110	4.3405	0.4069	761	15
111	4.5158	0.4311	767	16
112	4.7144	0.4610	774	17
113	4.9442	0.4994	783	18
114	5.2189	0.5511	793	20
115	5.5629	0.6263	805	23
116	6.0303	0.7517	822	27
117	6.7922	1.0355	850	37
118	8.2348	2.0194	902	73

Table A12: Form C 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	-4.4402	2.0552	443	74
1	-2.8868	1.1051	499	40
2	-1.9748	0.8466	532	31
3	-1.3543	0.7394	555	27
4	-0.8548	0.6782	573	25
5	-0.4238	0.6367	589	23
6	-0.0389	0.6053	603	22
7	0.3121	0.5804	615	21
8	0.6373	0.5609	627	20
9	0.9434	0.5464	638	20
10	1.2362	0.5366	649	19
11	1.5211	0.5317	659	19
12	1.8034	0.5318	669	19
13	2.0887	0.5376	680	19
14	2.3839	0.5503	690	20
15	2.6979	0.5722	702	21
16	3.0445	0.6080	714	22
17	3.4485	0.6680	729	24
18	3.9636	0.7789	747	28
19	4.7606	1.0494	776	38
20	6.2213	2.0242	829	73

Table A13: Form C 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	-4.6167	1.9989	437	72
1	-3.2321	0.9999	487	36
2	-2.5370	0.7099	512	26
3	-2.1257	0.5838	527	21
4	-1.8290	0.5104	538	18
5	-1.5940	0.4617	546	17
6	-1.3973	0.4268	553	15
7	-1.2267	0.4006	560	14
8	-1.0744	0.3802	565	14
9	-0.9362	0.3642	570	13
10	-0.8084	0.3512	575	13
11	-0.6888	0.3408	579	12
12	-0.5756	0.3324	583	12
13	-0.4674	0.3257	587	12
14	-0.3630	0.3205	591	12
15	-0.2616	0.3165	594	11
16	-0.1625	0.3138	598	11
17	-0.0646	0.3122	602	11
18	0.0327	0.3116	605	11
19	0.1299	0.3122	609	11
20	0.2278	0.3138	612	11
21	0.3271	0.3166	616	11
22	0.4285	0.3205	619	12
23	0.5328	0.3256	623	12
24	0.6409	0.3322	627	12
25	0.7539	0.3402	631	12
26	0.8728	0.3499	636	13
27	0.9993	0.3615	640	13
28	1.1348	0.3753	645	14
29	1.2818	0.3916	650	14
30	1.4427	0.4111	656	15
31	1.6211	0.4344	663	16
32	1.8221	0.4630	670	17
33	2.0528	0.4993	678	18
34	2.3259	0.5481	688	20
35	2.6642	0.6197	700	22
36	3.1201	0.7418	717	27
37	3.8630	1.0242	744	37
38	5.2852	2.0118	795	73

Table A14: Form C 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	-4.4062	2.0618	445	75
1	-2.8332	1.1157	501	40
2	-1.9045	0.8514	535	31
3	-1.2866	0.7303	557	26
4	-0.8105	0.6538	575	24
5	-0.4193	0.5996	589	22
6	-0.0845	0.5594	601	20
7	0.2108	0.5287	612	19
8	0.4775	0.5050	621	18
9	0.7229	0.4864	630	18
10	0.9521	0.4717	638	17
11	1.1690	0.4602	646	17
12	1.3766	0.4514	654	16
13	1.5772	0.4448	661	16
14	1.7729	0.4404	668	16
15	1.9657	0.4381	675	16
16	2.1574	0.4379	682	16
17	2.3498	0.4400	689	16
18	2.5452	0.4446	696	16
19	2.7461	0.4521	703	16
20	2.9552	0.4633	711	17
21	3.1769	0.4791	719	17
22	3.4164	0.5009	728	18
23	3.6821	0.5316	737	19
24	3.9874	0.5760	748	21
25	4.3569	0.6444	762	23
26	4.8448	0.7637	779	28
27	5.6225	1.0421	807	38
28	7.0744	2.0221	860	73

Table A15: Form C 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	-4.4750	2.0235	442	73
1	-3.0210	1.0422	495	38
2	-2.2468	0.7591	523	27
3	-1.7690	0.6341	540	23
4	-1.4151	0.5605	553	20
5	-1.1289	0.5121	563	19
6	-0.8843	0.4787	572	17
7	-0.6668	0.4553	580	16
8	-0.4673	0.4388	587	16
9	-0.2801	0.4274	594	15
10	-0.1009	0.4198	600	15
11	0.0733	0.4151	607	15
12	0.2443	0.4124	613	15
13	0.4139	0.4114	619	15
14	0.5831	0.4116	625	15
15	0.7529	0.4128	631	15
16	0.9241	0.4150	637	15
17	1.0976	0.4182	644	15
18	1.2743	0.4227	650	15
19	1.4554	0.4287	657	16
20	1.6424	0.4364	663	16
21	1.8369	0.4463	670	16
22	2.0417	0.4590	678	17
23	2.2596	0.4752	686	17
24	2.4948	0.4958	694	18
25	2.7535	0.5222	704	19
26	3.0436	0.5565	714	20
27	3.3780	0.6018	726	22
28	3.7760	0.6627	741	24
29	4.2702	0.7473	758	27
30	4.9205	0.8739	782	32
31	5.8906	1.1340	817	41
32	7.4910	2.0693	875	75

Table A16: Form C 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	-5.6641	2.0133	399	73
1	-4.2388	1.0252	451	37
2	-3.4959	0.7405	477	27
3	-3.0433	0.6157	494	22
4	-2.7112	0.5414	506	20
5	-2.4461	0.4906	515	18
6	-2.2242	0.4529	523	16
7	-2.0327	0.4233	530	15
8	-1.8639	0.3993	536	14
9	-1.7126	0.3791	542	14
10	-1.5754	0.3619	547	13
11	-1.4499	0.3471	551	13
12	-1.3339	0.3340	556	12
13	-1.2264	0.3224	560	12
14	-1.1258	0.3120	563	11
15	-1.0314	0.3027	567	11
16	-0.9422	0.2943	570	11
17	-0.8580	0.2866	573	10
18	-0.7778	0.2796	576	10
19	-0.7014	0.2732	579	10
20	-0.6284	0.2674	581	10
21	-0.5583	0.2620	584	9
22	-0.4910	0.2570	586	9
23	-0.4261	0.2524	589	9
24	-0.3635	0.2481	591	9
25	-0.3030	0.2442	593	9
26	-0.2443	0.2405	595	9
27	-0.1873	0.2371	597	9
28	-0.1319	0.2339	599	8
29	-0.0778	0.2310	601	8
30	-0.0251	0.2283	603	8
31	0.0265	0.2258	605	8
32	0.0769	0.2234	607	8
33	0.1264	0.2213	609	8
34	0.1749	0.2193	610	8
35	0.2226	0.2174	612	8
36	0.2694	0.2158	614	8
37	0.3157	0.2142	616	8
38	0.3613	0.2128	617	8
39	0.4063	0.2115	619	8
40	0.4507	0.2103	620	8

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
41	0.4947	0.2093	622	8
42	0.5383	0.2084	623	8
43	0.5816	0.2076	625	8
44	0.6245	0.2068	627	7
45	0.6672	0.2062	628	7
46	0.7096	0.2057	630	7
47	0.7518	0.2053	631	7
48	0.7939	0.2050	633	7
49	0.8359	0.2048	634	7
50	0.8778	0.2047	636	7
51	0.9197	0.2047	637	7
52	0.9616	0.2047	639	7
53	1.0035	0.2048	640	7
54	1.0456	0.2050	642	7
55	1.0877	0.2053	643	7
56	1.1299	0.2057	645	7
57	1.1723	0.2061	646	7
58	1.2149	0.2066	648	7
59	1.2577	0.2072	649	7
60	1.3007	0.2078	651	8
61	1.3440	0.2085	653	8
62	1.3876	0.2092	654	8
63	1.4315	0.2100	656	8
64	1.4758	0.2109	657	8
65	1.5204	0.2118	659	8
66	1.5655	0.2127	661	8
67	1.6109	0.2137	662	8
68	1.6569	0.2148	664	8
69	1.7032	0.2159	666	8
70	1.7501	0.2170	667	8
71	1.7974	0.2182	669	8
72	1.8454	0.2194	671	8
73	1.8937	0.2207	672	8
74	1.9427	0.2220	675	8
75	1.9923	0.2234	676	8
76	2.0426	0.2248	678	8
77	2.0934	0.2263	680	8
78	2.1450	0.2278	682	8
79	2.1973	0.2294	683	8
80	2.2503	0.2311	685	8
81	2.3040	0.2328	687	8

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

Raw Score	Theta	Std. Error Theta	Scale Score	Std. Error Scale Score
82	2.3586	0.2345	689	8
83	2.4140	0.2364	691	9
84	2.4704	0.2383	693	9
85	2.5277	0.2403	695	9
86	2.5859	0.2425	697	9
87	2.6454	0.2447	700	9
88	2.7058	0.2470	702	9
89	2.7673	0.2495	704	9
90	2.8302	0.2521	706	9
91	2.8945	0.2548	709	9
92	2.9602	0.2578	711	9
93	3.0275	0.2609	713	9
94	3.0964	0.2642	716	10
95	3.1671	0.2678	719	10
96	3.2398	0.2716	721	10
97	3.3147	0.2757	724	10
98	3.3918	0.2801	727	10
99	3.4717	0.2850	731	10
100	3.5543	0.2902	733	10
101	3.6403	0.2960	736	11
102	3.7298	0.3023	740	11
103	3.8232	0.3093	742	11
104	3.9212	0.3171	746	11
105	4.0245	0.3258	750	12
106	4.1339	0.3357	753	12
107	4.2504	0.3471	758	13
108	4.3753	0.3601	762	13
109	4.5104	0.3754	767	14
110	4.6582	0.3937	772	14
111	4.8217	0.4160	778	15
112	5.0062	0.4439	785	16
113	5.2190	0.4802	793	17
114	5.4729	0.5299	802	19
115	5.7915	0.6036	813	22
116	6.2281	0.7289	829	26
117	6.9526	1.0157	855	37
118	8.3626	2.0079	906	73

