# TABLE OF CONTENTS

TABLE OF CONTENTS ........................................................................................................ III
LIST OF TABLES .................................................................................................................. V
LIST OF FIGURES ................................................................................................................. VII
ABBREVIATIONS AND GLOSSARY ..................................................................................... VIII
PURPOSE OF THE TECHNICAL REPORT ............................................................................. 1

PART 1: OVERVIEW OF THE STATE ASSESSMENT PROGRAM ......................................... 2

ELEMENTS OF THE WASHINGTON ASSESSMENT SYSTEM .............................................. 3
State-Level Assessments in Reading, Writing, Mathematics, and Science .......................... 3
Classroom-Based Assessment .......................................................................................... 4
Professional Development ............................................................................................. 5
Certificate of Academic Achievement .......................................................................... 5
School and District Accountability System .................................................................. 5
Components of the Alternate Assessment System ....................................................... 5

CRITERION-REFERENCED TESTING ............................................................................ 7

APPROPRIATE USE OF TEST SCORES ........................................................................ 8

DESCRIPTION OF THE 2009 TESTS ........................................................................... 9

SCHEDULE FOR TESTING – GRADE 4 - SPRING 2009 ............................................. 11

SUMMARY ....................................................................................................................... 11

PART 2: TEST DEVELOPMENT ...................................................................................... 12

ITEM AND TEST SPECIFICATIONS ............................................................................... 12

CONTENT REVIEWS & BIAS AND FAIRNESS REVIEWS .................................................. 16

ITEM PILOTS .................................................................................................................. 17

CALIBRATION, SCALING, AND ITEM ANALYSIS ......................................................... 18

Traditional Item Analysis ............................................................................................ 18
IRT Analysis .................................................................................................................. 19
Bias Analysis ................................................................................................................. 21
Data Reviews ............................................................................................................... 24
Item Selection ............................................................................................................... 24

TEST CONSTRUCTION ............................................................................................... 24

PART 3: VALIDITY ......................................................................................................... 27

PURPOSE OF THE TEST ............................................................................................... 27

The Constructs To Be Measured .................................................................................. 27
The Scores Generated and the Interpretations and Uses of These Scores ................. 28

EVIDENCE BASED ON TEST CONTENT .................................................................. 28

EVIDENCE BASED ON INTERNAL STRUCTURE ....................................................... 28

Correlations Among WASI Strand Scores .................................................................. 29
Factor Analysis of Strand Scores .............................................................................. 31
Examining Construct Validity Through Confirmatory Factor Analysis .................... 33

PERFORMANCE IN DIFFERENT POPULATIONS ...................................................... 38

ADDITIONAL VALIDITY EVIDENCE ........................................................................ 38

SUMMARY ....................................................................................................................... 38

PART 4: RELIABILITY ..................................................................................................... 40

INTERNAL CONSISTENCY ......................................................................................... 40

STANDARD ERROR OF MEASUREMENT .................................................................. 41

INTERJUDGE AGREEMENT ....................................................................................... 43

DECISION CONSISTENCY AND ACCURACY ............................................................ 44
LIST OF TABLES

Table 1. 2009 Grade 4 Reading Items - Content Classification ........................................... 9
Table 2. 2009 Grade 4 Writing Prompts - Content Classification ........................................... 9
Table 3. 2009 Grade 4 Mathematics Items - Content Classification ....................................... 10
Table 4. 2009 Grade 4 State Standardized Testing Schedule .................................................. 11
Table 5. Grade 4 Reading Test Design ...................................................................................... 14
Table 6. Grade 4 Writing Test Design ...................................................................................... 14
Table 7. Grade 4 Mathematics Test Design ................................................................................ 15
Table 8. DIF Categories for Multiple-Choice Items ................................................................ 23
Table 9. DIF Categories for Constructed-Response Items ........................................................ 23
Table 10. 2009 Grade 4 Strand Score Intercorrelations ......................................................... 31
Table 11. 2009 Grade 4 Rotated Factor Pattern on Tests for Three-Factor Solution .................. 33
Table 12. 2009 Grade 4 Model Goodness-of-Fit Statistics ...................................................... 37
Table 13. 2009 Grade 4 Test & Content Strand Reliability Estimates ...................................... 42
Table 14. 2009 Grade 4 Summary Decision Consistency & Accuracy Index ............................ 46
Table 15. Theta to Scaled Score Linear Transformation Equations ............................................ 50
Table 16. Scaled Score Ranges for Performance Level Categories ............................................ 50
Table 17. Empirical Weighted Mean Rasch of 2000 – 2009 Grade 4 Reading & Mathematics Tests ......................................................................................................................... 54
Table 18. 2009 Grade 4 Reading Raw Score (RAW) to Scaled Scores (SS) with Conditional Standard Error of Measurement (SEM) .................................................................................. 56
Table 19. 2009 Grade 4 Mathematics Raw Score (RAW) to Scaled Scores (SS) with Conditional Standard Error of Measurement (SEM) ................................................................. 57
Table 20. 2009 Grade 4 Mathematics Braille Raw Score (RAW) to Scaled Scores (SS) with Conditional Standard Error of Measurement (SEM) ....................................................... 58
Table 21. 2009 Grade 4 Mathematics Large Print Raw Score (RAW) to Scaled Scores (SS) with Conditional Standard Error of Measurement (SEM) .................................................... 58
Table 22. 2009 Grade 4 Writing Raw Scores (RAW) with Conditional Standard Errors of Measurement (SEM) ....................................................................................................................... 59
Table 23. 2009 Grade 4 Content Strand Cut-Points ................................................................. 59
Table 24. 2009 Grade 4 Mathematics Braille Content Strand Cut-Points ................................... 60
Table 25. 2009 Grade 4 Mathematics Large Print Content Strand Cut-Points ........................... 60
Table 26. 2009 Grade 4 Reading - Interrater Percent Agreement ............................................. 69
Table 27. 2009 Grade 4 Writing - Interrater Percent Agreement .............................................. 70
Table 28. 2009 Grade 4 Mathematics - Interrater Percent Agreement ....................................... 70
Table 29. 2009 Grade 4 Means & Standard Deviations (SD) Test Scores ............................... 71
Table 30. 2009 Grade 4 Raw Test Score Summaries, Percent Students With Strength in Strand ................................................................................................................................. 72
Table 31. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Gender ................................................................................................................................. 72
Table 32. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Ethnic Group ................................................................................................................................. 72
Table 33. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Categorical Program ................................................................................................................................. 73
Table 34. 2009 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Gender ................................................................................................................................. 73
Table 35. 2009 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Ethnic Group ................................................................................................................................. 73
Table 36. 2009 Grade 4 Writing – Raw Score Means & Standard Deviations (SD) by Categorical Program ................................................................................................................................. 74
Table 37. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Gender ................................................................................................................................. 74
Table 38. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Ethnic Group ................................................................................................................................. 74
Table 39. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Categorical Program ................................................................................................................................. 75
Table 40. 2009 Grade 4 Reading – Percent Meeting Standards by Gender .................................. 76
Table 41. 2009 Grade 4 Reading – Percent Meeting Standards by Ethnic Group ...................... 77
Table 42. 2009 Grade 4 Reading – Percent Meeting Standards by Categorical Program .......... 77
TABLE 43. 2009 Grade 4 Writing – Percent Meeting Standards by Gender ........................................ 78
TABLE 44. 2009 Grade 4 Writing – Percent Meeting Standards by Ethnic Group ............................ 78
TABLE 45. 2009 Grade 4 Writing – Percent Meeting Standards by Categorical Program .................... 79
TABLE 46. 2009 Grade 4 Mathematics – Percent Meeting Standards by Gender ............................. 79
TABLE 47. 2009 Grade 4 Mathematics – Percent Meeting Standards by Ethnic Group ..................... 80
TABLE 48. 2009 Grade 4 Mathematics – Percent Meeting Standards by Categorical Program .......... 80
TABLE 49. Grade 4 Percentage of Students Meeting Standard from 1996-97 through 2008-09 ............ 81
TABLE 50. 2009 Grade 4 Summary Statistics – Operational Items ................................................... 83
TABLE 51. 2009 Grade 4 Omit Rate – Operational Item Statistics ....................................................... 83
TABLE 52. 2009 Grade 4 Rasch Item Difficulty Summary Statistics – Operational Items ...................... 84
TABLE 53. 2009 Grade 4 Writing – Operational Item Statistics ........................................................ 84
TABLE 54. 2009 Grade 4 Reading – Operational Item Statistics .......................................................... 85
TABLE 55. 2009 Grade 4 Mathematics – Operational Item Statistics .................................................. 86
TABLE 56. 2009 Grade 4 Reading Test Reliability Estimates ............................................................... 88
TABLE 57. 2009 Grade 4 Writing Test Reliability Estimates ............................................................... 89
TABLE 58. 2009 Grade 4 Mathematics Test Reliability Estimates ....................................................... 90
TABLE 59. 2009 Grade 4 Summary Statistics for Embedded Pilot Items ........................................... 93
TABLE 60. 2009 Grade 4 Omit Rate Statistics for Embedded Pilot Items ........................................ 93
TABLE 61. 2009 Grade 4 Rasch Item Difficulty Summary Statistics for Embedded Pilot Items ........ 94
TABLE 62. 2009 Grade 4 Reading Differential Item Functioning Summary Statistics for Embedded Pilot Items ........................................................................................................... 95
TABLE 63. 2009 Grade 4 Mathematics Differential Item Functioning Summary Statistics for Embedded Pilot Items ........................................................................................................... 96
TABLE 64. 2009 Grade 4 Summary of Items Flagged for Embedded Pilot Items ................................. 97
LIST OF FIGURES

FIGURE 1. LOCATION OF EXAMINEE $\beta_1$ ON TWO TESTS WITH DIFFERENT ITEMS .......................................................... 20
FIGURE 2. LOCATION OF EXAMINEE $\beta_1$ ON THE SAME “MATHEMATICS TEST” SCALE ............................................. 20
FIGURE 3. GRADE 4 SCREE PLOT OF EIGENVALUES ........................................................................................................ 32
FIGURE 4. SINGLE-FACTOR STRUCTURAL MODEL ............................................................................................................. 34
FIGURE 5. MULTI-FACTOR STRUCTURAL MODEL ............................................................................................................... 35
FIGURE 6. ACCURACY OR CONSISTENCY AROUND CRITICAL CUT POINT ................................................................. 45
FIGURE 7. HYPOTHETICAL RANGE OF MATHEMATICS STRAND ITEM DIFFICULTIES (8) ........................................... 51
FIGURE 8. SAMPLE SCORE DISTRIBUTION OF CONTRASTING GROUPS – COS STRAND ........................................... 52
FIGURE 9. GRADE 4 RESULTS FOR 1996-97 THROUGH 2008-09 BY CONTENT AREA .................................................. 81
# ABBREVIATIONS AND GLOSSARY

<table>
<thead>
<tr>
<th>Abbreviation or Term</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALY (RD)</td>
<td>Analysis strand</td>
</tr>
<tr>
<td>AP (SC)</td>
<td>Application of Science strand</td>
</tr>
<tr>
<td>AS</td>
<td>Algebraic Sense (content) Mathematics strand</td>
</tr>
<tr>
<td>CMP (RD)</td>
<td>Comprehension strand</td>
</tr>
<tr>
<td>CONV</td>
<td>Writing Conventions strand</td>
</tr>
<tr>
<td>COS</td>
<td>Content, Organization, &amp; Style Writing strand</td>
</tr>
<tr>
<td>CSEM</td>
<td>Conditional standard error of measurement</td>
</tr>
<tr>
<td>CT (RD)</td>
<td>Critical Thinking strand</td>
</tr>
<tr>
<td>CT (MA)</td>
<td>Content strand</td>
</tr>
<tr>
<td>CU</td>
<td>Communicates Understanding (process) Mathematics strand</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential item functioning</td>
</tr>
<tr>
<td>DRC</td>
<td>Data Recognition Corporation</td>
</tr>
<tr>
<td>EALR</td>
<td>Essential Academic Learning Requirements</td>
</tr>
<tr>
<td>ETS</td>
<td>Educational Testing Service</td>
</tr>
<tr>
<td>FORM</td>
<td>Operational items and imbedded pilot items that uniquely define a (test) form</td>
</tr>
<tr>
<td>GLE</td>
<td>Grade Level Expectation</td>
</tr>
<tr>
<td>GS</td>
<td>Geometric Sense (content) Mathematics strand</td>
</tr>
<tr>
<td>IA</td>
<td>Informational Analysis Reading strand</td>
</tr>
<tr>
<td>IC</td>
<td>Informational Comprehension Reading strand</td>
</tr>
<tr>
<td>IEP</td>
<td>Individual Education Program</td>
</tr>
<tr>
<td>IN (SC)</td>
<td>Inquiry in Science strand</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>IT</td>
<td>Informational Thinking Critically Reading strand</td>
</tr>
<tr>
<td>LA</td>
<td>Literary Analysis Reading strand</td>
</tr>
<tr>
<td>LC</td>
<td>Literary Comprehension Reading strand</td>
</tr>
<tr>
<td>LT</td>
<td>Literary Thinking Critically Reading strand</td>
</tr>
<tr>
<td>MC</td>
<td>Makes Connections (process) Mathematics strand</td>
</tr>
<tr>
<td>ME</td>
<td>Measurement (content) Mathematics strand</td>
</tr>
<tr>
<td>NS</td>
<td>Number Sense (content) Mathematics strand</td>
</tr>
<tr>
<td>OSPI</td>
<td>Office of the Superintendent of Public Instruction</td>
</tr>
<tr>
<td>PAS</td>
<td>Performance Assessment Services (DRC)</td>
</tr>
<tr>
<td>PC (MA)</td>
<td>Process strand</td>
</tr>
<tr>
<td>PCM</td>
<td>Partial Credit Model</td>
</tr>
<tr>
<td>PS</td>
<td>Probability and Statistics (content) Mathematics strand</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard Error of Measurement</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SR</td>
<td>Solves Problems &amp; Reasons Logically (process) Mathematics strand</td>
</tr>
<tr>
<td>SYS (SC)</td>
<td>Systems of Science strand</td>
</tr>
<tr>
<td>TEST</td>
<td>Operational test items in a testing booklet that contribute to reported student scores</td>
</tr>
<tr>
<td>WAAS</td>
<td>Washington Alternate Assessment System</td>
</tr>
<tr>
<td>WASL</td>
<td>Washington Assessment of Student Learning</td>
</tr>
</tbody>
</table>
PURPOSE OF THE TECHNICAL REPORT

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) identifies professional standards, criteria, and recommendations for test developers and test publishers. One of those standards is to provide sufficient documentation that enables potential test users to evaluate the quality of a test, including evidence for the reliability and validity of test scores. This annual technical report follows the format and composition of previous technical reports and is one component of a suite of reports that documents the properties and characteristics of the 2009 Washington Assessment of Student Learning Grade 4 Assessment for Reading, Writing, and Mathematics.

Unless otherwise noted, the analysis results and summaries about test performance are derived from the most recently available statewide student data file. Inclusion and exclusion rules to aggregate the data for purposes of these analyses may not necessarily coincide with the rules applied to produce operationally published score reports.
PART 1: OVERVIEW OF THE STATE ASSESSMENT PROGRAM

In 1993, Washington State embarked on the development of a comprehensive school change effort with the primary goal to improve teaching and learning. Created by the state legislature in 1993 and sunset in 1999, the Commission on Student Learning was charged with three important tasks to support this school change effort.

• Establish Essential Academic Learning Requirements (EALRs) that describe what all students should know and be able to do in eight content areas—Reading, Writing, Communication, Mathematics, Science, Health/Fitness, Social Studies, and the Arts.

• Develop an assessment system to measure student progress at three grade levels towards achieving the EALRs.

• Recommend an accountability system that recognizes and rewards successful schools and provides support and assistance to less successful schools.

The EALRs in Reading, Writing, Communications, and Mathematics were adopted in 1995 and revised in 1997. The EALRs for Science, Social Studies, Health/Fitness, and the Arts were adopted in 1996 and revised in 1997. (See http://www.k12.wa.us/curriculuminstruct for links to the EALRs and Grade Level Expectations (GLEs) in all subject areas.) Performance “benchmarks” were previously established at three grade levels – elementary (Grade 4), middle (Grade 7), and high school (Grade 10).

The assessments for Reading, Writing, and Mathematics were developed at Grades 4 and 7 and were operationalized in Spring 1998. The Grade 10 assessment in these same content areas was pilot-tested in Spring 1998, and was operationalized in Spring 1999. Participation in the Grade 4 assessment became mandatory for all public schools in Spring 1998. Participation in the Grade 7 and 10 assessments was voluntary until Spring 2000. Participation in the Grade 3, 5, 6, and 8 Reading and Mathematics assessments were voluntary in 2004 and 2005, and became mandatory for first operational administration in Spring 2006.

Science was implemented as a voluntary operational administration for Grades 8 and 10 in Spring 2003 and became mandatory in 2004. Grade 5 Science was a voluntary operational administration in Spring 2004 with mandatory implementation in Spring 2005.