Table A17: Form C 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	-3.2359	2.0361	487	74
1	-1.7405	1.0700	541	39
2	-0.9038	0.8012	571	29
3	-0.3585	0.6864	591	25
4	0.0652	0.6199	606	22
5	0.4214	0.5764	619	21
6	0.7356	0.5464	631	20
7	1.0224	0.5258	641	19
8	1.2912	0.5123	651	19
9	1.5493	0.5045	660	18
10	1.8020	0.5017	669	18
11	2.0542	0.5037	678	18
12	2.3111	0.5106	688	18
13	2.5776	0.5229	697	19
14	2.8604	0.5418	707	20
15	3.1682	0.5695	719	21
16	3.5148	0.6106	731	22
17	3.9248	0.6749	746	24
18	4.4525	0.7889	765	29
19	5.2685	1.0599	795	38
20	6.7477	2.0309	848	73

Table A18: Form C 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	-4.2160	2.0065	451	73
1	-2.8103	1.0129	502	37
2	-2.0909	0.7256	528	26
3	-1.6590	0.5999	544	22
4	-1.3447	0.5259	555	19
5	-1.0949	0.4761	564	17
6	-0.8859	0.4397	572	16
7	-0.7050	0.4119	578	15
8	-0.5446	0.3899	584	14
9	-0.3997	0.3721	589	13
10	-0.2668	0.3574	594	13
11	-0.1435	0.3453	599	12
12	-0.0277	0.3352	603	12
13	0.0817	0.3269	607	12
14	0.1862	0.3201	611	12
15	0.2869	0.3146	614	11
16	0.3845	0.3103	618	11
17	0.4798	0.3072	621	11
18	0.5735	0.3053	625	11
19	0.6664	0.3044	628	11
20	0.7590	0.3046	631	11
21	0.8521	0.3059	635	11
22	0.9463	0.3084	638	11
23	1.0425	0.3121	642	11
24	1.1415	0.3172	645	11
25	1.2442	0.3238	649	12
26	1.3516	0.3321	653	12
27	1.4652	0.3424	657	12
28	1.5867	0.3550	661	13
29	1.7182	0.3705	666	13
30	1.8623	0.3894	671	14
31	2.0228	0.4128	677	15
32	2.2052	0.4423	684	16
33	2.4173	0.4804	691	17
34	2.6723	0.5320	701	19
35	2.9943	0.6075	712	22
36	3.4369	0.7338	728	27
37	4.1699	1.0203	755	37
38	5.5877	2.0107	806	73

Table A19: Form C 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	-4.7571	2.0591	432	74
1	-3.1832	1.1222	489	41
2	-2.2221	0.8793	524	32
3	-1.5462	0.7728	548	28
4	-1.0070	0.6983	568	25
5	-0.5608	0.6395	584	23
6	-0.1818	0.5936	597	21
7	0.1491	0.5582	609	20
8	0.4450	0.5310	620	19
9	0.7154	0.5098	630	18
10	0.9667	0.4933	639	18
11	1.2034	0.4804	647	17
12	1.4292	0.4703	656	17
13	1.6467	0.4628	664	17
14	1.8583	0.4574	671	17
15	2.0660	0.4542	679	16
16	2.2715	0.4530	686	16
17	2.4769	0.4540	694	16
18	2.6845	0.4574	701	17
19	2.8964	0.4637	709	17
20	3.1156	0.4733	717	17
21	3.3458	0.4872	725	18
22	3.5924	0.5070	734	18
23	3.8632	0.5354	744	19
24	4.1712	0.5772	755	21
25	4.5405	0.6429	768	23
26	5.0244	0.7595	786	27
27	5.7930	1.0364	814	37
28	7.2342	2.0180	866	73

Table A20: Form C 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	-4.2690	2.0368	449	74
1	-2.7730	1.0694	504	39
2	-1.9417	0.7949	534	29
3	-1.4119	0.6708	553	24
4	-1.0148	0.5937	567	21
5	-0.6948	0.5402	579	20
6	-0.4244	0.5017	589	18
7	-0.1872	0.4738	597	17
8	0.0274	0.4536	605	16
9	0.2263	0.4391	612	16
10	0.4144	0.4291	619	16
11	0.5956	0.4227	625	15
12	0.7726	0.4194	632	15
13	0.9479	0.4183	638	15
14	1.1231	0.4191	645	15
15	1.2997	0.4212	651	15
16	1.4783	0.4243	657	15
17	1.6598	0.4280	664	15
18	1.8450	0.4325	671	16
19	2.0342	0.4377	678	16
20	2.2285	0.4440	685	16
21	2.4289	0.4517	692	16
22	2.6371	0.4612	699	17
23	2.8550	0.4729	707	17
24	3.0855	0.4877	716	18
25	3.3321	0.5062	724	18
26	3.6000	0.5300	734	19
27	3.8971	0.5614	745	20
28	4.2359	0.6050	757	22
29	4.6399	0.6707	772	24
30	5.1619	0.7851	791	28
31	5.9711	1.0561	820	38
32	7.4433	2.0282	873	73

APPENDIX B: WLPT-II FORM C ITEM DIFFICULTY AND FIT STATISTICS

Table B1: Form C Primary (Grades K-2): *N* = 35,282

Item Sequence	Difficulty	INFIT	OUTFIT
1	-2.5081	0.89	0.89
2	-3.3771	0.87	0.68
3	-4.2613	0.80	0.54
4	-4.9846	0.89	0.43
5	-4.2231	0.93	0.72
6	-.74780	0.98	1.00
7	-1.9820	0.86	0.85
8	-4.7501	0.90	0.37
9	-4.7209	0.88	0.32
10	-3.4638	1.12	0.79
11	-4.5639	1.24	0.96
12	-2.0961	0.99	0.98
13	-2.8921	0.84	0.90
14	.37830	1.24	1.40
15	.57310	1.32	1.56
16	-.00050	1.35	1.55
17	-3.5433	0.97	1.09
18	-1.8336	0.95	0.98
19	-1.6941	0.90	0.88
20	-.73570	1.34	1.55
21	-.38850	0.90	0.88
22	-2.1881	0.95	0.70
23	-1.8935	0.85	0.62
24	-1.9148	0.70	0.50
25	-1.1531	0.90	0.78
26	-.26850	1.14	1.20
27	-.44830	0.88	0.84
28	-.85700	0.77	0.69
29	-.20820	0.96	0.96
30	1.6585	1.16	1.40
31	-.46020	0.97	0.94
32	-.17190	0.87	0.85
33	.16260	0.77	0.74
34	-.35260	0.83	0.80
35	-.60360	0.76	0.68
36	-1.9650	0.87	0.62
37	-1.0802	0.89	0.78
38	-1.3610	0.87	0.72
39	-1.1537	0.87	0.72
40	-.98010	0.73	0.61

Table B1: Form C Primary (Grades K-2): *N* = 35, 282 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
41	0.4513	0.80	0.82
42	-0.9306	0.95	0.99
43	-0.0375	0.87	0.86
44	-0.0005	0.95	0.97
45	-0.5438	1.09	1.12
46	-0.5558	0.96	0.91
47	-0.1476	0.92	0.89
48	0.3177	0.96	0.97
49	0.1372	0.88	0.86
50	0.7787	1.06	1.20
51	0.1372	1.00	1.00
52	0.1246	1.07	1.13
53	0.8599	1.15	1.21
54	0.7279	0.91	0.97
55	0.7037	1.06	1.15
56	0.3983	0.94	0.95
57	0.8055	1.17	1.35
58	0.7180	0.97	0.98
59	1.0347	1.10	1.28
60	-4.2314	1.02	1.40
61	-3.9520	1.01	1.48
62	-1.4268	1.11	1.13
63	-2.3246	1.22	2.70
64	0.1454	0.80	0.79
65	0.1453	0.73	0.75
66	0.0441	0.64	0.68
67	-3.1296	1.08	1.04
68	-3.1695	1.05	0.99
69	-3.0936	1.00	0.87
70	-2.1647	1.07	0.99
71	-2.1716	0.89	0.79
72	-1.4748	0.90	0.88
73	-1.3270	0.95	0.97
74	-1.3081	0.87	0.86
75	-1.5639	0.95	1.01
76	-1.4595	0.87	0.85
77	-0.6124	1.05	1.04
78	-0.6428	1.10	1.10
79	-1.0946	1.22	1.34
80	-1.4012	0.98	1.02
81	-1.0632	1.05	1.09
82	-1.7432	0.96	0.90
83	-1.0417	1.17	1.24

Table B2: Form C Elementary (Grades 3-5): *N* = 20,064

Item Sequence	Difficulty	INFIT	OUTFIT
1	-2.7463	1.16	1.02
2	-2.2355	1.00	1.01
3	-2.1844	0.74	0.60
4	-1.4546	0.90	0.88
5	0.0573	1.10	1.09
6	2.1859	1.09	1.30
7	1.8089	1.10	1.21
8	0.9646	0.96	0.97
9	0.7128	0.90	0.88
10	1.1959	1.11	1.16
11	-0.6562	0.90	0.80
12	0.4517	1.11	1.15
13	-0.4210	1.20	1.21
14	0.6787	1.09	1.11
15	0.2878	1.03	1.02
16	-0.6767	1.23	1.24
17	0.1746	0.94	0.94
18	1.4441	1.11	1.18
19	0.1902	1.12	1.22
20	2.1903	1.07	1.23
21	-1.5918	0.67	0.43
22	-1.0042	0.99	1.13
23	-1.3806	0.93	0.70
24	0.0173	0.79	0.72
25	0.3615	0.98	0.95
26	1.2567	0.96	0.97
27	-0.4452	1.00	0.91
28	0.6098	0.90	0.86
29	1.5952	1.09	1.18
30	0.6271	0.93	0.90
31	-0.0014	0.87	0.80
32	1.3610	0.96	0.97
33	1.7841	1.05	1.15
34	1.3776	1.04	1.09
35	0.3797	0.87	0.79
36	1.4108	0.97	0.99
37	0.6434	0.90	0.85
38	0.8647	0.92	0.91
39	2.5341	1.15	1.58
40	1.5926	1.18	1.30