During the regular Spring 2005 testing period, Grade 11 students were allowed to retake any of the Grade 10 subject tests on which they had not met standard. Since students at all high school grades will eventually be able to take the tests, the Grade 10 assessments became known as the High School Washington Assessment of Student Learning (WASL). In 2009, Grade 9 students were not allowed to take the WASL assessments.

This report is limited to the results of the students in Grade 4 who took the WASL assessments in Reading, Writing, and Mathematics during Spring 2009.
ELEMENTS OF THE WASHINGTON ASSESSMENT SYSTEM

The assessment system has several major components: state-level assessments, classroom-based assessments, professional development, alternate assessment programs, the Certificate of Academic Achievement, and the Accountability System. The scope and subject of this report is necessarily limited to the technical characteristics of the regular state-level assessments, administered to the majority of students at specified grade levels.

State-Level Assessments in Reading, Writing, Mathematics, and Science

The state-level assessments require students to select and to construct responses to demonstrate their knowledge, skills, and understanding in each of the EALRs – from multiple-choice and short-answer items to extended-responses, essays, and problem solving tasks. Student-, school-, district-, and state-level scores are reported for the operational assessments. The state-level operational test forms are standardized and “on demand,” meaning students are expected to respond to the same items, under the same conditions, and at the same time during the school year.

All of the state-level assessments are untimed; that is, students may have as much time as they reasonably need to complete their work. Guidelines for providing accommodations to students with special needs have been developed to encourage the inclusion of as many students as possible. Special needs students include those in special education programs, English language learners (ELL/bilingual), migrant students, and highly capable students. A broad range of accommodations allows nearly all students access to some or all parts of the assessment. (See Guidelines for Inclusion and Accommodations for Special Populations on State-Level Assessments.)

Classroom teachers and curriculum specialists throughout the State of Washington assisted with the development of all items for the state-level assessments. Content committees were created at each grade level and content area. Working with content and assessment specialists, these committees defined the test and item specifications consistent with the Washington State EALRs, reviewed all items prior to pilot testing, and provided final review and recommendations to approve selected items after pilot testing. A separate “bias and fairness” committee, comprised of individuals that reflect Washington’s diversity, also conducted a sensitivity review of all items for words or content, that might be potentially offensive to students or parents or might disadvantage some students for reasons unrelated to the assessed skill or concept. Part 2 of this report provides further details about the test development process.

Hundreds of items were developed and pilot-tested to populate a pool of items in each grade level and content area. New forms of the assessment are constructed each year with selections from the item pool. Statistical equating procedures are applied to ensure that the same level of performance is required to achieve standard from year to year. The state-level assessments in Reading, Mathematics, and Science include a mix of multiple-choice, short-answer, and extended-response items. The state-level assessments in Writing include two writing prompts in two different modalities, each scored for content and for writing conventions.
Following the first operational administration of each grade level content area assessment, a standard-setting panel recommended the level of performance to meet the standard on the EALRs. Additionally, “progress categories” above and below the standard were recommended in Reading, Mathematics, and Science. At the school and district levels, the percentage of students meeting the standard and in each progress category is reported. In preparation for the implementation of the Certificate of Academic Achievement, the standards for Reading, Writing, and Mathematics were revisited in February and March of 2004. Further details that describe the procedures, outline the recommendations, and summarize the results can be found in the WASL 2004 Report and Results from Revisiting of the Standards for Grades 4/7/10 Reading, Mathematics, and Writing.

Classroom-Based Assessment

There are several important reasons to include classroom-based assessment as part of a comprehensive assessment system. First, classroom-based assessments help students and teachers better understand the EALRs and recognize the characteristics of quality work that define good performance in each content area. Second, classroom-based assessments provide assessment of some of the EALRs for which state-level assessment is not feasible – oral presentations and group discussion, for example. Third, classroom-based assessments offer teachers and students opportunities to gather evidence of student achievement in ways that best fit the needs and interests of individual students. Fourth, classroom-based assessments help teachers become more effective in gathering valid evidence of student learning related to the EALRs. Effective classroom-based assessments can be sensitive to the developmental needs of students and provide the flexibility necessary to accommodate the learning styles of children with special needs. In addition to items that may be on the state-level assessments, classroom-based assessments can provide information from oral interviews and presentations, work products, experiments and projects, or exhibits of student work collected over a week, a month, or the entire school year.

Classroom-based assessment Tool Kits have been developed for the early and middle school years to provide teachers with examples of good assessment strategies. The Tool Kits include models for paper and pencil tasks, generic checklists of skills and traits, observation assessment strategies, simple rating scales, and generic protocols for oral communications and personal interviews. At the upper grades, classroom-based assessment strategies include models for developing and evaluating interdisciplinary performance-based tasks. Additionally, the Tool Kits provide content frameworks to assist teachers at all grade levels to relate their classroom learning goals and instruction to the appropriate grade level EALRs. (For links to the Tool Kits see http://www.k12.wa.us/assessment/toolkits/default.aspx.)
Professional Development

A third major component of the assessment system emphasizes the need for ongoing, comprehensive support and professional training for teachers and administrators to improve their understanding of the EALRs, the characteristics of sound assessments, and effective instructional strategies that will help students meet the standards. The Commission on Student Learning established fifteen “Learning and Assessment Centers” throughout the state. Most are managed through Washington’s nine Educational Service Districts and a few are managed by school district consortia. These Centers provide professional development and support to assist school and district staff:

- Link teaching and curriculum to high academic standards based on the EALRs;
- Learn and apply the principles of good assessment practice;
- Use a variety of assessment techniques and strategies;
- Judge student work by applying explicit scoring rules;
- Make instructional and curricular decisions based on reliable and valid assessment information; and
- Help students and parents understand the EALRs and how students can achieve them.

Certificate of Academic Achievement

Beginning in 2008, graduating seniors may earn a Certificate of Academic Achievement in addition to the high school diploma. The Certificate will serve as evidence that students have achieved Washington’s EALRs by meeting the standards set for the High School Reading and Writing assessments. However, the graduation requirement for meeting the Mathematics assessment standards has been delayed.

School and District Accountability System

The Academic Achievement and Accountability (A+) Commission developed recommendations for a school and district accountability system that recognizes schools who are successful in helping their students achieve the standards on the WASL assessments. These recommendations also address the need for assistance to those schools and districts in which students are not achieving the standards. The A+ Commission was dissolved in 2005 and their duties and responsibilities were transferred to the State Board of Education.

Components of the Alternate Assessment System

State assessment programs provide a vehicle to gauge student academic achievement in an educational system. The Washington State Assessment System provides accountability for instructional programs and educational opportunities for all students, including those receiving
special education services. Alternate assessment is one component of Washington’s comprehensive assessment system.

The Washington Alternate Assessment System (WAAS) program was developed by the Washington Alternate Assessment Task Force and expanded by Advisory Panels in response to requirements of the Individuals with Disabilities Education Act of 1997:

The State has established goals for the performance of children with disabilities in the state that . . . are consistent, to the maximum extent appropriate, and with other goals and standards for children established by the state.

The alternate assessments are based on Washington’s EALRs in the content areas of Reading, Writing, Mathematics, and Science, and in this way, share a foundational link to the regular WASL assessments. The state prepared extensions for the EALRs that describe the critical function of the EALRs, the access skills, instructional activities, and assessment strategies that are designed to assist special education staff members to link functional Individual Education Plan (IEP) skills to the EALRs, to provide access to the general education curriculum, and to measure student progress toward achieving the EALRs.

The WAAS was designed for a small percentage of the total school population. Students with disabilities are expected to take the regular WASL tests, with or without necessary accommodations, unless the IEP team determines a student is unable to participate on one or more content areas of the WASL. In these instances, the IEP team may elect the WAAS portfolio assessment.

The Developmentally Appropriate WASL (DAW) and WASL-BASIC are alternatives to regular WASL administration for eligible students. The WASL-BASIC, previously called the WASL-MO (or WASL-Modified), is intended for students who take the WASL at grade level but the passing score is adjusted by the student’s IEP teams from Proficient (Level 3) to Basic (Level 2). Eligibility criteria, requirements, and resource information can be found at http://www.k12.wa.us/SpecialEd/assessment.aspx.
CRITERION-REFERENCED TESTING

The purpose of an achievement test is to determine how well a student has learned important concepts and skills. Test scores are used to make inferences about students’ overall performance in a particular domain. When we compare a student’s performance to a target performance, this is considered a criterion-referenced interpretation. When we compare a student’s performance relative to the performance of other students, this is considered a norm-referenced interpretation.

Criterion-referenced tests can measure the degree to which students have achieved a desired set of learning targets, conceptual understanding, and skills that are at grade level or developmentally appropriate. Much care and attention is spent to ensure that the items on the test represent only the desired learning targets and that there are sufficient numbers of items for each learning target to make reliable statements about students’ degree of achievement related to that target. After a criterion-referenced test, examinee scores are used to make inferences about whether students have attained the desired level of achievement. Test scores are used to make statements like, “This student meets the minimum mathematics requirements for this class,” or “This student knows how to apply computational skills to solve a complex word problem.”

Norm-referenced tests provide a general measure of some achievement domain relative to the performance of other students, schools, and districts. Much care and attention is spent to create items that vary in difficulty to measure a broad range of ability levels. Items are included on the test that measure below grade level, on grade level, and above grade level concepts and skills. Items are distributed broadly across the domain. While some norm-referenced tests provide objectives-level information, items for each objective may represent concepts and skills that are not easily learned by most students until their later years in school. Examinee scores on a norm-referenced test are compared to the performance of a norm group or a representative group of students of similar age and grade. Norm groups may be local (other students in a district or state) or national (representative samples of students from throughout the United States). Scores on norm-referenced tests are used to make statements like, “This student is the best student in the class,” or “This student knows mathematical concepts better than 75% of the students in the norm group.”

To test all of the desired concepts and skills in a domain, testing time would be inordinate. Well designed state or national achievement tests, whether norm-or criterion-referenced, always include samples from the domain of desired concepts and skills. Therefore, when state or national achievement tests are used, we generalize from a student’s performance on the sample of items in the test and estimate how the student would perform in the overall domain. For a broader measure of student achievement in a specific domain, it is necessary to use more than one assessment. District and classroom assessments are both useful and necessary to supplement information that is derived from state or national achievement tests.

It is possible and sometimes even desirable to have both norm-referenced and criterion-referenced information about students’ performance. The referencing scheme is best determined by the intended use of the test, and this is generally determined by how the test is constructed. If tests are being used to make decisions about the success or the usefulness of an instructional or administrative program, or the degree to which students have attained a set of
desired learning targets, then criterion-referenced tests and interpretations are most useful. If the tests are being used to select students for particular programs or compare students, districts, and states, then norm-referenced tests and interpretations are useful. In some cases, both norm-referenced and criterion-referenced interpretations can be made from the same achievement measures. The WASL state level assessment is a criterion-referenced test. Student performance should be interpreted in terms of how well students have achieved the Washington State EALRs.

**APPROPRIATE USE OF TEST SCORES**

Once tests are administered, total test scaled scores and strand scores are generated for each content area test and the performance is reported at the individual, school, district, and state levels. The total test scale score is used to classify students into performance levels in terms of their level of knowledge and skill in the subject area. Additionally, strand scores are used to draw inferences about a student’s achievement in each of several specific knowledge or skill areas covered by each test. Strand score indicators (+ or -) are reported to provide teachers, parents, and students more detailed information about students’ learning and performance on the test.

The information in these reports (scaled score, performance levels, and strand score indicators) can be used with other assessment information to help with school, district, and state curriculum planning and classroom instructional decisions.

While school and district scores may be useful in curriculum and instructional planning, it is important to exercise extreme caution when interpreting individual reports. The items included on WASL tests are samples from a larger domain. Scores from one test given on a single occasion should never be used to make important decisions about students’ placement, the type of instruction they receive, or retention in a given grade level in school. It is important to corroborate individual scores on WASL tests with classroom-based and other local evidence of student learning (e.g., scores from district testing programs). When making decisions about individuals, multiple sources of information should be used. Multiple individuals who are familiar with the student’s progress and achievement – including parents, teachers, school counselors, school psychologists, specialist teachers, and perhaps the students themselves – should be brought together to collaboratively make such decisions.
DESCRIPTION OF THE 2009 TESTS

The Grade 4 2009 Washington Assessment of Student Learning (WASL) tests measure students’ achievement of the EALRs in Reading, Writing, and Mathematics. Tables 1 to 3 indicate the EALRs measured by each of the three tests, the test “strands,” and the number of items per strand in the 2009 test.

Table 1. 2009 Grade 4 Reading Items - Content Classification

<table>
<thead>
<tr>
<th>Type of Reading Passage</th>
<th>Test Strand</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary ‡</td>
<td>Comprehension †</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Analysis †</td>
<td>8</td>
</tr>
<tr>
<td>Informational ‡</td>
<td>Comprehension †</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Analysis †</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total Number of Items</strong></td>
<td></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>

*Reading EALR 1: The student understands and uses different skills and strategies to read.
†Reading EALR 2: The student understands the meaning of what is read.
‡Reading EALR 3: The student reads different materials for a variety of purposes.

Table 2. 2009 Grade 4 Writing Prompts - Content Classification

<table>
<thead>
<tr>
<th>Task</th>
<th>Purposes 1</th>
<th>Process 2</th>
<th>Number of Prompts</th>
<th>Scores 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended Piece</td>
<td>Narrative</td>
<td>• prewrite• first draft• revise• edit• final draft</td>
<td>1</td>
<td>• Content, Organization &amp; Style• Writing Conventions</td>
</tr>
<tr>
<td></td>
<td>Inform</td>
<td>• prewrite• first draft• revise• edit• final draft</td>
<td>1</td>
<td>• Content, Organization &amp; Style• Writing Conventions</td>
</tr>
<tr>
<td><strong>Total Number of Prompts</strong></td>
<td></td>
<td></td>
<td><strong>2</strong></td>
<td></td>
</tr>
</tbody>
</table>

1 Writing EALR 1: The student writes clearly and effectively (concept & design, style [word choice, sentence fluency, voice], and conventions).
2 Writing EALR 2: The student writes in a variety of forms for different audiences and purposes.
3 Writing EALR 3: The student understands and uses the steps of a writing process.
<table>
<thead>
<tr>
<th>Process Strand</th>
<th>Concept Strand</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts &amp; Procedures</td>
<td>Number Sense $^1$</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Measurement $^1$</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Geometric Sense $^1$</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Probability and Statistics $^1$</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Algebraic Sense $^1$</td>
<td>5</td>
</tr>
<tr>
<td>Solves Problems $^2$ &amp; Reasons Logically $^3$</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Communicates Understanding $^4$</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Making Connections $^5$</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Total Number of Items</strong></td>
<td></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>

$^1$ Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.

$^2$ Mathematics EALR 2: The student uses mathematics to define and solve problems.

$^3$ Mathematics EALR 3 The student uses mathematical reasoning.

$^4$ Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

$^5$ Mathematics EALR 5: The student makes mathematical connections.
SCHEDULE FOR TESTING – GRADE 4 - SPRING 2009

Grade 4 Reading, Writing, and Mathematics tests were administered within the April 13 – May 1 testing window. Specific test administration schedules within that window were determined locally and approved by District Assessment Coordinators. All students within a grade level at a school were required to take the same test on the same day. There were two reading test administration sessions, and the estimated working time for each session was 50 – 70 minutes. There were two writing test administration sessions, and the estimated time for each session was 120 minutes. There were two mathematics test administration sessions, and the estimated working time for each session was 45 – 60 minutes. Table 4 shows the schedule as provided in the Washington Assessment of Student Learning Assessment Coordinator’s Manual Administration Schedules.

Table 4. 2009 Grade 4 State Standardized Testing Schedule

<table>
<thead>
<tr>
<th>Subject</th>
<th>Testing Window</th>
<th>Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>April 13 – May 1</td>
<td>Approved locally</td>
</tr>
<tr>
<td>Writing</td>
<td>April 13 – May 1</td>
<td>Approved locally</td>
</tr>
<tr>
<td>Mathematics</td>
<td>April 13 – May 1</td>
<td>Approved locally</td>
</tr>
</tbody>
</table>

SUMMARY

The Office of the Superintendent of Public Instruction is committed to developing an instructionally relevant, performance-based assessment system that enhances instruction and student learning. The assessments are based on the EALRs. Teachers and other professionals who provide pre-service and in-service training to teachers should be thoroughly familiar with the EALRs and the assessments that measure them. Teachers and administrators, at all grade levels, need to think and talk together about what they must do to prepare students to achieve the EALRs, and to demonstrate their achievement on classroom-based and state-level assessments.
PART 2: TEST DEVELOPMENT

The content of the WASL state assessment is derived from the Washington State EALRs (see www.k12.wa.us/curriculuminstruct for links to the EALRs in all subject areas). These EALRs define what Washington students should know and be able to do by the end of Grades 3-8 and 10 in Reading, Writing, Communications, and Mathematics, and by the end of Grades 5, 8, and 10 in Social Studies, Science, the Arts, Health and Fitness. The 2009 WASL tests measured EALRs for Reading and Mathematics in Grades 3-8 and 10, for Science in Grades 5, 8, and 10, and for Writing in Grades 4, 7, and 10. The following sections describe the test development process used to develop the 2009 test forms, using previously administered operational and pilot items.

ITEM AND TEST SPECIFICATIONS

The first step in the test development process was to select “Content Committees” to work with staff from the Office of the Superintendent of Public Instruction (OSPI), and the previous vendor, to develop the test items which make up the assessments at each grade level. Each Content Committee included 20 to 25 persons from throughout the state, most of whom were classroom teachers and curriculum specialists, with teaching experience at or near the grades and in the content areas that were to be assessed.

The second step in the development process was attaining common agreement about the meaning and interpretation of the EALRs and identifying which EALRs could be assessed on a statewide test. It was important that the contractor, the Content Committees and OSPI staff were in agreement about what students were expected to know and be able to do and how these skills and knowledge would be assessed. Benchmark indicators were combined in various ways to create testing targets for which items would be written.

Next, test specifications were prepared. Test specifications define the kinds and numbers of items on the assessment, the blueprint and physical layout of the assessment, the amount of time to be devoted to each content area, and the scores to be generated once the test is administered. It was important at this stage to define the goals of the assessment, and the ways in which the results will be used to ensure the structure of the test would support the intended uses. The test specifications are the building blocks to develop equivalent test forms in subsequent years, and to create new items to supplement the item pool. The final test specifications document the following topics:

- Purpose of the assessment;
- Strands;
- Item types;
- General considerations of testing time and style;
- Test scoring;
- Distribution of test items by item type.
The WASL uses three types of items on the Reading and Mathematics tests: multiple-choice, short-answer, and extended-response. For each multiple-choice item, students in Grades 3 through 5 select the one best answer from among three choices provided, and students in Grades 6 through 8 and 10 select the one best answer from among four choices provided. Each multiple-choice item is worth one point. These items are machine scanned and scored.

The other two “open-ended” item types – short-answer and extended-response – require students to produce their own response in words, numbers, or pictures (including graphs or charts). Short-answer items are worth a maximum of two points (scored 0, 1, or 2) and extended-response items are worth a maximum of four points (scored 0, 1, 2, 3, or 4). Student responses are assigned partial or full credit based on carefully defined scoring rules. These items cannot be scored by machine and require hand-scoring by well-trained professional scorers. Part 7 provides further detail about the hand-scoring process and results for the different subject area tests.

For Writing, students are asked to complete two writing prompts. For the Grade 4 test, students write one narrative piece and one expository piece. The writing prompts may require students to write a letter requesting information, describe an important event or situation, or explain a procedure for completing a task or project. Each written piece is worth six points and is hand-scored for content, organization, and style (1, 2, 3, or 4 points) and for mechanics and spelling (0, 1, or 2 points).

Tables 5 through 7 represent the test blueprints for item content and item types for the Grade 4 Reading, Writing, and Mathematics tests. Item specifications were developed from clarification of the EALRs and the test specifications. Item specifications provide sufficient detail including sample items to help item writers develop appropriate test items for each assessment strand. Separate specifications were produced for different item formats and different testing targets. The test and item specifications documents are not only essential for WASL test construction, but both are tools that teachers can use to develop their own assessments and administrators can use to evaluate instructional programs. Test and item specifications are updated annually as needed. The most recent versions of these specifications are available through the website for the Washington State Office of the Superintendent of Public Instruction. (See http://www.k12.wa.us/assessment/WASL/testspec.aspx for test and item specifications in all subjects).

---

1 Extended-response items are not used in any of the Grades 3-5 tests in 2009.
2 It is important to note that, as more is understood about how to develop high quality items that assess the Washington State EALRs, item and test specifications must continually be refined. Refinements have been made annually since 2000. These refinements are an important part of the test development process and reflect what has been learned through ongoing studies of item level data from 1999 to the present and through external reviewers’ item evaluations. (See the Fourth Grade Mathematics Study conducted by the Northwest Regional Education Laboratory in 2000 and the Seventh and Tenth Grade Mathematics Study conducted by Stanford Research Institute in 2005 for examples.)
Table 5. Grade 4 Reading Test Design

<table>
<thead>
<tr>
<th>Text types/Strands</th>
<th>No. of Reading Selections</th>
<th>No. of Words Per Passage</th>
<th>No. of Multiple-Choice Items</th>
<th>No. of Short-answer Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary ‡</td>
<td></td>
<td></td>
<td>14</td>
<td>2-4</td>
</tr>
<tr>
<td>Comprehension †</td>
<td>3</td>
<td>50-800</td>
<td>7</td>
<td>1-2</td>
</tr>
<tr>
<td>Analysis †</td>
<td></td>
<td></td>
<td>7</td>
<td>1-2</td>
</tr>
<tr>
<td>Informational</td>
<td></td>
<td></td>
<td>10</td>
<td>2-4</td>
</tr>
<tr>
<td>Comprehension †</td>
<td>3</td>
<td>150-700</td>
<td>5</td>
<td>1-2</td>
</tr>
<tr>
<td>Analysis †</td>
<td></td>
<td></td>
<td>5</td>
<td>1-2</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>1800-2200</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>Total Points</td>
<td></td>
<td></td>
<td>24</td>
<td>12</td>
</tr>
</tbody>
</table>

* Reading EALR 1: The student understands and uses different skills and strategies to read.
† Reading EALR 2: The student understands the meaning of what is read.
‡ Reading EALR 3: The student reads different materials for a variety of purposes.

Table 6. Grade 4 Writing Test Design

<table>
<thead>
<tr>
<th>Strands</th>
<th>Scored 0-2 Points</th>
<th>Scored 0-4 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COnent &amp; Style</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CONVentions &amp; Mechanics</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Expository</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COnent &amp; Style</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CONVentions &amp; Mechanics</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total Points</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

1 Writing EALR 1: The student writes clearly and effectively (concept & design, style [word choice, sentence fluency, voice], and conventions).
2 Writing EALR 2: The student writes in a variety of forms for different audiences and purposes.
3 Writing EALR 3: The student understands and uses the steps of a writing process.
<table>
<thead>
<tr>
<th>Strands</th>
<th>Multiple-choice</th>
<th>Short-answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Sense 1</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td>Measurement Concepts 1</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td>Geometric Sense 1</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td>Probability and Statistics Procedures 1</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td>Algebraic Sense 1</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td>Solves Problems 2 &amp; Reasons Logically 3</td>
<td>0-2</td>
<td>1-2</td>
</tr>
<tr>
<td>Communicates Understanding 4</td>
<td>0</td>
<td>1-2</td>
</tr>
<tr>
<td>Making Connections 5</td>
<td>2-4</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Number of Items</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>Maximum Number of Points</td>
<td>22</td>
<td>16</td>
</tr>
</tbody>
</table>

1 Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.
2 Mathematics EALR 2: The student uses mathematics to define and solve problems.
3 Mathematics EALR 3: The student uses mathematical reasoning.
4 Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.
5 Mathematics EALR 5: The student makes mathematical connections.
CONTENT REVIEWS & BIAS AND FAIRNESS REVIEWS

Using test and item specifications, item writers prepare new items and scoring rubrics. Item writers include committees of Washington teachers who participate in item writer workshops for professional development opportunities, and Content Specialists. Washington teacher item-writers include novice and experienced item writers, who all receive focused training during Washington item writer workshops. Raw items are initially produced during these workshops, and later refined by full-time staff of Content Specialist professionals who have, on average, 14 years of classroom and pedagogical experience. All item writers receive in-depth training before actively working on a contract as Content Specialists. Half of the Content Specialists assigned to the Washington contract have advanced degrees in curriculum, instruction, assessment, or their subject area specialty.

Item writers develop items, passages, and scenarios that:

- Match the passage, scenario, and item specifications;
- Fulfill the test map specifications;
- Display content accurately and clearly;
- Are within the grade level reading range;
- Are free of bias;
- Are sensitive to students with special needs.

Before an item may be piloted, it must be reviewed and approved by the Content Committee and the Bias and Fairness Committee. A Content Committee’s task is to review the item content and scoring rubric to assure that each item:

- Is an appropriate measure of the intended content (EALR);
- Is appropriate in difficulty for the grade level of the examinees;
- Has only one correct or best answer for each multiple-choice item;
- Has an appropriate and complete scoring guideline for open response items.

The Content Committees can make one of three decisions about each item: approve the item and scoring rubric as presented, conditionally approve the item and scoring rubric with recommended changes or item edits to improve the fit to the EALRs and the specifications, or eliminate the item from further consideration.

Based on content reviews, items may be revised. Each test item is coded by content area (EALR) and by item type (multiple-choice, short-answer, extended-response), and presented to the OSPI Assessment Specialist for final review and approval before pilot testing. The final review includes a review of graphics, art work, and page layout.

The Bias and Fairness Committee reviews each item to identify language or content that might be inappropriate or offensive to students, parents, or community members; or items which might contain “stereotypic” or biased references to gender, ethnicity, or culture. The Bias and Fairness Committee reviews each item and accepts, edits, or rejects it for use in item pilots.
ITEM PILOTS

Once an item has been approved for placement on a pilot test, pilot test forms are constructed by the contractor and must be approved by OSPI. Items are pilot tested with a sample of students from across the state. Pilot Reading and Mathematics items are included in operational testing sessions, but do not contribute to reported scores. Pilot Science items were previously administered in a stand-alone pilot testing program, but beginning in Spring 2006, they were also embedded in the operational test. Pilot items are presented in similar locations across operational forms. No more than 7 items are piloted in any single test form, so no student is administered more than 7 pilot items. Since pilot items are administered together with operational test items, students tend to complete pilot items with the same level of motivation and attention they give to the operational test items. The data for these pilot items are considered to provide reasonable estimates to the item difficulty when the items become operational. A test form is defined by different sets of pilot items and a common set of operational items. Placing pilot items on the operational form and systematically distributing the pilot forms, yields a statewide representative, randomly equivalent sample of students that respond to each pilot item. For the Grade 4 Writing program, new pilot prompts were last administered to a volunteer sample in a stand-alone pilot program in Fall 2006.

For each pilot form, at least 1,200 student responses are scored. Of the 1,200 scored student responses, and as a function of the number of total pilot forms administered at a grade level, approximately 100 responses per pilot item come from each of the OSPI-designated ethnic groups (African American, Asian/Pacific Islander, Native American, and Hispanic). A statewide representative sampling framework – specified by geographic region, district density, building enrollment type, grade level enrollment, proportion of ethnic groups within grade level, and percent of students receiving AFDC – is used to develop an intended sampling plan to distribute the pilot forms. Further details about the sampling framework and annual pilot form distribution plans are described in Blue Dot Rotation Documentation.

The following section of this report describes the statistical analyses (e.g., traditional item analyses, differential item functioning, and item response theory (IRT) scaling) conducted for the embedded pilot items. Tables providing statistical summary information for these embedded pilot items are provided in Appendix B (see Tables 59 – 64).
CALIBRATION, SCALING, AND ITEM ANALYSIS

After each administration, student responses are scored using the scoring rubrics approved by the Content Committees. Statistical analyses are completed using procedures based on classical test theory and modern item response theory to evaluate the effectiveness of the items and to empirically examine the presence of differential item functioning or “item bias.”

Two types of item analyses are completed for all items: Traditional item analysis and item response theory analysis using the Rasch Partial Credit Model. Traditional item analysis statistics, based on classical test theory, include item means, item-test correlations, percent of students at each response option or score level, and percent of students omitting the item. The Rasch Partial Credit Model is one class of mathematical models, based on modern item response theory, that is used to estimate item location and item fit statistics.

In addition, the pilot items Differential Item Functioning (DIF) analyses were conducted to flag items that might contain item bias, so that they could be further reviewed. DIF is observed when examinees from different demographic groups, with the same ability (students matched on operational total test score), perform differently on the same item.

Traditional Item Analysis

Traditional (or classical) item analyses involve computing, for every item in each form, a set of statistics based on classical test theory. Each statistic is designed to provide some key information about the quality of each item from an empirical perspective and includes item means, item-test correlations percent of students at each response option or score level, and percent of students omitting the item.

Item means or p-values and item-test correlations or point-biserials are computed for the multiple-choice items. These are the classical test theory equivalents of item difficulties and item discriminations. The item p-value is the percentage of examinees that selected the correct answer choice, and ranges from 0.00 to 1.00. The point-biserial is an index of the relationship between performance on an item and overall performance on the test. Point-biserials can range from -1.00 to 1.00. Point-biserials are usually greater than 0.20, but these values can be deflated when item content is unfamiliar to all examinees, regardless of performance on the total test or when the item does not distinguish between higher and lower test performance sufficiently well. Option biserials are correlations between incorrect answer choices and the overall test, and typically exhibit negative values.

Item means for short-answer and extended-response item types reflect the average earned item score for examinees. For two-point items, item means can range from 0 to 2. For four-point items, item means can range from 0 to 4. Item-test correlations for polytomous items indicate the relationship between item performance and overall test performance. As with multiple-choice items, item-test correlations can range from -1.00 to 1.00.

Unlike IRT item statistics, item means and item-test correlations are dependent on the particular group of examinees who took the test. When examinees are exceptionally well
schooled in the concepts and skills tested, item means will be fairly high and the items will appear to be easy. When examinees are not well schooled in the concepts and skills tested, item means will be fairly low and items will appear to be difficult. When one group’s performance on an item does not relate well to performance on the test as a whole, the item-test correlation will be artificially low. Since scaled IRT item parameters can provide information about a pilot item relative to a larger item pool, both Rasch and classical item statistics are computed to evaluate the quality of items and their inclusion in the larger item pool.

Summary statistics of the item means and item-test correlations are provided in Table 50. The item means for the short-answer and extended-response items were divided by the maximum possible score for that item, expressing all average item scores as a proportion of the maximum item score, like the multiple-choice items.

Additionally, for multiple-choice, short-answer, and extended-response items, the percent of students omitting the item are computed. The Grade 4 WASL Reading and Mathematics tests are intended to provide sufficient time for all students to respond to all of the questions. Table 51 lists the summary of the omit rate analyses.

**IRT Analysis**

The Rasch Partial Credit Model is a class of IRT models used to place all items with a common construct on the same scale. Differences between grade level development and subject area constructs frequently necessitate the development of separate grade level/subject area scales. Elementary grade level mathematics items, for example, are typically on a separate scale from elementary grade level reading items. Examinee abilities and item difficulty parameters share the same scale, and unlike traditional item means, IRT item difficulty parameters are essentially sample-independent. Stated another way, an item difficulty parameter is the same for different groups of examinees. Equations 1 and 2 specify the Rasch Partial Credit Model, defined by the probability of person $n$ scoring $x$ on item $i$ as:

$$P_{nx} = \frac{\exp \sum_{j=0}^{m-1} (B_n - D_j)}{\sum_{k=0}^{m-1} \exp \sum_{j=0}^{m-1} (B_n - D_j)}$$

(Equation 1)

where $x = 0, 1, 2, \ldots, m - 1$;

$B_n$ = person parameter;

$D_j$ = item-category parameter; and

$$\sum_{j=0}^{m-1} (B_n - D_j) = 0$$

(Equation 2)

Item difficulties and examinee abilities can be estimated for a test using this mathematical model. The item difficulty is the location on the ability scale where examinees have a 50/50 chance of
answering an item correctly. Figure 1 illustrates the relationship between examinee ability and item difficulty from two different tests.

**Figure 1. Location of examinee $\beta_1$ on two tests with different items**

Mathematics Test 1

![Diagram of Mathematics Test 1 with items $\delta_1$ to $\delta_{10}$]

Mathematics Test 2

![Diagram of Mathematics Test 2 with items $\delta_1$ to $\delta_{10}$]

Test scores can be conveyed in scaled scores or number correct scores. In Figure 1, above, an examinee correctly answered the first eight items on Mathematics Test 1 and the first six items on Mathematics Test 2. This example illustrates how number correct scores for the same examinee is a function of the particular set of items on a test. When all Mathematics items ($\beta_1, \beta_2, \beta_3, \ldots, \beta_{10}$) are placed on the same scale, the examinee’s ability can be reported relative to an underlying, common scale – a value between $\delta_8$ (from Test 1) and $\delta_7$ (from Test 2).

**Figure 2. Location of examinee $\beta_1$ on the same “Mathematics Test” scale**

![Diagram of “Mathematics Test” scale with items $\delta_1$ to $\delta_{10}$]

When a collection of items shares a construct, calibrating and scaling items with the Rasch model places the items on the same scale, so that examinee test scores reflect their location on the underlying scale, rather than the number of items answered correctly on a particular test.

For polytomously scored items, the Rasch Partial Credit Model estimates the step difficulties for each item-category. For example, items with 3 possible score points (0, 1, 2) can have two step categories. The first step is the location on the scale where examinees with abilities equal to that location have an equal chance of getting a score of 0 or 1. The second step is the point on the scale where examinees with abilities equal to that location have equal probability of earning a score of 1 or 2.
For dichotomously scored, multiple-choice items, the Rasch Partial Credit Model becomes a special case of the Rasch one-parameter model:

\[
P_{ai} = \frac{\exp(B_i - D_j)}{1 + \exp(B_i - D_j)}
\]  

(Equation 3)

where \( B_i \) = person parameter;
\( D_j \) = item parameter.