Table B2: Form C Elementary (Grades 3-5): *N* = 20,064 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
41	-1.9148	0.92	0.62
42	-1.9964	0.96	1.02
43	-1.7984	0.79	0.51
44	-1.5446	1.26	1.27
45	-1.2895	0.91	0.80
46	1.1959	0.98	0.99
47	1.8367	1.09	1.18
48	1.4774	0.94	0.97
49	1.4910	1.03	1.06
50	1.1629	1.03	1.07
51	1.7841	0.97	1.00
52	1.8367	1.06	1.14
53	1.7493	0.94	0.98
54	2.6463	1.09	1.37
55	1.7147	1.01	1.06
56	1.8975	1.09	1.18
57	2.1212	0.98	1.05
58	2.4463	1.07	1.25
59	0.7736	1.06	1.08
60	1.7921	1.02	1.09
61	2.0319	1.01	1.08
62	1.9062	1.04	1.11
63	2.0205	1.05	1.14
64	0.6437	0.92	0.92
65	0.3673	1.00	1.01
66	-1.9821	1.01	1.00
67	-2.1586	0.89	0.77
68	-1.9458	0.95	0.93
69	-1.8950	0.75	0.63
70	-1.1433	0.81	0.76
71	-0.3887	0.85	0.88
72	-0.7074	0.86	0.86
73	-0.8317	0.69	0.68
74	-0.2626	0.83	0.92
75	-0.4845	0.80	0.85
76	0.1434	0.98	1.06
77	0.4647	1.01	1.04
78	-0.7346	0.95	0.98
79	-0.5364	0.91	0.91
80	-0.5657	0.82	0.89
81	-0.3877	0.99	1.14
82	-0.2451	0.74	0.73

Table B3: Form C Middle Grades (Grades 6-8): *N* = 11,856

Item Sequence	Difficulty	INFIT	OUTFIT
1	-2.2191	1.10	0.82
2	-0.6739	0.75	0.47
3	-1.2386	0.73	0.53
4	-1.0137	0.88	0.50
5	-0.6739	0.77	0.45
6	2.1413	1.08	1.12
7	2.9386	1.20	1.49
8	2.8338	1.13	1.30
9	0.9067	1.07	1.12
10	2.3801	1.22	1.39
11	2.5174	1.06	1.13
12	0.9436	1.17	1.22
13	2.2996	1.11	1.19
14	1.0244	1.08	1.13
15	1.7488	0.97	0.97
16	1.6861	1.04	1.07
17	1.5547	1.08	1.11
18	2.9638	1.04	1.14
19	0.4975	0.93	0.88
20	1.5758	1.12	1.16
21	-0.1975	0.98	0.96
22	-1.4508	1.17	0.73
23	-1.6494	0.93	0.72
24	-0.7253	0.91	0.61
25	0.7180	0.95	0.93
26	0.0004	0.81	0.68
27	0.3163	0.91	0.82
28	0.0480	1.03	0.93
29	0.3290	1.02	1.01
30	1.3407	0.93	0.89
31	1.6571	1.16	1.24
32	1.3180	0.90	0.86
33	2.1048	0.91	0.94
34	2.1431	1.01	1.05
35	1.0740	1.03	1.05
36	0.8984	1.06	1.04
37	1.1976	0.87	0.83
38	1.1976	0.89	0.82
39	1.7344	1.13	1.17
40	3.2616	1.17	1.59

Table B3: Form C Middle Grades (Grades 6-8): *N* = 11,856 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
41	1.0533	0.88	0.85
42	1.0533	0.92	0.80
43	1.9172	1.12	1.15
44	3.3925	1.25	1.75
45	-2.2503	0.96	0.82
46	0.5628	0.82	0.72
47	-1.5478	0.88	0.50
48	-0.8153	0.65	0.34
49	0.2733	0.88	0.81
50	2.3801	0.87	0.89
51	0.7568	0.94	0.89
52	1.8355	1.07	1.07
53	1.2640	0.94	0.89
54	2.4531	1.01	1.05
55	0.2604	0.98	0.84
56	1.0316	0.84	0.72
57	0.7871	0.89	0.75
58	4.2552	1.07	1.78
59	2.6329	1.08	1.18
60	2.9047	1.16	1.40
61	2.2207	0.98	0.98
62	2.6329	1.01	1.10
63	1.1666	0.97	0.91
64	2.9914	1.03	1.18
65	2.1941	1.19	1.27
66	2.9889	1.01	1.09
67	3.2013	1.06	1.24
68	2.4641	0.93	0.95
69	2.1681	1.03	1.05
70	2.5696	1.00	1.12
71	2.6410	1.09	1.20
72	2.0314	1.10	1.14
73	1.2958	0.91	0.93
74	1.6867	0.94	0.96
75	-0.6174	0.92	0.98
76	-0.8114	0.91	0.69
77	-0.6817	0.76	0.64
78	-0.6922	0.79	0.65
79	-0.4875	0.80	0.76
80	0.6042	0.88	0.88

Table B3: Form C Middle Grades (Grades 6-8): *N* = 11,856 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
81	0.7987	0.73	0.71
82	0.2452	0.86	0.87
83	0.6108	0.73	0.72
84	0.3657	0.71	0.68
85	0.8683	0.86	0.88
86	1.0818	0.89	0.95
87	-0.3064	0.92	0.96
88	0.0729	0.82	0.78
89	0.4356	0.89	0.91
90	0.1041	0.80	0.67
91	0.0924	0.82	0.73

Table B4: Form C High School (Grades 9-12): *N* = 11,727

Item Sequence	Difficulty	INFIT	OUTFIT
1	1.2062	0.94	0.89
2	-0.5014	0.88	0.57
3	-0.4554	0.90	0.67
4	0.2660	0.96	0.83
5	1.4899	1.15	1.24
6	3.0604	1.12	1.23
7	1.6993	0.95	0.94
8	1.3597	1.04	1.03
9	1.7028	0.94	0.94
10	2.5030	1.23	1.30
11	3.2662	1.14	1.32
12	2.1053	0.95	0.94
13	3.3681	1.16	1.35
14	2.5442	1.06	1.09
15	1.5546	0.90	0.84
16	1.5175	0.88	0.84
17	4.0650	1.28	1.74
18	1.5432	0.93	0.91
19	2.2003	1.20	1.28
20	1.2763	0.98	0.97
21	-1.5511	0.93	0.60
22	-0.8425	0.89	0.49
23	-0.0370	0.89	0.64
24	-0.6167	0.98	0.91
25	-1.5127	0.95	0.67
26	3.1046	1.14	1.24
27	0.9273	0.99	0.97
28	0.7979	0.86	0.84
29	2.8994	1.16	1.30
30	4.0339	1.23	1.85
31	1.6108	1.20	1.28
32	1.0194	0.79	0.64
33	2.5030	1.05	1.07
34	1.2896	1.10	1.10
35	2.0612	0.87	0.85
36	2.5476	0.94	0.94
37	1.6570	1.09	1.12
38	1.6844	0.84	0.77
39	1.9281	0.86	0.82
40	2.9080	1.02	1.08

Table B4: Form C High School (Grades 9-12): *N* = 11,727 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
41	1.7178	0.84	0.76
42	2.7739	1.02	1.08
43	1.4702	1.12	1.13
44	3.9321	1.17	1.52
45	-1.8183	1.00	1.46
46	-2.2332	0.94	0.69
47	-2.2676	0.93	0.60
48	-0.0238	0.72	0.49
49	0.0588	1.12	1.13
50	1.5644	1.13	1.16
51	2.6247	0.96	0.96
52	0.5018	0.99	0.96
53	1.7398	0.96	0.92
54	1.9860	0.94	0.91
55	0.7278	0.83	0.62
56	0.7517	0.85	0.65
57	2.5924	1.09	1.12
58	1.9281	1.11	1.12
59	2.0506	1.18	1.24
60	3.2539	0.99	1.09
61	0.9327	0.92	0.80
62	2.7281	0.91	0.94
63	1.0990	0.92	0.82
64	3.5731	1.03	1.18
65	2.6422	1.04	1.06
66	3.4325	1.21	1.45
67	2.3148	1.02	1.02
68	3.6034	1.07	1.25
69	3.8064	0.99	1.14
70	3.2202	1.20	1.42
71	2.8203	1.09	1.16
72	3.1759	1.21	1.36
73	1.4403	0.86	0.88
74	1.8920	0.93	1.04
75	-0.2922	0.94	0.96
76	-0.6622	0.89	0.75
77	0.1083	0.93	0.96
78	-0.0505	0.97	1.04
79	0.1842	0.76	0.78
80	0.6868	0.85	0.85

Table B4: Form C High School (Grades 9-12): *N* = 11,727 (continued)

Item Sequence	Difficulty	INFIT	OUTFIT
81	1.0402	0.90	0.88
82	0.7514	0.82	0.83
83	0.6183	0.76	0.69
84	1.0707	0.98	1.00
85	1.2523	0.79	0.91
86	1.3648	0.80	0.82
87	0.8050	0.81	0.77
88	0.7559	0.79	0.64
89	0.8370	0.62	0.53
90	0.7669	0.61	0.60
91	0.7765	0.79	0.68

APPENDIX C: WLPT-II FORM C CLASSICAL ITEM ANALYSIS STATISTICS

Table C1: Form C Grade K (*N* = 12,795)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.74	0.38
Listening	2	0.81	0.36
Listening	3	0.91	0.33
Listening	4	0.95	0.31
Listening	5	0.89	0.38
Listening	6	0.49	0.38
Listening	7	0.78	0.35
Listening	8	0.93	0.40
Listening	9	0.93	0.40
Listening	10	0.80	0.41
Listening	11	0.89	0.36
Listening	12	0.62	0.35
Listening	13	0.81	0.35
Listening	14	0.29	0.17
Listening	15	0.29	0.16
Listening	16	0.46	0.15
Listening	17	0.83	0.37
Listening	18	0.67	0.33
Listening	19	0.61	0.37
Listening	20	0.48	0.21
Writing Conventions	21	0.25	0.19
Writing Conventions	22	0.52	0.24
Writing Conventions	23	0.37	0.39
Writing Conventions	24	0.46	0.42
Writing Conventions	25	0.34	0.16
Writing Conventions	26	0.20	0.23
Writing Conventions	27	0.17	0.27
Writing Conventions	28	0.27	0.27
Writing Conventions	29	0.15	0.26
Writing Conventions	30	0.06	0.17
Writing Conventions	31	0.17	0.27
Writing Conventions	32	0.16	0.27
Writing Conventions	33	0.09	0.29
Writing Conventions	34	0.20	0.23
Writing Conventions	35	0.19	0.25
Reading	36	0.31	0.36
Reading	37	0.33	0.33
Reading	38	0.30	0.33
Reading	39	0.35	0.17
Reading	40	0.19	0.33