When item scores are placed on a scale, items are assessed for statistical fit to the Rasch model. In order for items to be included in the operational item pool, they must measure relevant knowledge and skill, represent desired locations on the ability scale, and fit the Rasch model.

IRT analyses are completed separately by grade level for each WASL content area. The adequacy of item fit depends on whether the items in a scale all measure a similar construct or whether the scale is essentially unidimensional. Just as height, weight, and body temperature are different dimensions of the human body, so are Reading, Writing, Mathematics, and Science different dimensions of achievement.

In order to place all grade level/content area pilot items from different test forms on the same Rasch scale, all test forms shared a common set of anchor (or linking) items. For the Reading and Mathematics tests, the same set of operational items appeared in all test forms, but different sets of pilot items were embedded in or appended to the operational sections. Pilot items were then calibrated and scaled to the grade level/content area scale through the operational items. Summary statistics for operational items are provided in Table 52 and for the embedded pilot items in Table 61.

Bias Analysis

In addition to traditional item analyses and IRT analyses, Bias analyses, or Differential Item Functioning statistics, are also conducted on the embedded pilot items. Bias analyses are used to identify those items that identifiable groups of students (e.g. males, females) with the same underlying level of ability have different probabilities of answering correctly. Examinees are separated into relevant subgroups based on ethnicity or gender. Then examinees in each subgroup are ranked relative to their total test score (conditioning on ability). Examinees in the focal group (e.g., females) are compared to examinees in the reference group (e.g., males) relative to their performance on individual items.

If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, the item may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify items that are potentially biased. Subsequent review by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences. For the Spring 2009 Grade 4 WASL Reading and
Mathematics tests, DIF analyses were conducted for gender (male/female) and ethnicity (White/Asian, White/African American, White/Hispanic, and White/Native American.

Statistics from two DIF detection methods were computed: the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) and the standardization procedure (Dorans & Kulick, 1983, 1986). As part of the Mantel-Haenszel procedure, the statistic described by Holland Thayer (1988), known as MH D-DIF, was used. This statistic is expressed as the differences between members of the “focal group” (female, Asian, African American, Hispanic, and Native American) and members of the “reference group” (males and White) after conditioning on total operational test score. This statistic is reported on the ETS delta scale, which is a normalized transformation of item difficulty (p-value) with a mean of 13 and a standard deviation of 4. Negative MH D-DIF statistics favor the reference group and positive values favor the focal group. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically significantly different based on the MH D-DIF (p > 0.05) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences (p < 0.05), the effect size is used to determine the direction and severity of the DIF. For CR items, the Mantel-Haenszel procedure was executed where item categories are treated as integer scores and a chi-square test was carried out with one degree of freedom. DIF analyses were not conducted if the sample size for either the reference group or focal group was less than 100 and the sample size for the two groups combined was less than 400.

Based on these DIF statistics, items are classified into one of three categories and assigned values of A, B, or C. Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of DIF. Negative values imply that, conditional on the matching variable, the focal group has a lower mean item score than the reference group. In contrast, a positive value implies that, conditional on total test score; the reference group has lower mean item score than the focal group. The flagging criteria for multiple-choice items are provided in Table 8.

---

3The formula for the estimate of constant odds ratio is:

\[ \alpha_{MH} = \left( \frac{\sum_m R_{rm} W_{fm}}{N_m} \right) \left( \frac{\sum_m R_{fm} W_{rm}}{N_m} \right) \]

where

- \( R_{rm} \) = number in reference group at ability level \( m \) answering the item right,
- \( W_{fm} \) = number in focal group at ability level \( m \) answering the item wrong,
- \( R_{fm} \) = number in focal group at ability level \( m \) answering the item right,
- \( W_{rm} \) = number in reference group at ability level \( m \) answering the item wrong,
- \( N_m \) = total group at ability level \( m \).

This can then be used in the following formula (Holland & Thayer, 1985):

\[ MH\ D\ -\ DIF = -2.35 \ln(\alpha_{MH}). \]
Table 8. DIF Categories for Multiple-Choice Items

<table>
<thead>
<tr>
<th>DIF CATEGORY</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (negligible)</td>
<td>Absolute value of the MH D-DIF is not significantly different from zero, or is less than one. Positive values are classified as “A+” and negative values as “A-.”</td>
</tr>
<tr>
<td>B (slight to moderate)</td>
<td>1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; or 2. Absolute value of the MH D-DIF is significantly different from one, but is less than 1.5. Positive values are classified as “B+” and negative values as “B-.”</td>
</tr>
<tr>
<td>C (moderate to large)</td>
<td>Absolute value of the MH D-DIF is significantly different from one, and is at least 1.5. Positive values are classified as “C+” and negative values as “C-.”</td>
</tr>
</tbody>
</table>

For constructed response (CR) items, the MH D-DIF statistic is not calculated; instead the standardized mean difference (SMD) (Zwick, Donoghue, & Grima, 1993), in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959), was used in classifying items into A, B or C DIF categories. The flagging criteria for constructed response items are provided in Table 9.

Table 9. DIF Categories for Constructed-Response Items

<table>
<thead>
<tr>
<th>DIF CATEGORY</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (negligible)</td>
<td>Mantel Chi-square p-value &gt;0.05 and</td>
</tr>
<tr>
<td>B (slight to moderate)</td>
<td>Mantel Chi-square p-value &lt;0.05 and</td>
</tr>
<tr>
<td>C (moderate to large)</td>
<td>Mantel Chi-square p-value &lt;0.05 and</td>
</tr>
</tbody>
</table>

SMD is the Standardized Mean Difference index, and SD is the total group standard deviation of the item score (in its original metric). A negative SMD value shows that the question is more difficult for the focal group whereas a positive value indicates that it is more difficult for the reference group.

DIF statistics are computed for all pilot items and reviewed at Data Review as part of the evaluation process for inclusion into the active item pool. DIF statistics are not computed on operational items. Tables 62 and 63 (in Appendix B) summarizes the number and percentage of items by DIF category from the 2009 pilot. The 2009 operational tests are comprised of items that were piloted in years prior to 2008, which were reviewed and approved by Content Review, Bias and Fairness Review, and Data Review Committees.
Data Reviews

After statistical analyses for pilot items have been completed, Data Review Committees review these results to evaluate item quality and appropriateness for inclusion in the larger item pool and candidacy for future operational use. Examples of Data Review reports used by Data Review Committees are provided in Appendix C. These committees include Washington educators, curriculum specialists, and educational administrators with grade-level and subject matter expertise relevant to the specific data review grade levels. All committee members are selected by OSPI from a recommendation pool of professional Washington education organizations and from a pool of Washington educators who complete an application to participate in OSPI professional development activities. OSPI content specialists and content specialists and psychometricians from the testing vendor facilitate the Data Reviews. Pilot items and scoring rubrics are re-evaluated to confirm fit to the EALRs, pilot item statistics are reviewed to determine whether content or language may have contributed to any significant DIF statistics. During these committee reviews, items are either accepted into or rejected from the active item pool. Items may also be selected to be revised and repiloted.

Item Selection

Statistical review of items involves examining item means, Rasch item difficulties, and item-test correlations to determine whether items are functioning well. Statistical review also requires examining the adequacy of the model fit to the data. Items that exhibit poor fit to the model may need to be revised or removed from the item pool. Items that function poorly (too easy, too difficult, or have low or negative item-test correlations) may also need to be revised or removed from the item pool. Finally, items that are flagged for bias against any group are examined closely to decide whether they will be removed from the pool. Operational test forms are constructed with items from the active item pool.

TEST CONSTRUCTION

As described above, traditional item statistics and IRT-based item statistics were computed to evaluate the quality of pilot items and their eligibility for future operational use. Pilot items that met quality standards, statistical requirements, and content criteria were retained in the item pool for future operational use. Approved items from the pool were selected to construct the 2009 tests.

New operational forms are created for each test administration, usually sometime in the Spring after Data Review. OSPI content specialists, testing vendor content specialists, and psychometricians jointly select items according to test build specifications and test blueprints. There are a number of factors that must be considered during the test construction process. Operational test forms are constructed according to the requirements outlined in the WASL test blueprints, test specifications, and test maps. Items are selected to satisfy the test map, meet target test difficulty, and represent an overall test with balanced content. A test development checklist is used to review the initial test pulled during the test build. Test build is an iterative process to balance test content and its statistical properties.
Test specifications guide the item selection process to ensure that all relevant strands are represented in each operational form. Representation of all gender and ethnic groups – in character names, topics of reading passages, and item contexts – is reviewed to ensure that Reading passages, scenarios in Science, and stimulus materials used in the Mathematics and Writing tests include balanced representations of groups. The WASL is a criterion-referenced assessment with defined performance level standards on each operational test. Items are selected to cover a range of difficulty levels on each of the Reading, Mathematics, and Science scales.

When a new operational form is created for each test administration, test scores must be equated to the baseline scale to maintain interpretability over time. The baseline scale is determined when performance level standards are defined, typically following the first operational test administration until performance level standards are revisited or redefined. The test developer’s objective is to construct a new, parallel operational test form for each administration with target statistical characteristics and criteria to allow for comparability across test administrations. The better the match to these criteria, the better the equating accuracy of test scores between different test administrations.

Operational test forms are constructed such that the operational test forms across administrations have difficulties that are as similar as possible. The weighted mean Rasch difficulty is used as a statistical target for evaluating the test form’s difficulty. The weighted Rasch item difficulty of each operational item is multiplied by the maximum raw item score to obtain its weighted Rasch difficulty. The sum of weighted item Rasch difficulties is divided by the maximum total raw test score to compute the weighted mean Rasch difficulty for the test. The weighted mean Rasch difficulty for an operational form should approximate historical weighted mean Rasch difficulties unless there is a purposeful effort to shift the targeted difficulty level of a test. During the early years of a new assessment program, the target weighted mean Rasch frequently is near zero (0). Over time, however, item and test difficulties may tend to shift. The historical and 2009 weighted mean Rasch difficulty values can be found in Table 17, Part 5.

REFERENCES


PART 3: VALIDITY

Validity is one of the most important attributes of assessment quality. Validity refers to the degree to which each interpretation or use of a test score is supported by evidence that is gathered (APA, AERA, NCME, 1999; ETS, 2002). It is a central concern underlying the development, administration, scoring of a test and the uses and interpretations of test scores.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. It does not involve a single study or gathering one particular kind of evidence. Validation involves multiple investigations and various kinds of evidence (APA, AERA, & NCME, 1999; Cronbach 1971; ETS, 2002; Kane, 2006). The process begins with test design and continues through the entire assessment process, including item development and pilot testing, analyses of item and test data, test scaling, scoring and score reporting.

This section presents the evidence gathered to support the intended uses and interpretations of scores for the WASL testing program. This description is organized in the manner prescribed by APA, AERA, and NCME’s Test Standards for Educational and Psychological Testing (1999). These standards require a clear definition of the purpose of the test, which includes a description of the qualities, called constructs, that are to be assessed by a test, the population(s) to be assessed, as well as how the scores are to be interpreted and used.

In the text below, the purpose of the test is defined, and then the kinds of validity evidence that has been gathered are described and discussed.

PURPOSE OF THE TEST

The Constructs To Be Measured

As described in Part 1 of this technical report the WASL tests are criterion-referenced assessments designed to determine how well a student has learned important concepts and skills. Specifically the WASL are intended to measure the Essential Academic Learning Requirements (EALRs) that describe what all students should know and be able to do. The grade 4 WASL tests include Reading, Writing, and Mathematics.

Test blueprints and specifications provide an operational definition of each construct (Cronbach, 1971, p.449). That is, they define for each subject area to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score examinee responses. They control as many aspects of the measurement procedure as possible, so that the testing conditions will remain the same over test administrations (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to minimize construct irrelevant score variance (Messick, 1989).

More information about the State’s EALRs can be found at http://www.k12.wa.us/curriculuminstruct. The test blueprints can be found in Part 2 of this technical report.
The Scores Generated and the Interpretations and Uses of These Scores

As described in Part 1, total test scaled scores, performance levels, and strand score indicators are generated and provided to students, schools, and other test users (see Appropriate Use of Test Scores in Part 1).

EVIDENCE BASED ON TEST CONTENT

According to the Standards analyses of the relationship between a test’s content and the construct that the test was designed to measure can provide important evidence of validity. In current K12 testing the construct of interest is operationally defined by state content standards and the test blueprints that specify the content, format, and scoring of items that are admissible measures of the knowledge and skills described in the content standards. Evidence that the items meet these specifications and represent the domain of knowledge and skills, referenced by the standards, provides evidence to support the inference that students’ scores on these items can appropriately be regarded as measures of the intended construct.

As noted in the Standards, evidence based on test content may involve logical analyses of test content in which experts judge the adequacy with which the test content conforms to the test specifications and represents the intended domain of content. Such reviews can also be used to determine whether the test content contains material that is not relevant to the construct of interest. Analyses of test content may also involve the use of empirical evidence of item quality.

Also to be considered in evaluating test content are the procedures used for test administration and test scoring. As Kane (2006, p.29) has noted, although evidence that appropriate administration and scoring procedures have been used does not provide compelling evidence to support a particular score interpretation or use, such evidence may prove useful in refuting rival explanations of test results.

Part 2 of this technical report, “Test Development,” describes the processes used to ensure valid content representation, alignment, and conformity to the defined content area domains. Test blueprints, test specifications, and test maps define the framework of all WASL test development and test construction. Throughout the test development process, committees of professional educators, content area experts, and professionally trained test developers all provide on-going review, verification, and confirmation to ensure content validity of test content is aligned with the EALRs.

EVIDENCE BASED ON INTERNAL STRUCTURE

As suggested by the Standards, evidence of validity can also be obtained from studies of the properties of the item scores and the relationship between these scores and scores on components of the test. To the extent that the score properties and relationships found are consistent with the definition of the construct measured by test, support is gained for interpreting these score as measures of the construct.
In K12 testing, it is usually assumed that there is a dominant construct or dimension that underlies the total scores obtained on each test. Evidence to support this assumption can be gathered from item analyses, evaluation of internal consistency, and studies of model-data fit, dimensionality and reliability.

With respect to the strand scores that are reported, these scores are intended to reflect examinees' knowledge and/or skill in an area that is part of the construct underlying the total test. Analyses of the intercorrelations among the strand scores themselves and between the strand scores and total test score can be used for this purpose. It is also useful to provide information about the internal consistency of the items on which each strand score is based.

Studies were previously conducted to gather construct validity evidence for the Grade 4 WASL Reading, Writing, and Mathematics. The *WASL Technical Reports for Grade 4* from 1999 to 2002 provide construct validity information for the 1999 through 2002 Grade 4 data. The internal structure of tests was evaluated by examining the correlations among strand scores for the WASL content area strands and by factor analyses of the strand scores. The relationship of the WASL to external measures has been studied through correlational analysis of WASL scores and, in 2001 and 2005, with scores on the *Iowa Test of Educational Development*. In this technical report, there are three types of evidence of Internal Structure:

1. Correlations Among WASL Strand Scores;
2. Factor Analysis of Strand Scores;
3. Confirmatory Factor Analysis;

**Correlations Among WASL Strand Scores**

Table 10 lists the intercorrelations of strand scores between the different 2009 WASL content area tests. These correlations are derived from the statewide student data file received from Data Recognition Corporation on July 20, 2009. Students were included in these tables based on the follow criteria:

a) Students whose reporting grade is 4;
b) Student’s test type is either WABA or WASL;
c) Student’s attempt value is TS (Tested);
d) Student not missing test pages4; and
e) For Mathematics only, student did not take the Braille or the Large Print version5.

---

4 During the scanning of the test booklets, it was noted that twelve students were missing one or more pages from their test booklets. These students are excluded from this report.
5 Some operational items were excluded from the Mathematics Braille and Large Print forms, thus special raw score to scaled score relationship tables were created for these forms. All items were included in the Reading and Science Braille and Large Print forms, therefore no special scoring tables were produced for Reading or Science.
Table 10 lists the intercorrelations of strand scores between different 2009 WASL content area tests. These intercorrelations were completed only using the 74,068 cases for which all three Grade 4 WASL content area scores were available for analysis. The results of the intercorrelations analyses (see Table 10) reveal the following relationships:

a) The correlation between Reading Analysis and Reading Comprehension is 0.75;
b) Writing COS strand score correlated 0.56 with the CONV strand score;
c) The correlation between Mathematics Content and Process is 0.71;
d) Intercorrelations between Reading strand scores and Mathematics strand scores range from 0.60 to 0.67;
e) Intercorrelations between Writing strand scores and Mathematics strand scores range from 0.47 to 0.52; and
f) Intercorrelations between Writing strand scores and Reading strand scores are between 0.54 and 0.55.