Table C1: Form C Grade K (*N* = 12,795) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.10	0.15
Reading	42	0.24	0.32
Reading	43	0.14	0.24
Reading	44	0.17	0.18
Reading	45	0.20	0.25
Reading	46	0.15	0.27
Reading	47	0.14	0.24
Reading	48	0.11	0.23
Reading	49	0.08	0.22
Reading	50	0.17	0.17
Reading	51	0.11	0.22
Reading	52	0.19	0.19
Reading	53	0.08	0.20
Reading	54	0.10	0.23
Reading	55	0.13	0.24
Reading	56	0.12	0.23
Reading	57	0.19	0.19
Reading	58	0.07	0.21
Reading	59	0.12	0.14
Writing	60	0.91	0.26
Writing	61	0.88	0.32
Writing	62	0.93	0.33
Writing	63	1.51	0.40
Writing	64	0.22	0.47
Writing	65	0.31	0.49
Writing	66	0.29	0.48
Speaking	67	1.84	0.42
Speaking	68	1.84	0.43
Speaking	69	1.80	0.47
Speaking	70	1.52	0.52
Speaking	71	1.55	0.53
Speaking	72	1.12	0.64
Speaking	73	1.10	0.63
Speaking	74	1.14	0.66
Speaking	75	1.21	0.59
Speaking	76	1.22	0.64
Speaking	77	1.85	0.71

Table C1: Form C Grade K (*N* = 12,795) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	78	1.80	0.71
Speaking	79	1.15	0.54
Speaking	80	1.17	0.65
Speaking	81	1.07	0.63
Speaking	82	1.44	0.57
Speaking	83	1.02	0.61

Table C2: Form C Grade 1 (*N* = 13,069)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.90	0.28
Listening	2	0.95	0.28
Listening	3	0.98	0.22
Listening	4	0.99	0.20
Listening	5	0.98	0.24
Listening	6	0.71	0.37
Listening	7	0.87	0.32
Listening	8	0.99	0.26
Listening	9	0.99	0.27
Listening	10	0.94	0.35
Listening	11	0.98	0.24
Listening	12	0.84	0.34
Listening	13	0.93	0.25
Listening	14	0.41	0.19
Listening	15	0.39	0.14
Listening	16	0.50	0.13
Listening	17	0.96	0.22
Listening	18	0.83	0.27
Listening	19	0.82	0.32
Listening	20	0.58	0.20
Writing Conventions	21	0.47	0.41
Writing Conventions	22	0.83	0.36
Writing Conventions	23	0.80	0.44
Writing Conventions	24	0.86	0.42
Writing Conventions	25	0.66	0.45
Writing Conventions	26	0.44	0.26
Writing Conventions	27	0.47	0.42
Writing Conventions	28	0.67	0.47
Writing Conventions	29	0.41	0.37
Writing Conventions	30	0.18	0.13
Writing Conventions	31	0.46	0.35
Writing Conventions	32	0.48	0.44
Writing Conventions	33	0.42	0.54
Writing Conventions	34	0.50	0.48
Writing Conventions	35	0.58	0.50
Reading	36	0.80	0.43
Reading	37	0.68	0.38
Reading	38	0.71	0.28
Reading	39	0.62	0.28
Reading	40	0.67	0.22

Table C2: Form C Grade 1 (*N* = 13,069) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.24	0.43
Reading	42	0.68	0.46
Reading	43	0.42	0.49
Reading	44	0.38	0.43
Reading	45	0.47	0.39
Reading	46	0.49	0.42
Reading	47	0.36	0.36
Reading	48	0.30	0.34
Reading	49	0.30	0.37
Reading	50	0.29	0.36
Reading	51	0.32	0.29
Reading	52	0.35	0.37
Reading	53	0.30	0.23
Reading	54	0.25	0.29
Reading	55	0.31	0.25
Reading	56	0.31	0.26
Reading	57	0.30	0.29
Reading	58	0.26	0.21
Reading	59	0.25	0.31
Writing	60	0.97	0.16
Writing	61	0.97	0.31
Writing	62	1.25	0.19
Writing	63	1.90	0.16
Writing	64	0.90	0.15
Writing	65	1.43	0.26
Writing	66	1.58	0.31
Speaking	67	1.92	0.58
Speaking	68	1.92	0.66
Speaking	69	1.91	0.70
Speaking	70	1.75	0.34
Speaking	71	1.76	0.35
Speaking	72	1.48	0.39
Speaking	73	1.42	0.46
Speaking	74	1.48	0.48
Speaking	75	1.51	0.59
Speaking	76	1.55	0.55
Speaking	77	2.52	0.58

Table C2: Form C Grade 1 (*N* = 13,069) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	78	2.48	0.52
Speaking	79	1.42	0.56
Speaking	80	1.55	0.64
Speaking	81	1.46	0.64
Speaking	82	1.74	0.46
Speaking	83	1.38	0.54

Table C3: Form C Grade 2 (*N* = 9,795)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.95	0.27
Listening	2	0.98	0.26
Listening	3	0.99	0.20
Listening	4	0.99	0.25
Listening	5	0.99	0.20
Listening	6	0.78	0.32
Listening	7	0.90	0.34
Listening	8	0.99	0.27
Listening	9	0.99	0.31
Listening	10	0.96	0.40
Listening	11	0.99	0.23
Listening	12	0.91	0.33
Listening	13	0.96	0.27
Listening	14	0.49	0.18
Listening	15	0.43	0.12
Listening	16	0.53	0.14
Listening	17	0.97	0.20
Listening	18	0.90	0.28
Listening	19	0.89	0.36
Listening	20	0.63	0.23
Writing Conventions	21	0.75	0.47
Writing Conventions	22	0.95	0.36
Writing Conventions	23	0.94	0.43
Writing Conventions	24	0.95	0.38
Writing Conventions	25	0.85	0.49
Writing Conventions	26	0.61	0.27
Writing Conventions	27	0.75	0.48
Writing Conventions	28	0.87	0.49
Writing Conventions	29	0.67	0.41
Writing Conventions	30	0.27	0.14
Writing Conventions	31	0.74	0.39
Writing Conventions	32	0.75	0.44
Writing Conventions	33	0.76	0.56
Writing Conventions	34	0.78	0.50
Writing Conventions	35	0.85	0.53
Reading	36	0.95	0.38
Reading	37	0.86	0.41
Reading	38	0.89	0.45
Reading	39	0.88	0.53
Reading	40	0.91	0.48

Table C3: Form C Grade 2 (*N* = 9,795) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.58	0.52
Reading	42	0.82	0.27
Reading	43	0.70	0.48
Reading	44	0.63	0.42
Reading	45	0.67	0.34
Reading	46	0.74	0.42
Reading	47	0.66	0.48
Reading	48	0.54	0.38
Reading	49	0.60	0.46
Reading	50	0.45	0.35
Reading	51	0.55	0.36
Reading	52	0.57	0.31
Reading	53	0.44	0.16
Reading	54	0.52	0.43
Reading	55	0.47	0.30
Reading	56	0.59	0.41
Reading	57	0.43	0.23
Reading	58	0.52	0.35
Reading	59	0.40	0.24
Writing	60	0.98	0.12
Writing	61	0.98	0.11
Writing	62	1.42	0.26
Writing	63	1.95	0.27
Writing	64	1.18	0.51
Writing	65	2.31	0.68
Writing	66	2.51	0.71
Speaking	67	1.96	0.33
Speaking	68	1.96	0.35
Speaking	69	1.95	0.39
Speaking	70	1.86	0.47
Speaking	71	1.87	0.46
Speaking	72	1.68	0.56
Speaking	73	1.60	0.52
Speaking	74	1.65	0.55
Speaking	75	1.69	0.51
Speaking	76	1.69	0.55
Speaking	77	2.90	0.64

Table C3: Form C Grade 2 (*N* = 9,795) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	78	2.84	0.64
Speaking	79	1.59	0.45
Speaking	80	1.72	0.52
Speaking	81	1.61	0.50
Speaking	82	1.83	0.45
Speaking	83	1.58	0.49

Table C4: Form C Grade 3 (*N* = 7,818)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.96	0.23
Listening	2	0.95	0.24
Listening	3	0.96	0.34
Listening	4	0.92	0.30
Listening	5	0.65	0.41
Listening	6	0.25	0.06
Listening	7	0.38	0.20
Listening	8	0.56	0.36
Listening	9	0.65	0.40
Listening	10	0.46	0.23
Listening	11	0.82	0.48
Listening	12	0.58	0.35
Listening	13	0.75	0.39
Listening	14	0.55	0.31
Listening	15	0.65	0.37
Listening	16	0.77	0.38
Listening	17	0.73	0.39
Listening	18	0.42	0.22
Listening	19	0.72	0.24
Listening	20	0.29	0.19
Writing Conventions	21	0.94	0.41
Writing Conventions	22	0.87	0.31
Writing Conventions	23	0.90	0.41
Writing Conventions	24	0.79	0.44
Writing Conventions	25	0.62	0.43
Writing Conventions	26	0.48	0.38
Writing Conventions	27	0.77	0.47
Writing Conventions	28	0.63	0.44
Writing Conventions	29	0.48	0.27
Writing Conventions	30	0.60	0.43
Writing Conventions	31	0.73	0.52
Writing Conventions	32	0.54	0.39
Writing Conventions	33	0.32	0.16
Writing Conventions	34	0.48	0.27
Writing Conventions	35	0.62	0.51
Writing Conventions	36	0.47	0.36
Writing Conventions	37	0.59	0.44
Writing Conventions	38	0.58	0.43
Writing Conventions	39	0.24	0.02
Writing Conventions	40	0.42	0.12