These intercorrelations suggest that, for Reading, Writing, and Mathematics tests are moderately related. To further investigate the relationships between Reading, Writing, and Mathematics, an exploratory factor analysis was completed on the content area strand scores.
Table 10. 2009 Grade 4 Strand Score Intercorrelations

<table>
<thead>
<tr>
<th>Test</th>
<th>Reading</th>
<th>Writing</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strand</td>
<td>ALY (RD)</td>
<td>CMP (RD)</td>
<td>COS (WR)</td>
</tr>
<tr>
<td>ALY (RD)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMP (RD)</td>
<td>0.75</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>COS (WR)</td>
<td>0.54</td>
<td>0.54</td>
<td>1</td>
</tr>
<tr>
<td>CONV (WR)</td>
<td>0.55</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>CT (MA)</td>
<td>0.66</td>
<td>0.67</td>
<td>0.51</td>
</tr>
<tr>
<td>PC (MA)</td>
<td>0.60</td>
<td>0.61</td>
<td>0.47</td>
</tr>
</tbody>
</table>

ALY (RD) - Analysis  
CMP (RD) - Comprehension  
COS (WR) - Content, Organization, & Style  
CONV (WR) - Writing Conventions  
CT (MA) - Content  
PC (MA) - Process

Factor Analysis of Strand Scores

The relationships between the WASL strand scores were investigated with a principal components analysis, followed by a common factor model analysis using PROC FACTOR in SAS v 9.1. The number of factors was defined using two criteria – a scree plot, and a solution in which at least 60% of the variance is explained. The eigenvalues suggested a three-factor solution that explained 84% of the total variance. Rotation is a step in factor analysis that facilitates the identification of meaningful factor descriptions, and for ease of interpretation, an orthogonal varimax rotation was used. Table 11 lists the rotated factor pattern for the three-factor solution. These patterns indicate distinct constructs between the Mathematics, Reading, and Writing strand scores. For these analyses, a scree plot exhibited two prominent factors, and the presence of a third, less prominent factor. The first two factors alone accounted for 76% of the total variance.
Figure 3. Grade 4 Scree Plot of Eigenvalues
Table 11. 2009 Grade 4 Rotated Factor Pattern on Tests for Three-Factor Solution

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALY (RD)</td>
<td>0.82*</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>COMP (RD)</td>
<td>0.81</td>
<td>0.35</td>
<td>0.31</td>
</tr>
<tr>
<td>CT (MA)</td>
<td>0.42</td>
<td>0.75</td>
<td>0.29</td>
</tr>
<tr>
<td>PC (MA)</td>
<td>0.28</td>
<td>0.88*</td>
<td>0.24^</td>
</tr>
<tr>
<td>Content, Organization, &amp; Style (WR)</td>
<td>0.31</td>
<td>0.18^</td>
<td>0.81*</td>
</tr>
<tr>
<td>Conventions (WR)</td>
<td>0.22^</td>
<td>0.29</td>
<td>0.81*</td>
</tr>
</tbody>
</table>

* Largest loading within a common factor
^ Smallest loading within a common factor

Examining Construct Validity Through Confirmatory Factor Analysis

It is assumed that each WASL test is unidimensional to measure a specific content domain (e.g., mathematics or reading). Each WASL test is also designed to measure different sub-areas or strands within a specific content domain. For example, the WASL mathematics test includes items designed to assess students’ knowledge about mathematical content strands (number sense, measurement, geometric sense, probability/statistics, and algebraic sense) and mathematical process strands (solve problems, communicate understanding, and make connections). These content and process strands represent different mathematical knowledge and skills but are correlated to some degree. Strand score indicators (+ or -) are reported to provide teachers, parents, and students more detailed information about students’ learning and performance on the test.

Traditional approaches to evaluate construct validity include examining inter-item correlations and conducting exploratory factor analysis. These methods, however, offer limited information to compare and test various structural models about a test’s underlying construct. Confirmatory factor analysis offers a method to compare and test models of constructs.

Two hypothetical constructs are statistically tested and compared to examine the structure of the WASL tests.

1. The WASL is strictly unidimensional where all items in a test measure a single knowledge and skill. As illustrated in Figure 4, this is a single-factor structural model in which all items load on a general factor. This model presumes all modeled items contribute to the estimation of a general ability factor.
2. The second structural model supports strand score reporting and hypothesizes that each WASL test measures several distinct but correlated knowledge and skills. This is a multi-factor model where an item loads on the strand to which it corresponds. The strands are correlated with each other as illustrated in Figure 5.

Note: *Subject area may be Reading, Mathematics, or Science depending upon the grade.
LISREL 8.52 (Joreskog & Sorbom, 1993) was used to fit specified factor models to the data. For each subject and form, two models were fit to the data: a one-factor model; and a multi-factor model corresponding to the items in each strand. Parameter estimation was performed using the maximum likelihood (ML) method.

To compare model fit to the data for each hypothesized model, several goodness-of-fit indices were examined including: Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), Chi-Square ($\chi^2$), and the chi-square statistic divided by its associated degrees of freedom (df). To evaluate model fit, the one-factor and multi-factor fit statistics may be compared. In general, if fit statistics are adequate for the one factor model and improvement in fit statistics are small for the multi-factor model, then the results suggest that the data are essentially unidimensional.

Following are the general criteria to assess the indices of fit in this analysis:
1. The CFI is derived from the comparison of a hypothesized model with the independence model and provides a measure of complete covariation in the data. CFI values range from 0 to 1.0. A CFI value > .90 indicates acceptable fit to the data. Bentler (1990) has suggested that the CFI should be the index of choice (cited in Byrne, 1998);
2. Chi-square divided by associated degrees of freedom less than 2.0 indicates acceptable model fit (Arbuckle, 1997);
3. For RMSEA, a value less than 0.05 indicates a good fit, a value between 0.05 and 0.10 indicates a reasonable fit, and a value above 0.10 indicates poor fit;
4. If the $\chi^2$ value of the more complex model (more parameters to be estimated) is significantly smaller than the $\chi^2$ value of the more parsimonious model, the more complex model will be considered a better fitting model and thus better represents the data.
Due to the large number of cases analyzed, the chi-square statistics, which are sensitive to sample size, were all high (see Table 12). However, the other fit statistics (CFI and RMSEA) are within acceptable ranges for good model fit. Since the single-factor model fits reasonably well to the data for all of the grade level-subject tests, the unidimensionality assumption and the IRT-based ability estimation are both supported.

When comparing the correlated-multifactor to the single-factor model, the WASL tests do not show significantly better fit for the multifactor, strand-based structural model. While good model fit supports the current practice of reporting strand score indicators, the modest inter-strand correlations suggest caution in the separate interpretation of these strand scores when they are interdependent.
Table 12. 2009 Grade 4 Model Goodness-of-Fit Statistics

<table>
<thead>
<tr>
<th>Subject</th>
<th>Model</th>
<th>N</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2$/df</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Single-factor</td>
<td>75,823</td>
<td>64,384.57</td>
<td>405</td>
<td>158.97</td>
<td>.98</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>Multi-factor (Strand-based)</td>
<td>75,823</td>
<td>64,361.11</td>
<td>404</td>
<td>159.31</td>
<td>.98</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td></td>
<td></td>
<td></td>
<td>23.46</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>Single-factor</td>
<td>76,083</td>
<td>34,461.93</td>
<td>405</td>
<td>85.09</td>
<td>.99</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>Multi-factor (Strand-based)</td>
<td>76,083</td>
<td>34,120.35</td>
<td>404</td>
<td>84.46</td>
<td>.99</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td></td>
<td></td>
<td></td>
<td>341.58</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Note: Chi-square is inflated due to large sample size. All available cases were used in the analyses.
PERFORMANCE IN DIFFERENT POPULATIONS

The validity of the WASL assessments lies primarily in the content tested, which is based on a statewide curriculum intended to be taught to all students. The WASL tests, therefore, are neither more nor less valid for any specific population.

Part 8 of this technical report includes summaries of examinee performance on the WASL according to particular categorical programs – Title I Reading, Title I Mathematics, LAP Reading, LAP Mathematics, Special Education, Highly Capable Students, ELL/Bilingual, and Title I Migrant. These data can be examined to determine whether patterns of performance are consistent with expectation based on examinees’ special needs. Students identified as “highly capable,” for example, are likely to outperform all other groups on all tests. Students who are in Title I Migrant and ELL/Bilingual programs frequently have difficulty with reading and writing performance. Females outperform males in Reading, Writing, and Mathematics. While females and males perform similarly in Mathematics, a higher proportion of females than males meet standard in Reading and Writing.

ADDITIONAL VALIDITY EVIDENCE

In addition to the evidence presented here and the validation documentation gathered and maintained by OSPI, other information in support of the WASL appears in the following sections.

- Part 4 provides detailed information regarding the internal consistency and the decision consistency and accuracy, providing evidence supporting the validity of inferences based on the total test scores and the proficiency levels.
- Part 5 provides information about the calibration, scaling, and equating procedures used to place scores on the base scales, thereby supporting the validity of inferences based on comparability of test scores.
- Part 7 provides information about the scoring of open-ended items, thereby supporting the validity of inferences based on the total test scores.

SUMMARY

The results of these analyses provide evidence of validity based on test content and content area constructs of the 2009 Grade 4 WASL. Although achievement in one subject area is generally related to achievement in other subject areas, an examination of WASL strand scores suggests that Reading, Writing, and Mathematics comprise different underlying dimensions of academic achievement and performance on the WASL tests.
REFERENCES


PART 4: RELIABILITY

The reliability of test scores is a measure of the degree to which the scores on the test are a “true” measure of the examinees’ knowledge and skill relevant to the tested knowledge and skills. In Classical Test Theory, reliability is the proportion of observed score variance that is true score variance.

There are several methods to estimate score reliability: test-retest, alternate forms, internal consistency, and generalizability analysis are among the most common. Test-retest estimates require administration of the same test at two different times. Alternate forms reliability estimates require administration of two parallel tests. These tests must be created in such a way that we have confidence they measure the same domain of knowledge and skills using different items. Both test-retest and alternate forms reliability estimates require significant examinee testing time and are generally avoided when there is potential impact from fatigue or loss of motivation.

The WASL is a system of rigorous measures that requires significant concentration on the part of students for a sustained period of time. For this reason, it was determined that test-retest and alternate forms reliability methods were unlikely to yield accurate estimates of score reliability. Internal consistency measures were used to estimate score reliability for Reading and Mathematics tests.

INTERNAL CONSISTENCY

Internal consistency reliability is an indication of how similarly students perform across items measuring the same knowledge and skills; that is, how consistently does each examinee perform on all of the items within a test. Internal consistency can be estimated by Cronbach’s coefficient alpha. There are two requirements to estimate score reliability:

1. The number of items should be sufficient to obtain stable estimates of students’ achievement; and
2. All test items should be homogeneous (similar in format and measure very similar knowledge and skills).

The WASL tests are complex measures that combine multiple-choice, short-answer, and extended-response items. The Reading and Mathematics tests measure different strands that are components of the Reading and Mathematics content domains. Examinee performance may differ markedly from one item to another due to interactions with prior knowledge, educational experiences, and exposure to similar content or item format. The heterogeneity of items in the Reading and Mathematics tests may tend to under-estimate the reliability of test scores estimated by Cronbach’s coefficient alpha. When items are heterogeneous in content and format as they are in the WASL, it is generally believed that the true score reliability is higher than the estimate computed by Cronbach’s coefficient alpha.

The WASL Writing test consists of two essays. There are four scores for the test (a COS and a CONV score for each essay), the items measure essentially the same ability and share the
same item format. For the Grade 4 Writing test, each essay is scored independently by readers for a maximum total score of 12 points. The number of total score points and test structure may be barely sufficient to justify the use of Cronbach’s alpha to compute an internal consistency estimate of reliability, but a more meaningful estimate of internal consistency may be obtained through applications of generalizability theory.

Cronbach’s coefficient alpha is represented by:

\[ r_{xx} = \left( \frac{N}{N-1} \right) \left( 1 - \frac{\sum s_i^2}{s_x^2} \right) \]  

(Equation 4)

where \( \sum s_i^2 \) = sum of all of the item variances
\( s_x^2 \) = observed score variance, and
\( N \) = the number of items on the test.

Cronbach’s coefficient alphas for each of the 2009 Grade 4 WASL tests, and content strands, are listed in Table 13. Cronbach’s coefficient alpha for the various subgroups are provided in Appendix A (Tables 56-58). The 2009 WASL scores from Reading and Mathematics, as well as the shorter Writing test all exhibit relatively high coefficient alphas to support the expectation items within a content area test work in concert to measure a similar construct.

**STANDARD ERROR OF MEASUREMENT**

One way to interpret the reliability of test scores is with the conditional standard error of measurement (SEM). The SEM is an estimate of the standardized distribution of error around a particular score. An observed score bounded by one SEM represents a 68 percent probability that, over repeated observations, an examinee’s true score estimate falls within the band. A two-SEM boundary represents a 95 percent probability that, over repeated observations, an examinee’s true score estimate falls within the band. Under Classical Test Theory and traditional item analysis, we obtain the SEM from:

\[ s.e.m. = s_x \sqrt{1-r_{xx}} \]  

(Equation 5)

where: \( s_x \) is the observed score standard deviation, and
\( r_{xx} \) is the reliability estimate or alpha coefficient.

In the item response theory (IRT) framework, SEM is estimated as a function of measured ability, and thus is often referred to as a conditional standard error of measurement (CSEM). CSEMs typically are smaller in scale score units towards the center of the scale where
there are more items and more test information and larger at the extremes where there are fewer items and less test information.

Table 13 includes the 2009 Grade 4 standard errors of measurement for the WASL Reading and Mathematics tests on the raw score metric. However, Tables 18 through 21 (in Part 5) list the conditional standard errors of measurement for the WASL Grade 4 Reading and Mathematics tests on the scaled score metric. Table 22 also includes the 2009 Grade 4 conditional standard errors of measurement for the WASL Writing test on the raw score metric.

### Table 13. 2009 Grade 4 Test & Content Strand Reliability Estimates

<table>
<thead>
<tr>
<th>Strand</th>
<th>N</th>
<th>Maximum Possible Raw Score</th>
<th>Alpha Coefficient</th>
<th>Raw Score Standard Error of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>75,823</td>
<td>36</td>
<td>0.86</td>
<td>2.48</td>
</tr>
<tr>
<td>CMP</td>
<td>75,823</td>
<td>18</td>
<td>0.75</td>
<td>1.80</td>
</tr>
<tr>
<td>ALY</td>
<td>75,823</td>
<td>18</td>
<td>0.74</td>
<td>1.71</td>
</tr>
<tr>
<td>Writing</td>
<td>74,358</td>
<td>12</td>
<td>0.77</td>
<td>1.02</td>
</tr>
<tr>
<td>COS</td>
<td>74,358</td>
<td>8</td>
<td>0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>CONV</td>
<td>74,358</td>
<td>4</td>
<td>0.72</td>
<td>0.52</td>
</tr>
<tr>
<td>Mathematics</td>
<td>76,083</td>
<td>38</td>
<td>0.87</td>
<td>2.69</td>
</tr>
<tr>
<td>CT</td>
<td>76,083</td>
<td>28</td>
<td>0.83</td>
<td>2.29</td>
</tr>
<tr>
<td>PC</td>
<td>76,083</td>
<td>10</td>
<td>0.63</td>
<td>1.41</td>
</tr>
</tbody>
</table>

ALY (RD) - Analysis  
CMP (RD) - Comprehension  
COS (WR) - Content, Organization, & Style  
CONV (WR) - Writing Conventions  
CT (MA) - Content  
PC (MA) - Process
INTERJUDGE AGREEMENT

Another aspect of reliability is interrater reliability. Interrater agreement is an important facet for the consistent application of scoring standards and the subsequent reliability of test scores because constructed response items are scored by trained human readers. When two trained judges independently assign the same score to a student’s item response, this is evidence of the consistent application of a scoring standard. The evidence is strengthened when it can be replicated by increasing the numbers of different items, judges, students’ responses, and ranges of item score points. The quality of interrater reliability can be evaluated empirically in three ways:

1. Percent agreement between two readers;
2. Validity paper hit rates or percent agreement for a reader on validity paper sets; and

Percent agreement between two readers is frequently defined as the percent of exact score and adjacent score agreement. Percent of exact score agreement is a stringent criterion which tends to decrease with increasing numbers of item score points. The fewer the item score points, the fewer degrees of freedom on which two readers can vary, and the higher the percent of agreement. WASL scores must be scored to satisfy a pre-defined level of exact + adjacent score agreement. Tables 26 through 28 provide information on the agreement rates and can be found in Part 7.

Validity papers are student papers that, according to a panel of trained content and scoring professionals, represent specific item score points. Validity sets represent the full range of item score points, as well as a range of performance within a given item score point (e.g., “high” 2-point papers, “low” 2-point papers, and mid-range 2-point papers to reflect the full range of a “2” item score point). These validity sets are imbedded throughout the operational scoring process to monitor rater drift to provide rater intervention and retraining or recalibration as necessary.

The Kappa coefficient is an index of interrater reliability that incorporates a correction for the rate of chance agreement. Kappa is computed by:

\[ \kappa = \frac{p_x - p_r}{1 - p_r} \]  

(Equation 6)

where \( p_x \) = overall proportion of exact agreement

\( p_r \) = overall proportion of chance agreement = \( \sum_{i=1}^{m} p_x p_x \), for item score points \( i \) to \( m \).