Table C4: Form C Grade 3 (*N* = 7,818) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.93	0.42
Reading	42	0.94	0.31
Reading	43	0.94	0.46
Reading	44	0.89	0.41
Reading	45	0.90	0.43
Reading	46	0.43	0.35
Reading	47	0.39	0.24
Reading	48	0.41	0.35
Reading	49	0.43	0.29
Reading	50	0.55	0.33
Reading	51	0.38	0.35
Reading	52	0.36	0.23
Reading	53	0.39	0.38
Reading	54	0.23	0.15
Reading	55	0.37	0.27
Reading	56	0.36	0.22
Reading	57	0.30	0.28
Reading	58	0.25	0.16
Reading	59	0.58	0.25
Reading	60	0.37	0.28
Reading	61	0.31	0.25
Reading	62	0.34	0.25
Writing	63	0.32	0.22
Writing	64	2.38	0.61
Speaking	65	2.44	0.62
Speaking	66	1.94	0.34
Speaking	67	1.95	0.35
Speaking	68	1.94	0.39
Speaking	69	1.94	0.39
Speaking	70	1.81	0.52
Speaking	71	1.59	0.55
Speaking	72	1.65	0.57
Speaking	73	1.79	0.55
Speaking	74	1.63	0.54
Speaking	75	1.68	0.54
Speaking	76	2.81	0.63
Speaking	77	2.71	0.61
Speaking	78	1.79	0.50
Speaking	79	1.72	0.54
Speaking	80	1.80	0.49

Table C4: Form C Grade 3 (*N* = 7,818) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.66	0.49
Speaking	82	1.74	0.53

Table C5: Form C Grade 4 (*N* = 6,533)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.98	0.28
Listening	2	0.97	0.23
Listening	3	0.98	0.32
Listening	4	0.94	0.33
Listening	5	0.76	0.45
Listening	6	0.26	0.05
Listening	7	0.43	0.23
Listening	8	0.66	0.40
Listening	9	0.75	0.42
Listening	10	0.55	0.27
Listening	11	0.89	0.45
Listening	12	0.68	0.37
Listening	13	0.81	0.36
Listening	14	0.64	0.36
Listening	15	0.75	0.39
Listening	16	0.85	0.41
Listening	17	0.79	0.38
Listening	18	0.47	0.24
Listening	19	0.75	0.23
Listening	20	0.37	0.21
Writing Conventions	21	0.97	0.43
Writing Conventions	22	0.92	0.29
Writing Conventions	23	0.94	0.42
Writing Conventions	24	0.84	0.45
Writing Conventions	25	0.74	0.45
Writing Conventions	26	0.57	0.42
Writing Conventions	27	0.84	0.49
Writing Conventions	28	0.72	0.47
Writing Conventions	29	0.51	0.27
Writing Conventions	30	0.71	0.45
Writing Conventions	31	0.81	0.50
Writing Conventions	32	0.63	0.44
Writing Conventions	33	0.42	0.26
Writing Conventions	34	0.58	0.33
Writing Conventions	35	0.77	0.57
Writing Conventions	36	0.56	0.42
Writing Conventions	37	0.73	0.51
Writing Conventions	38	0.68	0.42
Writing Conventions	39	0.27	0.12
Writing Conventions	40	0.46	0.15

Table C5: Form C Grade 4 (*N* = 6,533) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.97	0.42
Reading	42	0.97	0.30
Reading	43	0.96	0.47
Reading	44	0.92	0.42
Reading	45	0.93	0.43
Reading	46	0.54	0.38
Reading	47	0.47	0.28
Reading	48	0.55	0.41
Reading	49	0.51	0.32
Reading	50	0.61	0.32
Reading	51	0.53	0.41
Reading	52	0.42	0.29
Reading	53	0.52	0.43
Reading	54	0.31	0.25
Reading	55	0.47	0.34
Reading	56	0.41	0.25
Reading	57	0.38	0.33
Reading	58	0.30	0.22
Reading	59	0.67	0.34
Reading	60	0.46	0.32
Reading	61	0.42	0.32
Reading	62	0.44	0.30
Writing	63	0.41	0.28
Writing	64	2.74	0.63
Speaking	65	2.75	0.63
Speaking	66	1.95	0.38
Speaking	67	1.96	0.37
Speaking	68	1.95	0.39
Speaking	69	1.95	0.42
Speaking	70	1.85	0.54
Speaking	71	1.64	0.57
Speaking	72	1.71	0.59
Speaking	73	1.83	0.57
Speaking	74	1.70	0.56
Speaking	75	1.73	0.54
Speaking	76	3.02	0.65
Speaking	77	2.90	0.62
Speaking	78	1.83	0.53
Speaking	79	1.78	0.57
Speaking	80	1.84	0.53

Table C4: Form C Grade 4 (*N* =6,533) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.72	0.51
Speaking	82	1.79	0.57

Table C6: Form C Grade 5 (*N* = 5,713)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.99	0.28
Listening	2	0.98	0.24
Listening	3	0.98	0.35
Listening	4	0.95	0.33
Listening	5	0.82	0.46
Listening	6	0.28	0.09
Listening	7	0.48	0.25
Listening	8	0.73	0.44
Listening	9	0.81	0.43
Listening	10	0.60	0.28
Listening	11	0.92	0.43
Listening	12	0.73	0.35
Listening	13	0.84	0.35
Listening	14	0.70	0.36
Listening	15	0.79	0.42
Listening	16	0.89	0.44
Listening	17	0.81	0.38
Listening	18	0.50	0.24
Listening	19	0.75	0.22
Listening	20	0.45	0.27
Writing Conventions	21	0.97	0.42
Writing Conventions	22	0.93	0.25
Writing Conventions	23	0.96	0.41
Writing Conventions	24	0.87	0.48
Writing Conventions	25	0.81	0.48
Writing Conventions	26	0.63	0.45
Writing Conventions	27	0.89	0.47
Writing Conventions	28	0.78	0.50
Writing Conventions	29	0.54	0.32
Writing Conventions	30	0.78	0.48
Writing Conventions	31	0.84	0.52
Writing Conventions	32	0.68	0.46
Writing Conventions	33	0.52	0.27
Writing Conventions	34	0.65	0.36
Writing Conventions	35	0.85	0.62
Writing Conventions	36	0.64	0.40
Writing Conventions	37	0.81	0.52
Writing Conventions	38	0.73	0.46
Writing Conventions	39	0.35	0.18
Writing Conventions	40	0.55	0.20

Table C6: Form C Grade 5 (*N* = 5,713) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Reading	41	0.97	0.40
Reading	42	0.98	0.27
Reading	43	0.97	0.45
Reading	44	0.93	0.39
Reading	45	0.94	0.44
Reading	46	0.62	0.43
Reading	47	0.54	0.33
Reading	48	0.65	0.45
Reading	49	0.59	0.33
Reading	50	0.63	0.28
Reading	51	0.63	0.44
Reading	52	0.49	0.30
Reading	53	0.60	0.45
Reading	54	0.36	0.31
Reading	55	0.57	0.37
Reading	56	0.45	0.27
Reading	57	0.46	0.35
Reading	58	0.38	0.28
Reading	59	0.73	0.39
Reading	60	0.52	0.32
Reading	61	0.49	0.36
Reading	62	0.49	0.31
Writing	63	0.47	0.32
Writing	64	2.98	0.67
Speaking	65	2.96	0.66
Speaking	66	1.95	0.37
Speaking	67	1.96	0.36
Speaking	68	1.95	0.40
Speaking	69	1.95	0.38
Speaking	70	1.86	0.56
Speaking	71	1.67	0.59
Speaking	72	1.73	0.62
Speaking	73	1.84	0.60
Speaking	74	1.71	0.59
Speaking	75	1.74	0.57
Speaking	76	3.13	0.68
Speaking	77	2.96	0.65
Speaking	78	1.85	0.57
Speaking	79	1.80	0.58
Speaking	80	1.84	0.55

Table C4: Form C Grade 5 (*N* = 5,713) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.75	0.55
Speaking	82	1.81	0.60

Table C7: Form C Grade 6 (*N* = 4,652)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.98	0.26
Listening	2	0.96	0.44
Listening	3	0.97	0.31
Listening	4	0.96	0.43
Listening	5	0.96	0.52
Listening	6	0.56	0.30
Listening	7	0.41	0.10
Listening	8	0.41	0.20
Listening	9	0.75	0.32
Listening	10	0.48	0.13
Listening	11	0.48	0.25
Listening	12	0.71	0.41
Listening	13	0.52	0.31
Listening	14	0.80	0.36
Listening	15	0.65	0.42
Listening	16	0.68	0.34
Listening	17	0.68	0.33
Listening	18	0.40	0.25
Listening	19	0.86	0.47
Listening	20	0.65	0.25
Writing Conventions	21	0.90	0.38
Writing Conventions	22	0.96	0.40
Writing Conventions	23	0.97	0.35
Writing Conventions	24	0.93	0.44
Writing Conventions	25	0.83	0.43
Writing Conventions	26	0.91	0.47
Writing Conventions	27	0.86	0.50
Writing Conventions	28	0.87	0.47
Writing Conventions	29	0.83	0.37
Writing Conventions	30	0.73	0.45
Writing Conventions	31	0.65	0.28
Writing Conventions	32	0.74	0.49
Writing Conventions	33	0.59	0.38
Writing Conventions	34	0.60	0.38
Writing Conventions	35	0.75	0.36
Writing Conventions	36	0.78	0.31
Writing Conventions	37	0.76	0.45
Writing Conventions	38	0.73	0.44
Writing Conventions	39	0.60	0.40
Writing Conventions	40	0.31	0.08

Table C7: Form C Grade 6 (*N* = 4,652) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.75	0.45
Writing Conventions	42	0.76	0.47
Writing Conventions	43	0.53	0.32
Writing Conventions	44	0.31	0.09
Reading	45	0.99	0.29
Reading	46	0.85	0.49
Reading	47	0.98	0.41
Reading	48	0.96	0.54
Reading	49	0.88	0.45
Reading	50	0.57	0.47
Reading	51	0.79	0.41
Reading	52	0.59	0.34
Reading	53	0.70	0.41
Reading	54	0.47	0.32
Reading	55	0.84	0.44
Reading	56	0.77	0.52
Reading	57	0.80	0.52
Reading	58	0.14	0.08
Reading	59	0.35	0.19
Reading	60	0.37	0.13
Reading	61	0.46	0.38
Reading	62	0.44	0.31
Reading	63	0.71	0.46
Reading	64	0.38	0.26
Reading	65	0.52	0.21
Reading	66	0.36	0.28
Reading	67	0.33	0.21
Reading	68	0.47	0.40
Reading	69	0.55	0.36
Reading	70	0.47	0.34
Reading	71	0.42	0.24
Reading	72	0.56	0.28
Writing	73	2.67	0.61
Writing	74	2.47	0.60
Speaking	75	1.90	0.44
Speaking	76	1.93	0.53
Speaking	77	1.94	0.49
Speaking	78	1.93	0.54
Speaking	79	1.90	0.50
Speaking	80	1.62	0.60