Tables 26 through 28 provide information on the agreement rates and can be found in Part 7.
DECISION CONSISTENCY AND ACCURACY

Analyses were performed, using the computer program RelClass (ETS proprietary software), to estimate the accuracy and consistency of decisions about meeting standard on the WASL. The methods described by Livingston and Lewis (1995), and Young and Yoon (1998) were applied to complete these analyses.

Every discrete test administration will result in some error in the classification of examinees. When an assessment uses performance classifications as the primary method to report test results, accuracy and consistency of decisions become important indicators about the quality of the assessment. This section includes the results of decision consistency and accuracy analyses for the WASL tests administered in Spring 2009.

The accuracy of decisions is represented by the agreement between the classifications based on students’ observed scores on the actual test form and the classifications that would have been made based on students’ true scores. True scores are assumed to be errorless but are not a known entity. They can, however, be estimated based on the expected values of test scores over all possible forms of the test. A false positive decision results when a true score corresponds to a classification below a critical cut score (e.g., “does not meet standard”), but the observed score corresponds to a “meets standard” classification. A false negative decision results when a true score “meets standard”, but the observed score corresponds to a “does not meet standard” classification. Decision consistency is the agreement between two non-overlapping and equally difficult forms of the test. This index is estimated using response data from the actual test form and a hypothetical alternate form, based on the actual test form’s estimated reliability.

For each WASL test, the decision consistency and accuracy table includes the proportion of:
- Overall accurate classifications;
- False positives for accurate classifications;
- False negatives for accurate classifications;
- Overall consistent classifications;
- False positives for consistent classifications;
- False negatives for consistent classifications;
- Accuracy around critical cut point (“meets standard” vs. “does not meet standard”);
- Consistency around critical cut point (“meets standard” vs. “does not meet standard”).

A classification accuracy table is a cross-tabulation of the true score vs. observed score classifications. A classification consistency table is a cross-tabulation of the observed score vs. hypothetical alternate form score classifications.

The proportion of overall accuracy and consistency classifications is computed as the sum of the diagonal cell entries (agreement between observed & true score decisions for accuracy; agreement between observed & hypothetical alternate form score decisions for consistency).
Accuracy and consistency classifications around a critical cut point (e.g., “meets standard” vs. “does not meet standard”) is similarly computed by collapsing all classification decisions into a dichotomized distribution around the critical cut point. For WASL assessments, “below basic” and “basic” performance levels result in a “does not meet standard” classification; “proficient” and “advanced” performance levels result in the “meets standard” classification.

**Figure 6. Accuracy or Consistency Around Critical Cut Point**

<table>
<thead>
<tr>
<th></th>
<th>Below Basic</th>
<th>Basic</th>
<th>Proficient</th>
<th>Advanced</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td></td>
<td></td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results for the Spring 2009 administration are provided in Table 14. Decision accuracy, based on errorless true score classification, is typically higher than decision consistency, which is based on two types of test scores that both contain measurement error.
<table>
<thead>
<tr>
<th>Subject</th>
<th>N</th>
<th>Accuracy</th>
<th>Consistency</th>
<th>Cut Point Accuracy</th>
<th>Cut Point Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>False Positive</td>
<td>False Negative</td>
<td>Overall</td>
</tr>
<tr>
<td>Reading</td>
<td>75,823</td>
<td>0.76</td>
<td>0.12</td>
<td>0.13</td>
<td>0.66</td>
</tr>
<tr>
<td>Writing</td>
<td>74,358</td>
<td>0.64</td>
<td>0.18</td>
<td>0.18</td>
<td>0.53</td>
</tr>
<tr>
<td>Mathematics</td>
<td>76,083</td>
<td>0.71</td>
<td>0.14</td>
<td>0.15</td>
<td>0.62</td>
</tr>
</tbody>
</table>
SUMMARY

The interrater data indicate that scorers applied consistent scoring standards defined by the scoring rubrics. The alpha coefficients for overall content area tests and by content area strands reveal acceptable levels of internal consistency, supporting the intention for selected item sets to measure a related construct (see Table 13). Decision accuracy indices around the critical cut-point for all Grade 4 tests are in the mid 80s to low 0.90s, with corresponding decision consistency indices in the 0.80s (see Table 14). The conditional standard errors of measurement, however, are sufficiently large to warrant judicious interpretation when evaluating test scores and making decisions about individual student scores (see Table 13 and Tables 18 through 22 in Part 5).

REFERENCES


PART 5: SCALING AND EQUATING

The 2009 Grade 4 Reading, Writing, and Mathematics WASL item data and test scores were scaled to the results from the 2004 standards revisiting. Although very few adjustments to the standards were recommended, adopting those recommendations redefined the baseline scale from the initial 1999 definition to the scale defined in 2004 from standards revisiting.

All WASL tests are scaled so that a scaled score of 400 is the cut score for Level 3 or “Proficient” and a scaled score of 375 is the cut score for Level 2 or “Basic.” To “meet standard,” students must either be Level 3 (Proficient) or Level 4 (Advanced).

Prior to equating, all operational items are analyzed using the traditional and IRT analyses as described in Part 2. The traditional item analyses are used to verify answer keys and to evaluate item performance. Items with a p-value less than 0.25 or a point-biserial value less than 0.10 were reviewed by content specialists and psychometricians, prior to equating and all items were found to be acceptable for operational use.

SCALED SCORE DEVELOPMENT

Scores on the WASL are reported as scaled scores. Tables 18 and 21 provide the 2009 Grade 4 number correct to scaled scores conversions for each test used for scoring. The Rasch model and Master’s (1982) Partial Credit Model produce in an equal interval scale, much like a ruler marked in inches or centimeters, for each test for which items and student scores can be reported. The Partial Credit Model (PCM) accommodates polytomously scored constructed-response items. Calibrating a test with the PCM produces estimated parameters for item difficulty and the difficulty of item score points or steps. The scaled score range for each test is sufficient to describe levels of performance from the lowest possible earned scaled score to the highest possible earned scaled score, across all content areas tested.

Item Response Theory (IRT) uses mathematical models to describe the probability of choosing a response category as a function of a latent trait and item parameters. IRT models can be specified by three item parameters: item difficulty, item discrimination, and a “guessing” parameter. The Rasch and PCM models are one class of IRT models that also specifies theta (θ) for examinees. Rasch models do not explicitly parameterize item discrimination or guessing parameters (although empirical item discrimination and “guessing” can be evaluated by characteristics of Rasch fit statistics). This means that, unlike more complicated IRT models, there is a one-to-one relationship between the number correct score on a test and the θ score on the test.

Once θ scores are estimated, it is general practice to linearly transform θ to a positive whole number scale. The linear transformation preserves the original shape of the distribution, facilitates group-level computations, and conveys information about an ability scale that is intuitively more clear and accessible to non-technical audiences.
Because the scaled scores are on an equal interval scale, it is possible to compare score performance at different points on the scale. Much like a yardstick, differences are constant at different measurement points. For example, a difference of 2 inches between 12 and 14 inches is the same differences as a difference of 2 inches between 30 and 32 inches. Two inches is two inches. Similarly, for equal interval achievement scales, a difference of 20 scaled score points between 360 and 380 means the same difference in achievement as a difference of 400 and 420, except that the difference is in degree of achievement rather than length.

One limitation of scaled scores is that they are not well suited to making score interpretations beyond “how much more” and “how much less.” Administrators, parents, and students ask, “What score is good enough? How do we compare with other schools like ours? Is a 40 point difference between our school and another school a meaningful difference?” For this reason, scaled scores are usually interpreted by using performance standards or by converting them to percentile ranks.

Based on the content of the WASL, committees set the performance standards for each content area test that would represent acceptable performance for a well taught, hard working Grade 4 student. Standard setting committees also identified two performance levels below standard (Level 1 = Below Basic; Level 2 = Basic) and one above standard (Level 4 = Advanced).

The standard setting procedures identified the θ values associated with each committee’s recommended cut-score (i.e., the “Below Basic”/”Basic”, “Basic”/”Proficient”, and “Proficient”/”Advanced” cuts). These θ values defined the linear transformation system to derive scaled scores. To maintain the raw score to θ relationship, any two points on the θ scale can be fixed to any two specified scaled scores to define the linear transformation.

Following the standard setting and the standard revisiting process, a linear transformation was defined to convert the θ scores to a whole number scaled score. For all tests, the θ score from baseline associated with Level 3 “Proficient” was fixed to a WASL scaled score of 400. The θ score identified as Level 2 “Basic” was fixed to a WASL scaled score of 600.

---

6 Following are the general descriptions of the performance levels established for the WASL:

Level 4 – Advanced: This level represents superior performance, notably above that required for meeting the standard at Grade 4.

Level 3 – Proficient: This level represents solid academic performance for Grade 4. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 – Basic: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at Grade 4.

Level 1 – Below Basic: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at Grade 4.

In all content areas, the standard (Level 3) reflects what a well taught, hard working student should know and be able to do.
375. All θ scores are translated to scaled scores by specific linear transformation equations for each grade level content area test. The Level 4 “Advanced” scaled score varies by content area.

The general form of a linear equation of θ to scaled score is:

\[ a \theta + b = \text{scaled score} \]  

(Equation 7)

where \( a \) is the slope and \( b \) is the intercept of the linear transformation to scaled scores.

Because two points define any line, the linear transformation equation is defined by simultaneously solving the system of two equations for constants \( a \) and \( b \):

\[ a \theta \text{ associated with Level 3 “Proficient”} + b = 400 \]  

(Equation 8)

\[ a \theta \text{ associated with Level 2 “Basic”} + b = 375 \]

Table 15 lists the theta values at Level 2 “Basic” and Level 3 “Proficient” from the applicable baseline year used to define the θ to scaled score linear transformation equations for each content area. Because θ is equated to the baseline year θ scale, the same linear transformation is used from year to year until existing standards are revisited or new standards are set.

### Table 15. Theta to Scaled Score Linear Transformation Equations

<table>
<thead>
<tr>
<th>Content Area</th>
<th>θ at Level 2 “Basic” (Scaled Score 375)</th>
<th>θ at Level 3 “Proficient” (Scaled Score 400)</th>
<th>θ to Scaled Score Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>-0.331</td>
<td>0.952</td>
<td>Scaled Score = 19.4856*θ + 381.4497</td>
</tr>
<tr>
<td>Writing †</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>-0.090</td>
<td>0.572</td>
<td>Scaled Score = 37.76435*θ + 378.3988</td>
</tr>
</tbody>
</table>

† Writing results are reported on the total raw score metric.

In Reading and Mathematics, scaled scores below 375 are assigned to the Level 1 “Below Basic” performance level category. Scaled scores between 375 and 399, inclusive, are assigned to the Level 2 “Basic” category. Scaled score ranges assigned to the Level 3 “Proficient” category and Level 4 “Advanced” category varies according to content area test as illustrated in Table 16 below.

### Table 16. Scaled Score Ranges for Performance Level Categories

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Level 1 “Below Basic”</th>
<th>Level 2 “Basic”</th>
<th>Level 3 “Proficient”</th>
<th>Level 4 “Advanced”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>275-374</td>
<td>375-399</td>
<td>400-423</td>
<td>424-475</td>
</tr>
<tr>
<td>Writing †</td>
<td>0-6</td>
<td>7-8</td>
<td>9-10</td>
<td>11-12</td>
</tr>
<tr>
<td>Mathematics</td>
<td>200-374</td>
<td>375-399</td>
<td>400-426</td>
<td>427-550</td>
</tr>
</tbody>
</table>

† Writing results are reported on the total raw score metric.
CUT POINTS FOR CONTENT STRANDS

Cut points for content strands in Reading and Mathematics are defined relative to the total content area scale using the following steps. Writing tests are not equated from year to year, and strand scores are not provided for Writing.

1. Content area operational items are scaled and calibrated;
2. All candidate anchor items on the operational test are subjected to a stability analysis to determine the final anchor item set in the year-to-year common item equating;
3. Operational items are calibrated with the final anchor item set;
   [Further details about Steps 1-3 above are described in the annual equating reports, WASL Grade 3-8 Reading 2009 Equating Study Technical Report, and WASL Grade 3-8 Mathematics 2009 Equating Study Technical Report.]
4. Item parameter estimates resulting from Step 3 are used to score operational items specific to each content strand. This step produces a raw score-to-Θ table for each content strand;
5. Strand score Θs greater than or equal to the Level 3 “Proficient” Θ cut point (scaled score 400) from the baseline year is the “+/-” content strand cut point.

Table 23 lists the strand score and strand Θ ranges, and the raw cut points that operationalize the “+/-” content strand cut point. The Writing test is not equated from year to year on a scale score metric, and therefore have no corresponding “+/-” content strand cut points. Tables 24 and 25 contain the content strand cut points for the Mathematics Braille and Large Print forms, respectively.

Figure 7 is a hypothetical distribution of item difficulties for Mathematics strand items, illustrating how the range of item difficulties can differ for each strand. What may be less apparent is that the number of items below and above the Θ value of 0.572 (the Θ for Mathematics Level 3 “Proficient” from baseline 2003-04, standards revisiting) can also vary by strand. This example highlights differences between strand difficulties and a caution when interpreting strand-level results based on a limited sample of items from a strand domain.

Figure 7. Hypothetical Range of Mathematics Strand Item Difficulties (Θ)
The Writing test includes two strands from each of two writing prompts. Relatively few total score points on the total test limit the utility of explicitly equating test scores from year to year. All scaling was completed in the baseline year on the raw score scale. Performance level results on the raw score scale are applied to scored results from year to year.

Following standard setting in the baseline year, cut-scores for the two Writing strands were defined using a contrasting groups method. Total Writing scores were divided into two groups – those that “Meets Standard” and those that did not. For each group, raw strand score frequency distributions for Writing Content, Organization, and Style (COS) and for Writing Mechanics (CONV) were examined. Strand score cut-points were defined as the point with minimal overlap between the distributions of the two groups (see Figure 8).

**Figure 8. Sample Score Distribution of Contrasting Groups – COS Strand**

![Score Distribution Graph](image)

**EQUATING**

Reading and Mathematics tests were equated using similar designs and procedures. Multiple-choice, short-answer, and extended-response items in the first operational year were calibrated and scaled using the PCM to define the baseline scale.

To equate the second year operational test to the first year operational test and the baseline scale, an anchor item set was used to link tests between administration years. “Test” refers to the set of operational items administered to all students that contribute to reported scores. The anchor item set is first subjected to a stability analysis before proceeding with anchor item equating. This procedure enables equating operational test scores from year to year and enables initial calibration and scaling of imbedded pilot items to the baseline scale. This general design and procedure is replicated from year to year to equate current test scores to the baseline scale.

The equating is completed on a sample of ~10,000 available scored student records for each content area test. Logistic, operational processing, and score reporting schedules necessitate the completion of equating on a sample of the statewide population before the completion of scoring. OSPI and the previous vendor initiated a concerted effort in 2006 to enhance consistent statewide representation in the equating sample from year to year. Geographic region, population density, building enrollment type, grade level enrollments, ethnic minority composition, and past WASL achievement were included in the statewide sampling framework. Several equivalent samples of school rosters were developed from the statewide...
sampling framework for annual use on a rotating basis. The intention is to prioritize processing and scoring of identified schools on an annual early-return roster for inclusion in the final equating sample.

The operational item parameters resulting from the equating are used to develop the raw score to scale score conversion tables (see Table 18-22), which are used for score reporting. These tables are also used to identify the raw score cuts for each performance level.

The empirical weighted mean Rasch values and the corresponding raw score proficient cut scores based on the 2009 equating results are provided in Table 17 for the Reading and Mathematics tests. The cut scores are the raw scores required for students to be classified as “Meeting Standards” (Proficient and Advanced performance levels). The percent correct is simply the cut score as a percent of the total possible raw score. For the purpose of comparison, Table 17 also provides the same information for the tests administered from 2000 through 2008 as reported in the previous technical reports.

Further details are described in the *WASL 2009 Grade 3-8 Equating Study Technical Reports* and previous annual equating study technical reports.

**Equating the Writing Test**

For Writing, writing prompts were selected for the 2008 WASL that were of similar difficulty and purpose as those from the 2001 WASL. These prompt characteristics were evaluated from a stand-alone pilot administration from which Writing prompt pairs are selected and reserved for future operational use. The Grade 4 Writing Pilot was last administered in Fall 2006 to develop the current bank of operational prompt pairs.
Table 17. Empirical Weighted Mean Rasch of 2000 ~ 2009 Grade 4 Reading & Mathematics Tests

<table>
<thead>
<tr>
<th>Subject</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Mean Rasch</td>
<td>0.43</td>
<td>0.62</td>
<td>-0.01</td>
<td>0.04</td>
<td>0.44</td>
<td>0.07</td>
<td>0.27</td>
<td>0.38</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Cut Score</td>
<td>28 out of 43</td>
<td>24 out of 40</td>
<td>30 out of 40</td>
<td>28 out of 40</td>
<td>27 out of 40</td>
<td>30 out of 40</td>
<td>31 out of 42</td>
<td>28 out of 42</td>
<td>31 out of 42</td>
</tr>
<tr>
<td></td>
<td>Percent Correct</td>
<td>65.1%</td>
<td>60.0%</td>
<td>75.0%</td>
<td>70.0%</td>
<td>67.5%</td>
<td>75.0%</td>
<td>73.8%</td>
<td>66.7%</td>
<td>73.0%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Mean Rasch</td>
<td>0.37</td>
<td>0.23</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.02</td>
<td>0.21</td>
<td>0.11</td>
<td>-0.13</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>Cut Score</td>
<td>35 out of 62</td>
<td>33 out of 56</td>
<td>36 out of 55</td>
<td>34 out of 54</td>
<td>35 out of 55</td>
<td>33 out of 55</td>
<td>35 out of 55</td>
<td>36 out of 55</td>
<td>37 out of 55</td>
</tr>
<tr>
<td></td>
<td>Percent Correct</td>
<td>56.5%</td>
<td>58.9%</td>
<td>67.3%</td>
<td>63.0%</td>
<td>63.6%</td>
<td>60.0%</td>
<td>63.6%</td>
<td>65.5%</td>
<td>66.3%</td>
</tr>
</tbody>
</table>
NUMBER CORRECT SCORES TO SCALED SCORES

The raw score to scaled score relationship on each WASL test varies from year to year as a function of the particular operational items that comprise a test. The underlying scale and scaled score interpretations are the same from year to year until standards are revisited or new standards are defined.

Tables 18 to 21 include the raw score (Raw) to scaled score (SS) relationship for the 2009 Grade 4 Reading and Mathematics tests used for scoring. Scaled scores that were adjusted to reflect the minimum or maximum possible scaled score or the cut score at each proficiency level have been identified with an asterisk (*). Because the Writing test is already “scaled” on the raw score metric, there is no raw score to SS relationship. Table 22 lists the conditional standard errors of measurement at each Writing raw score point.

The raw score to scaled score tables for Reading and Mathematics are based on the equating files provided by DRC. 7 Students were included in the equating analyses, and therefore the production of the raw score to scaled score tables, if they met the following criteria:

a) Student’s reporting grade is 4;
   b) Student’s test type was either WABA or WASL;
   c) Student’s test attempt value was TS (Tested);
   d) Student responded to two or more items per test section; and
   e) For Mathematics only, student did not take the Braille or Large Print version. 8

Table 23 lists the strand score and strand 0 ranges, and the raw cut points that operationalize the “+/−” content strand cut point for Reading and Mathematics. Tables 24 and 25 contain the content strand cut points for the Mathematics Braille and Large Print forms, respectively.

---

7 The equating files for Reading and Mathematics were received on June 16, 2009.
8 Some operational items were excluded from the Mathematics Braille and Large Print forms, thus special raw score to scaled score relationship tables were created for these forms. All items were included in the Reading and Science Braille and Large Print forms, therefore no special scoring tables were produced for Reading or Science.
### Table 18. 2009 Grade 4 Reading Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (SEM)

<table>
<thead>
<tr>
<th>Raw</th>
<th>Reading SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>275*</td>
<td>35.892</td>
</tr>
<tr>
<td>1</td>
<td>311</td>
<td>20.051</td>
</tr>
<tr>
<td>2</td>
<td>326</td>
<td>14.556</td>
</tr>
<tr>
<td>3</td>
<td>335</td>
<td>12.178</td>
</tr>
<tr>
<td>4</td>
<td>341</td>
<td>10.776</td>
</tr>
<tr>
<td>5</td>
<td>347</td>
<td>9.860</td>
</tr>
<tr>
<td>6</td>
<td>352</td>
<td>9.197</td>
</tr>
<tr>
<td>7</td>
<td>356</td>
<td>8.691</td>
</tr>
<tr>
<td>8</td>
<td>359</td>
<td>8.301</td>
</tr>
<tr>
<td>9</td>
<td>363</td>
<td>7.989</td>
</tr>
<tr>
<td>10</td>
<td>366</td>
<td>7.736</td>
</tr>
<tr>
<td>11</td>
<td>369</td>
<td>7.541</td>
</tr>
<tr>
<td>12</td>
<td>372</td>
<td>7.385</td>
</tr>
<tr>
<td>13</td>
<td>375*</td>
<td>7.249</td>
</tr>
<tr>
<td>14</td>
<td>377</td>
<td>7.151</td>
</tr>
<tr>
<td>15</td>
<td>380</td>
<td>7.073</td>
</tr>
<tr>
<td>16</td>
<td>382</td>
<td>7.015</td>
</tr>
<tr>
<td>17</td>
<td>385</td>
<td>6.976</td>
</tr>
<tr>
<td>18</td>
<td>387</td>
<td>6.956</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw</th>
<th>Reading SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>390</td>
<td>6.956</td>
</tr>
<tr>
<td>20</td>
<td>392</td>
<td>6.976</td>
</tr>
<tr>
<td>21</td>
<td>395</td>
<td>7.015</td>
</tr>
<tr>
<td>22</td>
<td>397</td>
<td>7.093</td>
</tr>
<tr>
<td>23</td>
<td>400</td>
<td>7.171</td>
</tr>
<tr>
<td>24</td>
<td>403</td>
<td>7.288</td>
</tr>
<tr>
<td>25</td>
<td>405</td>
<td>7.424</td>
</tr>
<tr>
<td>26</td>
<td>408</td>
<td>7.599</td>
</tr>
<tr>
<td>27</td>
<td>411</td>
<td>7.833</td>
</tr>
<tr>
<td>28</td>
<td>415</td>
<td>8.125</td>
</tr>
<tr>
<td>29</td>
<td>418</td>
<td>8.496</td>
</tr>
<tr>
<td>30</td>
<td>424*</td>
<td>8.983</td>
</tr>
<tr>
<td>31</td>
<td>426</td>
<td>9.645</td>
</tr>
<tr>
<td>32</td>
<td>432</td>
<td>10.561</td>
</tr>
<tr>
<td>33</td>
<td>438</td>
<td>11.945</td>
</tr>
<tr>
<td>34</td>
<td>447</td>
<td>14.341</td>
</tr>
<tr>
<td>35</td>
<td>461</td>
<td>19.895</td>
</tr>
<tr>
<td>36</td>
<td>475*</td>
<td>35.795</td>
</tr>
</tbody>
</table>

*Note.* * represents adjusted scaled scores.
Table 19. 2009 Grade 4 Mathematics Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (SEM)

<table>
<thead>
<tr>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200*</td>
<td>69.600</td>
<td>0</td>
<td>200*</td>
<td>69.600</td>
</tr>
<tr>
<td>1</td>
<td>224</td>
<td>38.897</td>
<td>21</td>
<td>382</td>
<td>13.029</td>
</tr>
<tr>
<td>2</td>
<td>252</td>
<td>28.210</td>
<td>22</td>
<td>386</td>
<td>13.142</td>
</tr>
<tr>
<td>3</td>
<td>270</td>
<td>23.565</td>
<td>23</td>
<td>391</td>
<td>13.331</td>
</tr>
<tr>
<td>4</td>
<td>283</td>
<td>20.846</td>
<td>24</td>
<td>396</td>
<td>13.520</td>
</tr>
<tr>
<td>5</td>
<td>293</td>
<td>19.033</td>
<td>25</td>
<td>400</td>
<td>13.520</td>
</tr>
<tr>
<td>6</td>
<td>302</td>
<td>17.711</td>
<td>26</td>
<td>405</td>
<td>13.822</td>
</tr>
<tr>
<td>7</td>
<td>310</td>
<td>16.730</td>
<td>27</td>
<td>411</td>
<td>14.162</td>
</tr>
<tr>
<td>8</td>
<td>317</td>
<td>15.937</td>
<td>28</td>
<td>416</td>
<td>14.615</td>
</tr>
<tr>
<td>9</td>
<td>323</td>
<td>15.295</td>
<td>29</td>
<td>427*</td>
<td>15.144</td>
</tr>
<tr>
<td>10</td>
<td>329</td>
<td>14.804</td>
<td>30</td>
<td>428</td>
<td>15.823</td>
</tr>
<tr>
<td>12</td>
<td>340</td>
<td>14.011</td>
<td>32</td>
<td>443</td>
<td>17.711</td>
</tr>
<tr>
<td>13</td>
<td>345</td>
<td>13.746</td>
<td>33</td>
<td>452</td>
<td>19.147</td>
</tr>
<tr>
<td>14</td>
<td>350</td>
<td>13.482</td>
<td>34</td>
<td>463</td>
<td>21.035</td>
</tr>
<tr>
<td>15</td>
<td>355</td>
<td>13.293</td>
<td>35</td>
<td>476</td>
<td>23.867</td>
</tr>
<tr>
<td>16</td>
<td>360</td>
<td>13.180</td>
<td>36</td>
<td>494</td>
<td>28.550</td>
</tr>
<tr>
<td>17</td>
<td>364</td>
<td>13.066</td>
<td>37</td>
<td>523</td>
<td>39.237</td>
</tr>
<tr>
<td>18</td>
<td>369</td>
<td>12.991</td>
<td>38</td>
<td>550*</td>
<td>69.826</td>
</tr>
<tr>
<td>19</td>
<td>375*</td>
<td>12.953</td>
<td>19</td>
<td>375*</td>
<td>12.953</td>
</tr>
</tbody>
</table>
Table 20. 2009 Grade 4 Mathematics Braille Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (SEM)

<table>
<thead>
<tr>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200*</td>
<td>69.675</td>
</tr>
<tr>
<td>1</td>
<td>232</td>
<td>39.048</td>
</tr>
<tr>
<td>2</td>
<td>260</td>
<td>28.399</td>
</tr>
<tr>
<td>3</td>
<td>278</td>
<td>23.754</td>
</tr>
<tr>
<td>4</td>
<td>291</td>
<td>21.073</td>
</tr>
<tr>
<td>5</td>
<td>302</td>
<td>19.260</td>
</tr>
<tr>
<td>6</td>
<td>311</td>
<td>17.976</td>
</tr>
<tr>
<td>7</td>
<td>319</td>
<td>17.032</td>
</tr>
<tr>
<td>8</td>
<td>327</td>
<td>16.314</td>
</tr>
<tr>
<td>9</td>
<td>333</td>
<td>15.748</td>
</tr>
<tr>
<td>10</td>
<td>340</td>
<td>15.295</td>
</tr>
<tr>
<td>11</td>
<td>346</td>
<td>14.955</td>
</tr>
<tr>
<td>12</td>
<td>352</td>
<td>14.728</td>
</tr>
<tr>
<td>13</td>
<td>357</td>
<td>14.539</td>
</tr>
<tr>
<td>14</td>
<td>363</td>
<td>14.426</td>
</tr>
<tr>
<td>15</td>
<td>368</td>
<td>14.388</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>375*</td>
<td>14.388</td>
</tr>
<tr>
<td>17</td>
<td>379</td>
<td>14.464</td>
</tr>
<tr>
<td>18</td>
<td>385</td>
<td>14.615</td>
</tr>
<tr>
<td>19</td>
<td>391</td>
<td>14.841</td>
</tr>
<tr>
<td>20</td>
<td>400*</td>
<td>15.106</td>
</tr>
<tr>
<td>21</td>
<td>403</td>
<td>15.483</td>
</tr>
<tr>
<td>22</td>
<td>409</td>
<td>15.974</td>
</tr>
<tr>
<td>23</td>
<td>416</td>
<td>16.616</td>
</tr>
<tr>
<td>24</td>
<td>427*</td>
<td>17.409</td>
</tr>
<tr>
<td>25</td>
<td>433</td>
<td>18.429</td>
</tr>
<tr>
<td>26</td>
<td>442</td>
<td>19.789</td>
</tr>
<tr>
<td>27</td>
<td>454</td>
<td>21.677</td>
</tr>
<tr>
<td>28</td>
<td>468</td>
<td>24.471</td>
</tr>
<tr>
<td>29</td>
<td>486</td>
<td>29.116</td>
</tr>
<tr>
<td>30</td>
<td>516</td>
<td>39.690</td>
</tr>
<tr>
<td>31</td>
<td>550*</td>
<td>70.091</td>
</tr>
</tbody>
</table>

Table 21. 2009 Grade 4 Mathematics Large Print Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (SEM)

<table>
<thead>
<tr>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200*</td>
<td>69.675</td>
</tr>
<tr>
<td>1</td>
<td>227</td>
<td>39.011</td>
</tr>
<tr>
<td>2</td>
<td>255</td>
<td>28.399</td>
</tr>
<tr>
<td>3</td>
<td>273</td>
<td>23.792</td>
</tr>
<tr>
<td>4</td>
<td>286</td>
<td>21.073</td>
</tr>
<tr>
<td>5</td>
<td>297</td>
<td>19.260</td>
</tr>
<tr>
<td>6</td>
<td>306</td>
<td>17.976</td>
</tr>
<tr>
<td>7</td>
<td>314</td>
<td>16.994</td>
</tr>
<tr>
<td>8</td>
<td>321</td>
<td>16.201</td>
</tr>
<tr>
<td>9</td>
<td>328</td>
<td>15.597</td>
</tr>
<tr>
<td>10</td>
<td>334</td>
<td>15.106</td>
</tr>
<tr>
<td>11</td>
<td>340</td>
<td>14.728</td>
</tr>
<tr>
<td>12</td>
<td>346</td>
<td>14.388</td>
</tr>
<tr>
<td>13</td>
<td>351</td>
<td>14.162</td>
</tr>
<tr>
<td>14</td>
<td>356</td>
<td>13.973</td>
</tr>
<tr>
<td>15</td>
<td>362</td>
<td>13.822</td>
</tr>
<tr>
<td>16</td>
<td>367</td>
<td>13.708</td>
</tr>
<tr>
<td>17</td>
<td>375*</td>
<td>13.671</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw</th>
<th>Mathematics SS</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>376</td>
<td>13.671</td>
</tr>
<tr>
<td>19</td>
<td>381</td>
<td>13.708</td>
</tr>
<tr>
<td>20</td>
<td>386</td>
<td>13.822</td>
</tr>
<tr>
<td>21</td>
<td>392</td>
<td>13.973</td>
</tr>
<tr>
<td>22</td>
<td>400*</td>
<td>14.199</td>
</tr>
<tr>
<td>23</td>
<td>402</td>
<td>14.502</td>
</tr>
<tr>
<td>24</td>
<td>408</td>
<td>14.879</td>
</tr>
<tr>
<td>25</td>
<td>414</td>
<td>15.408</td>
</tr>
<tr>
<td>26</td>
<td>427*</td>
<td>16.050</td>
</tr>
<tr>
<td>27</td>
<td>428</td>
<td>16.881</td>
</tr>
<tr>
<td>28</td>
<td>436</td>
<td>17.938</td>
</tr>
<tr>
<td>29</td>
<td>445</td>
<td>19.373</td>
</tr>
<tr>
<td>30</td>
<td>456</td>
<td>21.337</td>
</tr>
<tr>
<td>31</td>
<td>469</td>
<td>24.169</td>
</tr>
<tr>
<td>32</td>
<td>488</td>
<td>28.890</td>
</tr>
<tr>
<td>33</td>
<td>517</td>
<td>39.539</td>
</tr>
<tr>
<td>34</td>
<td>550*</td>
<td>70.015</td>
</tr>
</tbody>
</table>
Table 22. 2009 Grade 4 Writing Raw Scores (Raw) with Conditional Standard Errors of Measurement (SEM)

<table>
<thead>
<tr>
<th>Raw</th>
<th>Conditional SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.553</td>
</tr>
<tr>
<td>1</td>
<td>1.013</td>
</tr>
<tr>
<td>2</td>
<td>0.815</td>
</tr>
<tr>
<td>3</td>
<td>0.729</td>
</tr>
<tr>
<td>4</td>
<td>0.715</td>
</tr>
<tr>
<td>5</td>
<td>0.767</td>
</tr>
<tr>
<td>6</td>
<td>0.827</td>
</tr>
<tr>
<td>7</td>
<td>0.780</td>
</tr>
<tr>
<td>8</td>
<td>0.760</td>
</tr>
<tr>
<td>9</td>
<td>0.858</td>
</tr>
<tr>
<td>10</td>
<td>1.127</td>
</tr>
<tr>
<td>11</td>
<td>1.122</td>
</tr>
<tr>
<td>12</td>
<td>1.574</td>
</tr>
</tbody>
</table>

Table 23. 2009 Grade 4 Content Strand Cut-Points

<table>
<thead>
<tr>
<th>Strand</th>
<th>θ Range</th>
<th>Max Raw Strand Score</th>
<th>“-” Strand</th>
<th>“+” Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMP</td>
<td>-3.939 ~ 4.611</td>
<td>18</td>
<td>0 - 11</td>
<td>12 - 18</td>
</tr>
<tr>
<td>ALY</td>
<td>-4.346 ~ 4.679</td>
<td>18</td>
<td>0 - 11</td>
<td>12 - 18</td>
</tr>
<tr>
<td>Mathematics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>-5.133 ~ 4.828</td>
<td>28</td>
<td>0 - 18</td>
<td>19 - 28</td>
</tr>
<tr>
<td>PC</td>
<td>-3.738 ~ 3.592</td>
<td>10</td>
<td>0 - 6</td>
<td>7 - 10</td>
</tr>
</tbody>
</table>
### Table 24. 2009 Grade 4 Mathematics Braille Content Strand Cut-Points

<table>
<thead>
<tr>
<th>Strand</th>
<th>$\theta$ Range</th>
<th>Max Raw Strand Score</th>
<th>“.” Strand</th>
<th>“+” Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics Braille Form</td>
<td>CT</td>
<td>-4.943 ~ 4.714</td>
<td>23</td>
<td>0 - 15</td>
</tr>
<tr>
<td></td>
<td>PC</td>
<td>-3.487 ~ 3.307</td>
<td>8</td>
<td>0 - 5</td>
</tr>
</tbody>
</table>

### Table 25. 2009 Grade 4 Mathematics Large Print Content Strand Cut-Points

<table>
<thead>
<tr>
<th>Strand</th>
<th>$\theta$ Range</th>
<th>Max Raw Strand Score</th>
<th>“.” Strand</th>
<th>“+” Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics Large Print Form</td>
<td>CT</td>
<td>-5.109 ~ 4.744</td>
<td>26</td>
<td>0 - 17</td>
</tr>
<tr>
<td></td>
<td>PC</td>
<td>-3.487 ~ 3.307</td>
<td>8</td>
<td>0 - 5</td>
</tr>
</tbody>
</table>

### Reference

PART 6: ESTABLISHING AND REVISITING STANDARDS

Standard setting for the Grade 4 WASL in Reading, Writing, and Mathematics was conducted in Summer 1997. Standard-setting for the Grades 8 and 10 WASL in Science took place in July 2003. Standard-setting for Science was completed after operational Spring 2003 test administration of the Grades 8 and 10 assessments and after the operational Spring 2004 test administration for Grade 5. Details of the standard setting procedures used for Reading, Mathematics, and Writing can be found in the 1999 through 2003 Washington Assessment of Student Learning Grade 4 Technical Reports. Details of the standard setting procedures used for Grades 8 and 10 Science can be found in the 2003 Washington Assessment of Student Learning Grade 10 Technical Report. The details of the standard setting procedures used for Grade 5 Science can be found in the 2004 Washington Assessment of Student Learning Grade 5 Technical Report.