Table C7: Form C Grade 6 (*N* = 4,652) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.68	0.64
Speaking	82	1.73	0.58
Speaking	83	1.72	0.63
Speaking	84	1.84	0.60
Speaking	85	3.27	0.70
Speaking	86	3.20	0.69
Speaking	87	1.91	0.50
Speaking	88	1.87	0.60
Speaking	89	1.80	0.57
Speaking	90	1.88	0.62
Speaking	91	1.86	0.59

Table C8: Form C Grade 7 (*N* = 3,642)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.98	0.36
Listening	2	0.94	0.45
Listening	3	0.97	0.37
Listening	4	0.95	0.48
Listening	5	0.94	0.54
Listening	6	0.56	0.29
Listening	7	0.41	0.11
Listening	8	0.42	0.20
Listening	9	0.79	0.35
Listening	10	0.48	0.14
Listening	11	0.47	0.32
Listening	12	0.71	0.42
Listening	13	0.48	0.28
Listening	14	0.76	0.35
Listening	15	0.64	0.42
Listening	16	0.65	0.38
Listening	17	0.65	0.36
Listening	18	0.39	0.27
Listening	19	0.84	0.50
Listening	20	0.68	0.33
Writing Conventions	21	0.90	0.43
Writing Conventions	22	0.95	0.48
Writing Conventions	23	0.97	0.39
Writing Conventions	24	0.93	0.47
Writing Conventions	25	0.81	0.45
Writing Conventions	26	0.90	0.48
Writing Conventions	27	0.85	0.53
Writing Conventions	28	0.86	0.52
Writing Conventions	29	0.85	0.42
Writing Conventions	30	0.72	0.45
Writing Conventions	31	0.66	0.27
Writing Conventions	32	0.72	0.49
Writing Conventions	33	0.61	0.43
Writing Conventions	34	0.55	0.38
Writing Conventions	35	0.76	0.41
Writing Conventions	36	0.77	0.38
Writing Conventions	37	0.76	0.50
Writing Conventions	38	0.75	0.49
Writing Conventions	39	0.56	0.38
Writing Conventions	40	0.33	0.10

Table C8: Form C Grade 7 (*N* = 3,642) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.78	0.49
Writing Conventions	42	0.76	0.49
Writing Conventions	43	0.53	0.34
Writing Conventions	44	0.34	0.11
Reading	45	0.98	0.35
Reading	46	0.84	0.52
Reading	47	0.97	0.46
Reading	48	0.95	0.52
Reading	49	0.86	0.47
Reading	50	0.57	0.49
Reading	51	0.82	0.44
Reading	52	0.62	0.34
Reading	53	0.72	0.47
Reading	54	0.48	0.35
Reading	55	0.86	0.46
Reading	56	0.76	0.56
Reading	57	0.79	0.55
Reading	58	0.17	0.10
Reading	59	0.35	0.21
Reading	60	0.40	0.15
Reading	61	0.49	0.39
Reading	62	0.47	0.32
Reading	63	0.73	0.45
Reading	64	0.37	0.28
Reading	65	0.54	0.20
Reading	66	0.38	0.31
Reading	67	0.33	0.21
Reading	68	0.47	0.40
Reading	69	0.54	0.34
Reading	70	0.46	0.30
Reading	71	0.46	0.27
Reading	72	0.57	0.26
Writing	73	2.73	0.66
Writing	74	2.55	0.66
Speaking	75	1.88	0.46
Speaking	76	1.90	0.59
Speaking	77	1.91	0.56
Speaking	78	1.90	0.60
Speaking	79	1.87	0.58
Speaking	80	1.57	0.65

Table C8: Form C Grade 7 (*N* = 3,642) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.64	0.67
Speaking	82	1.69	0.62
Speaking	83	1.68	0.67
Speaking	84	1.77	0.63
Speaking	85	3.15	0.75
Speaking	86	3.08	0.73
Speaking	87	1.89	0.53
Speaking	88	1.84	0.62
Speaking	89	1.76	0.63
Speaking	90	1.84	0.65
Speaking	91	1.83	0.64

Table C9: Form C Grade 8 (*N* = 3,562)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.98	0.35
Listening	2	0.95	0.46
Listening	3	0.97	0.36
Listening	4	0.95	0.52
Listening	5	0.94	0.54
Listening	6	0.59	0.35
Listening	7	0.40	0.13
Listening	8	0.44	0.22
Listening	9	0.81	0.33
Listening	10	0.47	0.17
Listening	11	0.52	0.31
Listening	12	0.77	0.40
Listening	13	0.46	0.29
Listening	14	0.77	0.34
Listening	15	0.66	0.46
Listening	16	0.67	0.39
Listening	17	0.67	0.38
Listening	18	0.41	0.32
Listening	19	0.84	0.47
Listening	20	0.70	0.31
Writing Conventions	21	0.91	0.40
Writing Conventions	22	0.96	0.49
Writing Conventions	23	0.98	0.37
Writing Conventions	24	0.95	0.48
Writing Conventions	25	0.81	0.48
Writing Conventions	26	0.91	0.46
Writing Conventions	27	0.86	0.54
Writing Conventions	28	0.88	0.50
Writing Conventions	29	0.88	0.38
Writing Conventions	30	0.74	0.49
Writing Conventions	31	0.68	0.25
Writing Conventions	32	0.73	0.52
Writing Conventions	33	0.67	0.47
Writing Conventions	34	0.58	0.37
Writing Conventions	35	0.78	0.38
Writing Conventions	36	0.81	0.36
Writing Conventions	37	0.78	0.50
Writing Conventions	38	0.78	0.49
Writing Conventions	39	0.58	0.37
Writing Conventions	40	0.36	0.12

Table C9: Form C Grade 8 (*N* = 3,562) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.80	0.50
Writing Conventions	42	0.79	0.48
Writing Conventions	43	0.55	0.34
Writing Conventions	44	0.37	0.08
Reading	45	0.99	0.34
Reading	46	0.86	0.52
Reading	47	0.97	0.47
Reading	48	0.96	0.54
Reading	49	0.88	0.46
Reading	50	0.61	0.49
Reading	51	0.84	0.43
Reading	52	0.66	0.34
Reading	53	0.76	0.49
Reading	54	0.54	0.36
Reading	55	0.88	0.43
Reading	56	0.78	0.59
Reading	57	0.82	0.53
Reading	58	0.19	0.12
Reading	59	0.43	0.25
Reading	60	0.44	0.18
Reading	61	0.56	0.44
Reading	62	0.48	0.35
Reading	63	0.79	0.44
Reading	64	0.42	0.29
Reading	65	0.59	0.21
Reading	66	0.42	0.32
Reading	67	0.38	0.26
Reading	68	0.55	0.43
Reading	69	0.59	0.38
Reading	70	0.51	0.36
Reading	71	0.49	0.27
Reading	72	0.63	0.32
Writing	73	2.83	0.65
Writing	74	2.67	0.63
Speaking	75	1.88	0.49
Speaking	76	1.90	0.56
Speaking	77	1.92	0.53
Speaking	78	1.90	0.60
Speaking	79	1.88	0.55
Speaking	80	1.59	0.65

Table C9: Form C Grade 8 (*N* = 3,562) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.65	0.69
Speaking	82	1.70	0.59
Speaking	83	1.68	0.67
Speaking	84	1.81	0.63
Speaking	85	3.15	0.75
Speaking	86	3.10	0.72
Speaking	87	1.89	0.53
Speaking	88	1.85	0.63
Speaking	89	1.78	0.63
Speaking	90	1.85	0.64
Speaking	91	1.84	0.63

Table C10: Form C Grade 9 (*N* = 4,090)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.80	0.55
Listening	2	0.93	0.54
Listening	3	0.93	0.52
Listening	4	0.88	0.47
Listening	5	0.77	0.33
Listening	6	0.41	0.23
Listening	7	0.67	0.49
Listening	8	0.73	0.38
Listening	9	0.71	0.43
Listening	10	0.49	0.21
Listening	11	0.38	0.15
Listening	12	0.62	0.45
Listening	13	0.37	0.19
Listening	14	0.52	0.32
Listening	15	0.71	0.54
Listening	16	0.75	0.48
Listening	17	0.28	0.09
Listening	18	0.70	0.52
Listening	19	0.64	0.26
Listening	20	0.76	0.46
Writing Conventions	21	0.97	0.37
Writing Conventions	22	0.94	0.51
Writing Conventions	23	0.89	0.53
Writing Conventions	24	0.94	0.37
Writing Conventions	25	0.97	0.36
Writing Conventions	26	0.38	0.20
Writing Conventions	27	0.80	0.42
Writing Conventions	28	0.85	0.42
Writing Conventions	29	0.41	0.14
Writing Conventions	30	0.24	-0.03
Writing Conventions	31	0.64	0.36
Writing Conventions	32	0.79	0.57
Writing Conventions	33	0.44	0.34
Writing Conventions	34	0.74	0.27
Writing Conventions	35	0.65	0.51
Writing Conventions	36	0.58	0.47
Writing Conventions	37	0.68	0.40
Writing Conventions	38	0.70	0.60
Writing Conventions	39	0.68	0.52
Writing Conventions	40	0.49	0.35

Table C10: Form C Grade 9 (*N* = 4,090) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.70	0.61
Writing Conventions	42	0.54	0.37
Writing Conventions	43	0.73	0.36
Reading	44	0.28	0.16
Reading	45	0.98	0.21
Reading	46	0.98	0.30
Reading	47	0.98	0.29
Reading	48	0.91	0.53
Reading	49	0.87	0.49
Reading	50	0.75	0.25
Reading	51	0.49	0.41
Reading	52	0.83	0.39
Reading	53	0.65	0.43
Reading	54	0.63	0.47
Reading	55	0.82	0.59
Reading	56	0.80	0.58
Reading	57	0.47	0.28
Reading	58	0.60	0.38
Reading	59	0.61	0.26
Reading	60	0.38	0.31
Reading	61	0.77	0.50
Reading	62	0.50	0.47
Reading	63	0.76	0.53
Reading	64	0.28	0.24
Reading	65	0.46	0.32
Reading	66	0.37	0.16
Reading	67	0.51	0.32
Reading	68	0.28	0.19
Reading	69	0.26	0.22
Reading	70	0.37	0.11
Reading	71	0.42	0.21
Reading	72	0.39	0.14
Reading	73	2.70	0.71
Writing	74	2.57	0.72
Writing	75	1.85	0.52
Speaking	76	1.90	0.55
Speaking	77	1.76	0.64
Speaking	78	1.84	0.61
Speaking	79	1.81	0.63
Speaking	80	1.61	0.72

Table C10: Form C Grade 9 (*N* = 4,090) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.51	0.73
Speaking	82	1.63	0.73
Speaking	83	1.72	0.75
Speaking	84	1.51	0.71
Speaking	85	3.09	0.80
Speaking	86	3.02	0.80
Speaking	87	1.69	0.73
Speaking	88	1.71	0.73
Speaking	89	1.72	0.76
Speaking	90	1.76	0.71
Speaking	91	1.74	0.71

Table C11: Form C Grade 10 (*N* = 3,265)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.83	0.51
Listening	2	0.95	0.48
Listening	3	0.95	0.39
Listening	4	0.90	0.42
Listening	5	0.77	0.28
Listening	6	0.46	0.24
Listening	7	0.73	0.47
Listening	8	0.78	0.35
Listening	9	0.75	0.44
Listening	10	0.51	0.21
Listening	11	0.43	0.22
Listening	12	0.67	0.47
Listening	13	0.40	0.19
Listening	14	0.57	0.32
Listening	15	0.76	0.52
Listening	16	0.79	0.44
Listening	17	0.30	0.07
Listening	18	0.77	0.50
Listening	19	0.64	0.23
Listening	20	0.81	0.43
Writing Conventions	21	0.98	0.27
Writing Conventions	22	0.96	0.41
Writing Conventions	23	0.92	0.48
Writing Conventions	24	0.96	0.29
Writing Conventions	25	0.98	0.26
Writing Conventions	26	0.40	0.19
Writing Conventions	27	0.84	0.40
Writing Conventions	28	0.88	0.35
Writing Conventions	29	0.50	0.16
Writing Conventions	30	0.26	0.01
Writing Conventions	31	0.69	0.32
Writing Conventions	32	0.85	0.53
Writing Conventions	33	0.52	0.37
Writing Conventions	34	0.79	0.30
Writing Conventions	35	0.71	0.53
Writing Conventions	36	0.64	0.46
Writing Conventions	37	0.72	0.40
Writing Conventions	38	0.76	0.58
Writing Conventions	39	0.72	0.55
Writing Conventions	40	0.51	0.35

Table C11: Form C Grade 10 (*N* = 3,265) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.74	0.58
Writing Conventions	42	0.56	0.40
Writing Conventions	43	0.75	0.36
Reading	44	0.31	0.16
Reading	45	0.99	0.12
Reading	46	0.99	0.22
Reading	47	0.99	0.22
Reading	48	0.94	0.46
Reading	49	0.89	0.43
Reading	50	0.77	0.26
Reading	51	0.57	0.42
Reading	52	0.88	0.39
Reading	53	0.74	0.46
Reading	54	0.69	0.49
Reading	55	0.86	0.59
Reading	56	0.86	0.55
Reading	57	0.53	0.34
Reading	58	0.66	0.34
Reading	59	0.65	0.22
Reading	60	0.48	0.39
Reading	61	0.85	0.51
Reading	62	0.58	0.50
Reading	63	0.83	0.52
Reading	64	0.35	0.27
Reading	65	0.56	0.36
Reading	66	0.39	0.13
Reading	67	0.61	0.37
Reading	68	0.34	0.22
Reading	69	0.32	0.27
Reading	70	0.43	0.15
Reading	71	0.49	0.23
Reading	72	0.42	0.16
Reading	73	2.91	0.67
Writing	74	2.78	0.68
Writing	75	1.88	0.48
Speaking	76	1.93	0.47
Speaking	77	1.80	0.59
Speaking	78	1.87	0.55
Speaking	79	1.85	0.56
Speaking	80	1.70	0.67

Table C11: Form C Grade 10 (*N* = 3,265) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.61	0.69
Speaking	82	1.70	0.68
Speaking	83	1.78	0.68
Speaking	84	1.59	0.68
Speaking	85	3.25	0.77
Speaking	86	3.17	0.77
Speaking	87	1.76	0.69
Speaking	88	1.78	0.68
Speaking	89	1.82	0.66
Speaking	90	1.84	0.61
Speaking	91	1.81	0.67

Table C12: Form C Grade 11 (*N* = 2,610)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.80	0.42
Listening	2	0.96	0.40
Listening	3	0.95	0.37
Listening	4	0.91	0.39
Listening	5	0.76	0.24
Listening	6	0.49	0.24
Listening	7	0.76	0.42
Listening	8	0.81	0.34
Listening	9	0.76	0.41
Listening	10	0.56	0.17
Listening	11	0.42	0.20
Listening	12	0.69	0.40
Listening	13	0.41	0.17
Listening	14	0.59	0.33
Listening	15	0.77	0.46
Listening	16	0.80	0.44
Listening	17	0.30	0.10
Listening	18	0.78	0.44
Listening	19	0.64	0.20
Listening	20	0.81	0.38
Writing Conventions	21	0.98	0.26
Writing Conventions	22	0.97	0.37
Writing Conventions	23	0.94	0.41
Writing Conventions	24	0.96	0.27
Writing Conventions	25	0.98	0.22
Writing Conventions	26	0.41	0.21
Writing Conventions	27	0.84	0.39
Writing Conventions	28	0.90	0.33
Writing Conventions	29	0.53	0.21
Writing Conventions	30	0.28	0.05
Writing Conventions	31	0.72	0.29
Writing Conventions	32	0.87	0.50
Writing Conventions	33	0.56	0.37
Writing Conventions	34	0.82	0.29
Writing Conventions	35	0.70	0.47
Writing Conventions	36	0.64	0.42
Writing Conventions	37	0.71	0.37
Writing Conventions	38	0.74	0.53
Writing Conventions	39	0.74	0.50
Writing Conventions	40	0.49	0.37

Table C12: Form C Grade 11 (*N* = 2,610) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.74	0.55
Writing Conventions	42	0.52	0.34
Writing Conventions	43	0.75	0.33
Reading	44	0.33	0.19
Reading	45	0.99	0.12
Reading	46	0.99	0.24
Reading	47	0.99	0.22
Reading	48	0.96	0.39
Reading	49	0.92	0.33
Reading	50	0.78	0.19
Reading	51	0.57	0.41
Reading	52	0.90	0.35
Reading	53	0.75	0.43
Reading	54	0.70	0.45
Reading	55	0.88	0.52
Reading	56	0.88	0.51
Reading	57	0.57	0.29
Reading	58	0.70	0.33
Reading	59	0.69	0.19
Reading	60	0.50	0.38
Reading	61	0.87	0.44
Reading	62	0.60	0.46
Reading	63	0.84	0.44
Reading	64	0.40	0.28
Reading	65	0.60	0.31
Reading	66	0.40	0.10
Reading	67	0.67	0.37
Reading	68	0.38	0.25
Reading	69	0.34	0.34
Reading	70	0.45	0.12
Reading	71	0.56	0.31
Reading	72	0.46	0.14
Reading	73	3.00	0.61
Writing	74	2.90	0.61
Writing	75	1.88	0.38
Speaking	76	1.93	0.41
Speaking	77	1.82	0.51
Speaking	78	1.88	0.47
Speaking	79	1.88	0.48
Speaking	80	1.71	0.58

Table C12: Form C Grade 11 (*N* = 2,610) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.62	0.61
Speaking	82	1.71	0.60
Speaking	83	1.81	0.61
Speaking	84	1.62	0.63
Speaking	85	3.29	0.70
Speaking	86	3.21	0.70
Speaking	87	1.79	0.59
Speaking	88	1.82	0.60
Speaking	89	1.85	0.59
Speaking	90	1.87	0.55
Speaking	91	1.84	0.58

Table C13: Form C Grade 12 (*N* = 1,762)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Listening	1	0.77	0.39
Listening	2	0.96	0.33
Listening	3	0.96	0.32
Listening	4	0.91	0.33
Listening	5	0.73	0.22
Listening	6	0.48	0.25
Listening	7	0.76	0.43
Listening	8	0.80	0.33
Listening	9	0.77	0.38
Listening	10	0.55	0.21
Listening	11	0.45	0.22
Listening	12	0.67	0.42
Listening	13	0.41	0.20
Listening	14	0.59	0.35
Listening	15	0.77	0.49
Listening	16	0.81	0.44
Listening	17	0.30	0.08
Listening	18	0.77	0.43
Listening	19	0.62	0.20
Listening	20	0.81	0.36
Writing Conventions	21	0.99	0.38
Writing Conventions	22	0.98	0.37
Writing Conventions	23	0.94	0.42
Writing Conventions	24	0.96	0.31
Writing Conventions	25	0.98	0.34
Writing Conventions	26	0.41	0.21
Writing Conventions	27	0.88	0.39
Writing Conventions	28	0.89	0.34
Writing Conventions	29	0.56	0.26
Writing Conventions	30	0.29	0.06
Writing Conventions	31	0.74	0.28
Writing Conventions	32	0.88	0.48
Writing Conventions	33	0.56	0.37
Writing Conventions	34	0.82	0.35
Writing Conventions	35	0.68	0.48
Writing Conventions	36	0.61	0.38
Writing Conventions	37	0.69	0.39
Writing Conventions	38	0.74	0.52
Writing Conventions	39	0.70	0.48
Writing Conventions	40	0.48	0.34

Table C13: Form C Grade 12 (*N* = 1,762) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Writing Conventions	41	0.72	0.51
Writing Conventions	42	0.50	0.34
Writing Conventions	43	0.75	0.31
Reading	44	0.34	0.21
Reading	45	0.99	0.18
Reading	46	0.99	0.30
Reading	47	0.99	0.37
Reading	48	0.95	0.37
Reading	49	0.91	0.33
Reading	50	0.78	0.25
Reading	51	0.60	0.41
Reading	52	0.91	0.35
Reading	53	0.75	0.42
Reading	54	0.69	0.43
Reading	55	0.88	0.54
Reading	56	0.89	0.52
Reading	57	0.57	0.36
Reading	58	0.66	0.35
Reading	59	0.71	0.25
Reading	60	0.51	0.40
Reading	61	0.87	0.45
Reading	62	0.61	0.46
Reading	63	0.84	0.44
Reading	64	0.40	0.28
Reading	65	0.60	0.37
Reading	66	0.37	0.15
Reading	67	0.69	0.38
Reading	68	0.39	0.28
Reading	69	0.34	0.34
Reading	70	0.47	0.16
Reading	71	0.59	0.32
Reading	72	0.47	0.18
Reading	73	3.06	0.61
Writing	74	2.91	0.59
Writing	75	1.87	0.41
Speaking	76	1.92	0.43
Speaking	77	1.78	0.51
Speaking	78	1.85	0.50
Speaking	79	1.85	0.52
Speaking	80	1.69	0.56

Table C13: Form C Grade 12 (*N* = 1,762) (continued)

Modality	Item Sequence	Item Mean	Item-Total Correlation
Speaking	81	1.62	0.58
Speaking	82	1.70	0.54
Speaking	83	1.82	0.59
Speaking	84	1.59	0.59
Speaking	85	3.23	0.69
Speaking	86	3.18	0.68
Speaking	87	1.80	0.55
Speaking	88	1.82	0.55
Speaking	89	1.84	0.55
Speaking	90	1.87	0.54
Speaking	91	1.85	0.55

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 2008 WLPT-II Form C Overall Performance Level Cut Scores

	Raw Score			Scale Score			Theta		
Grade	I	A	T	I	A	T	I	A	T
K	29	59	75	509	566	595	-2.6125	-1.0394	-0.2497
1	37	70	90	527	586	627	-2.1449	-0.5041	0.6271
2	46	79	98	544	603	650	-1.6735	-0.0365	1.2481
3	29	61	85	560	619	669	-1.2178	0.3927	1.8121
4	35	69	91	572	634	686	-0.8941	0.8218	2.2691
5	42	74	96	584	644	703	-0.5437	1.1083	2.7303
6	35	71	98	595	654	713	-0.2599	1.3936	3.0227
7	39	75	101	602	662	722	-0.0737	1.5969	3.2762
8	43	78	103	608	668	729	0.1069	1.7549	3.4642
9	36	73	99	614	672	731	0.2694	1.8937	3.4717
10	37	74	100	616	675	733	0.3157	1.9427	3.5543
11	38	74	101	617	675	736	0.3613	1.9427	3.6403
12	38	76	102	617	678	740	0.3613	2.0426	3.7298

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 B 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]	77	814	635	305	347	557.56	37.72
	Listening	20	738	678	316	347	559.50	55.62
	Reading	21	770	770	400	347	525.86	54.86
	Speaking	17	735	735	372	347	574.89	58.56
	Writing	19	784	671	414	347	557.96	37.02
	Comprehension [e]	41	783	672	313	347	542.21	46.11
	Social [f]	37	763	681	309	347	567.85	50.25
	Academic [g]	40	803	697	380	347	547.08	40.78
	Productive [h]	25	788	659	439	343	572.46	40.42
1	Composite [d]	77	814	700	305	336	600.78	45.06
	Listening	20	738	738	316	336	591.75	52.64
	Reading	21	770	717	400	336	588.61	49.82
	Speaking	17	735	735	372	336	609.69	68.03
	Writing	19	784	784	414	336	616.23	43.63
	Comprehension [e]	41	783	686	313	336	588.99	45.83
	Social [f]	37	763	763	309	336	600.32	55.03
	Academic [g]	40	803	697	380	336	605.07	41.31
	Productive [h]	25	788	788	369	335	612.70	50.06
2	Composite [d]	77	814	736	305	283	621.48	45.19
	Listening	20	738	678	316	283	606.73	49.40
	Reading	21	770	770	400	283	620.29	51.01
	Speaking	17	735	735	372	283	623.36	63.37
	Writing	19	784	784	414	283	642.65	51.41
	Comprehension [e]	41	783	703	313	283	614.41	45.66
	Social [f]	37	763	710	309	283	612.54	50.12
	Academic [g]	40	803	752	380	283	631.52	45.71
	Productive [h]	25	788	788	470	281	631.90	48.02
3	Composite [d]	81	845	732	365	212	633.92	40.03
	Listening	20	785	732	428	212	635.07	45.00
	Reading	22	806	806	414	212	630.04	44.99
	Speaking	17	764	764	406	212	653.71	63.89
	Writing	22	806	724	420	212	632.92	41.12
	Comprehension [e]	42	822	744	395	212	632.19	39.76
	Social [f]	37	802	750	390	212	639.07	45.48
	Academic [g]	44	831	725	391	212	631.58	39.59
	Productive [h]	19	802	802	495	205	644.70	43.99
4	Composite [d]	81	845	724	477	172	641.63	45.88
	Listening	20	785	732	482	172	641.78	51.02
	Reading	22	806	806	469	172	642.17	51.83
	Speaking	17	764	764	406	172	663.11	73.35
	Writing	22	806	806	420	172	641.30	53.88
	Comprehension [e]	42	822	728	475	172	641.31	45.81
	Social [f]	37	802	750	440	172	646.10	53.36
	Academic [g]	44	831	780	471	172	641.22	47.87
	Productive [h]	19	802	802	480	162	655.86	49.44

[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 B 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]	81	845	732	365	131	647.05	54.17
	Listening	20	785	785	428	131	648.17	56.14
	Reading	22	806	806	414	131	649.02	61.77
	Speaking	17	764	764	406	131	662.85	87.09
5	Writing	22	806	806	420	131	651.73	58.71
	Comprehension [e]	42	822	744	395	131	647.27	54.50
	Social [f]	37	802	802	390	131	650.05	61.85
	Academic [g]	44	831	753	391	131	648.70	55.84
	Productive [h]	19	802	802	406	125	657.70	70.31
	Composite [d]	91	881	768	376	150	659.98	59.56
	Listening	20	823	823	411	150	665.74	62.99
	Reading	28	834	834	446	150	656.82	64.68
	Speaking	17	793	793	422	150	672.94	96.85
6	Writing	26	850	798	438	150	664.67	59.00
	Comprehension [e]	48	854	776	399	150	658.87	59.43
	Social [f]	37	836	836	391	150	664.21	71.81
	Academic [g]	54	868	774	416	150	660.61	58.57
	Productive [h]	19	825	825	422	133	678.21	66.10
	Composite [d]	91	881	768	511	115	668.79	52.17
	Listening	20	823	823	500	115	671.03	58.97
	Reading	28	834	782	500	115	669.37	53.14
	Speaking	17	793	793	499	115	690.53	88.04
7	Writing	26	850	770	493	115	673.15	48.77
	Comprehension [e]	48	854	759	514	115	668.59	50.20
	Social [f]	37	836	836	510	115	673.41	65.56
	Academic [g]	54	868	774	514	115	670.94	46.92
	Productive [h]	19	825	825	499	106	686.49	63.51
	Composite [d]	91	881	788	517	130	683.50	53.47
	Listening	20	823	769	524	130	683.77	54.19
	Reading	28	834	834	446	130	684.00	62.84
	Speaking	17	793	793	422	130	695.90	86.78
8	Writing	26	850	850	438	130	688.77	55.79
	Comprehension [e]	48	854	803	514	130	682.41	51.99
	Social [f]	37	836	784	510	130	684.62	60.61
	Academic [g]	54	868	790	416	130	685.14	55.12
	Productive [h]	19	825	825	515	118	708.58	55.35

[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 B 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]	92	890	743	492	209	666.89	47.93
	Listening	20	824	824	487	209	667.43	50.86
	Reading	30	848	848	528	209	666.93	49.45
	Speaking	17	809	809	419	209	689.38	88.96
	Writing	25	859	745	456	209	667.55	44.75
	Comprehension [e]	50	863	769	524	209	666.73	44.74
	Social [f]	37	842	791	477	209	670.96	60.98
	Academic [g]	55	879	748	508	209	667.28	42.01
	Productive [h]	19	831	831	418	192	684.19	65.71
10	Composite [d]	92	890	777	566	134	685.10	41.44
	Listening	20	824	824	519	134	679.88	44.44
	Reading	30	848	848	507	134	683.99	51.94
	Speaking	17	809	809	534	134	714.64	73.97
	Writing	25	859	806	586	134	685.85	41.74
	Comprehension [e]	50	863	785	552	134	681.69	41.74
	Social [f]	37	842	842	542	134	690.70	50.93
	Academic [g]	55	879	801	580	134	684.66	41.99
	Productive [h]	19	831	831	588	129	708.12	58.42
11	Composite [d]	92	890	763	583	97	688.21	37.62
	Listening	20	824	824	559	97	687.26	48.43
	Reading	30	848	796	586	97	688.96	42.31
	Speaking	17	809	809	534	97	718.99	72.11
	Writing	25	859	777	611	97	685.35	37.28
	Comprehension [e]	50	863	769	587	97	687.22	38.39
	Social [f]	37	842	842	542	97	697.53	53.09
	Academic [g]	55	879	763	605	97	686.24	35.10
	Productive [h]	19	831	831	596	90	713.60	50.15
12	Composite [d]	92	890	757	589	63	692.78	33.79
	Listening	20	824	824	559	63	692.41	51.95
	Reading	30	848	796	546	63	697.27	47.56
	Speaking	17	809	809	562	63	730.27	71.99
	Writing	25	859	759	575	63	687.10	35.04
	Comprehension [e]	50	863	812	552	63	694.56	44.71
	Social [f]	37	842	842	573	63	701.27	48.04
	Academic [g]	55	879	748	596	63	691.16	35.22
	Productive [h]	19	831	831	607	61	714.52	51.44

[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 B

Grade	Beginner/ Advanced Beginner	Intermediate	Advanced	Transitional
K	9	46	31	14
1	8	21	43	29
2	6	17	53	24
3	6	20	59	16
4	10	24	52	14
5	11	20	60	10
6	17	21	43	19
7	13	21	49	17
8	10	17	55	18
9	17	25	55	3
10	8	25	54	9
11	5	24	63	8
12	3	22	68	6

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