It is recommended in the research literature that standards should be revisited over time and revised if necessary. Given the tenure of the assessments over a number of years, a history of education reform in the state, the requirements of the 2001 No Child Left Behind Act, and the introduction of high school graduation requirements, OSPI elected to revisit all of the standards for the existing Reading, Writing, and Mathematics tests. The revisiting of standards for Grades 4, 7, and 10 Reading, Writing, and Mathematics occurred in February and March of 2004. The 2004 Washington Assessment of Student Learning Grade 4 Technical Report provides details and results from the standard revisiting process.

The defined performance levels resulting from the initial standard setting and standards revisiting were based on criterion-referenced definitions and interpretations of content area performance. Following standards revisiting, an articulation committee comprised of all WASL content areas and grade levels considered all content/grade level performance levels descriptors, performance level cut points, and impact data in a total assessment system. Based on the standards revisiting recommendations and the articulation committee’s review, subsequent changes to the initial standard setting results were very minimal, lending further credence and validation of the existing standards and assessment system.
PART 7: SCORING THE WASL OPEN-ENDED ITEMS

During item development, item-specific scoring rubrics are written. During item reviews, scoring rubrics are reviewed along with item content. A central aspect of the validity of test scores is the degree to which scoring rubrics are related to the appropriate learning targets or EALRs. A key aspect of reliability is whether scoring rules are applied faithfully during scoring sessions. The following procedures are used to score the WASL items and apply to all content areas that include open-ended questions calling for student-constructed responses. These procedures are used for the full pool of items that were pilot tested as well as for the operational tests.

RANGEFINDING

After student answer documents were received and processed, DRC’s Performance Assessment Services (PAS) staff assembled groups of responses that exemplified the different score points represented in rubrics for Reading, Writing, Mathematics, and Science. Papers were pulled for the 2009 WASL assessment.

Once examples of all the score points were identified, packets, or sets, were put together for each item. These sets were copied for use at rangefinding, held at multiple dates and locations in the Winter and Spring, depending on the subject. The rangefinding committees consisted of Washington State educators, OSPI staff members, ETS Test Development staff, and DRC PAS staff.

Each committee began with a review of the item and the rubric. Copies of the student example sets were presented to the committees, one item at a time. The committees reviewed and scored several student samples together to ensure that everyone was interpreting the rubric consistently. Committee members then went on to score responses independently and those scores were discussed until a consensus was reached. Only responses for which a good agreement rate was attained were used in training the scorers. Discussions of the responses used rubric language, assuring OSPI and all involved that the score point examples clearly illustrated the specific requirements of each score level. DRC PAS staff made notes of how and why the committees arrived at score point decisions, and this information was used by the scoring directors in scorer training.

OSPI, DRC and ETS discussed rubric edits that the committees suggested. Changes were then made by ETS Test Development staff and approved by OSPI, and these final rubrics were used by PAS staff in the training of scorers.

SCORER RECRUITMENT/QUALIFICATIONS

DRC retains a number of experienced scorers from year to year, and those scorers made up approximately 55% of the scorer pool for 2009. To complete the scorer staffing for this project, DRC placed advertisements in local papers diversity publications, and at regional colleges and universities. Recruiting events were held and applications for scorer positions were
screened by DRC recruiting staff. Candidates were personally interviewed and a mandatory, on-demand writing sample, plus a Mathematics sample, were collected along with references and proof of a four-year college degree. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments and with subject specific degrees (Mathematics, Science, teaching, English, journalism, education). Since scorers had to have a strong content-specific background, the scorer pool consisted of educators, writers, editors, and other professionals who were valued for their experience, but who were also required to set aside their own biases about student performance and accept the scoring standards.

LEADERSHIP RECRUITMENT/QUALIFICATIONS

Scoring directors and team leaders were chosen by the content specialists from a pool consisting of experienced individuals who were successful scorers and leaders on previous DRC projects and had strong backgrounds in scoring content-specific projects. Those selected demonstrated strong organization, leadership, and management skills. All scoring directors, team leaders, and scorers were required to sign confidentiality agreements, prior to training with Washington materials or handling of secure materials.

Each room of scorers was assigned a scoring director or assistant scoring director. This individual was monitored by the content specialist and led the hand scoring for the duration of the project. The scoring director assisted in range finding, worked with supervisors to create training materials, conducted the team leader training, and was responsible for training the scorers.

Team leaders assisted the scoring directors/assistant scoring directors with scorer training and monitoring by working with their teams in small group discussions, and answering individual questions that scorers may not have felt comfortable asking in a large group. Once scorers qualified, the team leaders were responsible for maintaining accuracy and workload of team members. The ongoing monitoring identified those scorers who were having difficulty scoring and resulted in the scorer receiving one-on-one retraining. If this process did not correct inaccuracies in scoring, that scorer was released from the project.

TRAINING

To train the 2009 WASL items, DRC’s PAS staff used the approved rubrics and training materials. Responses that were relevant in terms of the scoring concepts they illustrated were annotated for use in the anchor set for Reading, Mathematics and Science. Writing training procedure is discussed more fully in the next section. The item-specific rubrics served as the scorer’s constant reference. Scorers were instructed on how to apply the rubrics, and were required to demonstrate a clear comprehension of each anchor set by performing well on the training materials that were presented for each grade and item.

Team leaders assisted the scoring directors with the training and monitoring of scorers. The scoring director conducted the team leader training prior to the scorer training. This
training followed much the same procedures as the scorer training, but additional time was allotted for review, discussion, and addressing anticipated scorer questions and concerns. Team leaders were required to annotate all of their training responses with the official annotations received from the content committee members at the rangefinding meetings. To facilitate scoring consistency, it is imperative that each team leader imparted the same rationale for each response that other team leaders used. Once the team leaders qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader for 8-10 scorers ensured adequate monitoring of the scorers.

Scorer training began with the scoring director providing an intensive review of the rubric and anchor papers with all scorers to help them internalize the scoring criteria. Next, the scorers “practiced” by independently scoring the responses in the practice set(s). Afterwards, the scoring director and team leaders led a thorough discussion of the set(s) with the entire group. All papers were discussed using the annotations from rangefinding.

Once the scoring guidelines and all the training sets were discussed, scorers were required to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the “true” scores decided upon at rangefinding) on at least one of the qualifying sets. Scorers who failed to achieve the level of agreement determined by OSPI were given additional training to acquire the highest degree of accuracy possible. Scorers who did not perform at the required level of agreement by the end of the qualifying process were not permitted to score “live” student work and were released from the project.

Training is an on-going, continuous process and thus, does not end after the qualifying rounds. There are several reliability checks that are performed throughout the project. Primarily, team leaders monitor scorers’ reliability through “read-behinds.” This is a process whereby team leaders re-read and check scores of each scorer on his team, approximately ten percent of each scorer’s work each day. This is to catch potential scorer drift so that the scorer can have immediate feedback and be retrained in a timely fashion. Scorers are removed from the project if they are unable to score consistently with the rubric and the anchor papers after re-training.

**TYPES OF RESPONSES**

Reading, Mathematics, and Science portions of the WASL contained three types of items: multiple-choice, short-answer (SA) and extended-response (ER). The writing assessment required students to write extended-responses to two different purposes of writing; Narrative and Expository in grade 4 and Expository and Persuasive in grades 7 and High School.

**Multiple-choice Items**

Multiple-choice items required students to select a correct answer from several alternatives. For grades three through five, students generally selected from three alternatives. For grades six through eight and High School, students generally selected from four alternatives.
Each multiple-choice item was scored as right or wrong and had a value of 1 point. Missing responses (items that a student did not answer) and multiple responses were scored as incorrect.

**Short-answer Items**

Short-answer items required a student to respond in a few words or sentences or to demonstrate a process. These items were scored using rubrics written specifically for each item.

**Extended-response Items**

The reading assessment included extended-response items which required the student to write on the basis of one or two passages. Responses were scored if they were legible and on topic and if there was enough original work to be evaluated. Grades 3-5 Reading and Mathematics did not contain extended-response items. The extended-responses for Science consisted of “attribute” items. The scorers were trained on each attribute or characteristic and assigned a score of 0 or 1, depending on if the attribute was present (1) in the response given or whether it was absent (0).

**SCORING FOR WRITING**

Although training to score Writing follows the same process as described above, the approach to scoring varied from other content areas. A “focused holistic” approach was used. This approach assesses writing fluency and measures the degree to which a writer has connected to the “audience”, namely, the scorer. Due to the unique, complex facets of a writing rubric, scorers were asked to take notes, using rubric language, to assist them with internalizing rubric requirements during training rather than having annotations provided in anchor sets. With holistic scoring, the overall effectiveness of the piece is assessed and the score is assigned with the overall quality in mind. All of the elements of that make up a “quality” piece are taken into account; content, organization, style, and mechanics. A domain-based scoring rubric was used to score the writing prompts. A student could earn four points (1-4) for one domain that included Composing, Organization and Style (COS) or two points (0-2) for Conventions (CONV). The High School writing was scored by two scorers. These four scores were combined to provide a maximum of 12 points on any one piece of writing, or a total of 24 points for the entire Writing test. Grades 4 and 7 writing was scored once, with a 10% double read. A total of 6 points could be earned on each piece of writing, yielding a maximum of 12 points for the entire Writing test.
HANDSCORING PROCESS

Student responses were scored independently and by multiple scorers. All responses for Reading, Mathematics, Science and grades 4 and 7 Writing were read once with a 10% double read or read-behind to ensure reliability. High School Writing responses had 100% double read to ensure reliability. The read-behinds were randomly chosen by the imaging system at the item/prompt level.

Scorers scored the imaged responses on PC monitors at the DRC Scoring Centers in Sharonville, Ohio; Minnetonka, Minnesota; and Woodbury, Minnesota. Scorers were seated, in comfortable adjustable chairs, at tables with two imaging stations per table. Image distribution was controlled, thus ensuring that student images were sent to designated groups of scorers qualified to score those items. Scorers read each response and then selected the appropriate score on the scoring screen.

To handle possible alerts (i.e., student responses indicating potential issues related to the student’s safety and well-being that may require attention at the state or local level), the imaging system allowed scorers to forward responses needing attention to the scoring director. These alerts were submitted to the content specialist, and then the director of PAS, who then notified OSPI of this occurrence. At no time did scorers know anything about the student’s personal identity.

Once handscoring was completed, PAS compiled reviews of the anchor items for Reading and Science (Mathematics and Writing did not do reviews of anchor/operational items) for all grade levels. This information was submitted to OSPI content specialists.

QUALITY CONTROL

Scorer accuracy was monitored throughout the scoring session by producing both daily and on-demand reports, ensuring that an acceptable level of scoring accuracy was maintained. Inter-scorer reliability was tracked and monitored with multiple quality control reports that were reviewed by quality assurance analysts. These reports were generated at the handscoring center and were reviewed by the scoring directors, team leaders, content specialists, and project directors. The following reports were used during the scoring of the constructed responses:

The Scorer Monitor Report monitored how often scorers were in exact agreement and ensured that an acceptable agreement rate was maintained. This report provided daily and cumulative exact and adjacent inter-scorer agreement and, in the case of High school Writing, the percentage of responses requiring resolution.

The Score Point Distribution Report monitored the percentage of responses given each of the score points by scorer and then as an average of all scorers. For example, reading daily and cumulative reports showed how many 0s, 1s, 2s, 3s, 4s, and non-score options, a scorer had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each scorer so that production rates could be monitored.
The Item Status Report monitored the progress of handscoring. This report tracked each item and indicated the status (e.g., “needs second reading,” “complete”). This report ensured that all discrepancies were resolved by the end of the project and accounted for all students.

The Read-Behind Log was used by the team leaders/scoring director to monitor scorer reliability. Student responses were randomly selected and team leaders read scored items from each team member. If the team leader disagreed with the scorer's score, remediation occurred. This proved to be a very effective type of feedback because it was done with “live” items scored by a particular scorer.

The Validity Reports tracked how the scorers performed by comparing pre-determined scored responses, approved by OSPI, to scorers’ scores for the same set of responses. If the scorers’ scoring fell outside of a determined percentage of agreement, remediation occurred and additional responses were given to individuals who needed to be monitored more closely.

TRANSITIONING FROM PREVIOUS VENDOR

To make the first year of the contract as cohesive and homogenous a transition as possible, DRC established expectations early, planned for all handscoring activities to mirror as closely as possible previous years and utilized many of the same processes as the previous vendor.

One of the processes DRC utilized to ensure scorer accuracy and detect drift was the validity process. The goal of the validity process is to ensure that scoring standards are maintained. Specifically, the objective is to make sure that scorers rate student responses in a manner consistent with statewide standards within a single administration of the WASL and across consecutive administrations, as well as, particularly in this transition year, from vendor-to-vendor.

The validity selection process began first by identifying an item as either anchor (previously operational) or non-anchor (previously piloted) for the WASL. If an item was identified as an anchor item, validity responses from the previous contractor/previous administration of the assessment were carried forward and placed in the validity “pool” that would eventually be distributed among all the scorers for that given item. The “true” scores or scores the response had received previously were carried forward and were not changed. If the item was a non-anchor item, the Content Specialist and Scoring Directors were asked to select 70 responses from “live” responses (responses from the current administration) after rangefinding and OSPI would make final approval. Those 70 responses along with the 30 scored responses from the rangefinding committee which made up anchor and qualifying sets would make up the validity pool for all newly operational items.

The validity papers were then implemented to test scorer accuracy. The responses were set up in the imaging system as validity responses and dispersed intermittently to the scorers. The dispersal rates varied by subject, but generally there were more validity responses assigned at
the beginning of the project and fewer as the project progressed; By the end of the project, scorers in all subjects had scored all validity papers for any item were qualified to score. These validity responses were “blind” reads, meaning that scorers saw these responses the same as they saw the actual live student responses; there was no distinguishable difference. This helped ensure the ’internal validity’ of the process. It is important to note that all scorers who received validity papers had already successfully completed the training/qualifying process.

Next, the scores that the scorers assigned to the validity papers were compared to the true scores in order to determine the “validity” of the scorers’ scores. Each response was listed by the lithocode so that an overview of scorer performance for the particular response could be reviewed. For each item, the percentage of exact agreement as well as the percentage of high and low scores was computed. The same kind of data was also computed for each specific scorer. The results of this data could be run “live”, from several different reports, throughout the day to monitor for potential drift from the very early stages of the project.

If the results indicated that there was drift for a particular response, item or scorer, immediate action was taken to correct it. Corrective action could mean individual scorer re-training, room-wide re-training/recalibration, and/or re-scoring responses where it was determined a scorer had been errantly assigning scores. Sometimes, when a particular validity paper generated low agreement, an example of a similar response could be found in the already-existent training materials. If this was the case, a quick review of that particular training paper was usually enough to get scorers re-aligned.

Validity was employed on all operational items. Each validity set was formulated to mimic the score point distribution that the item generated during its previous administration. Examples of different types of responses were included to ensure that scorers were tested on the full spectrum of response types. Occasionally, validity responses were replaced, always with OSPI making the final decision about removing and replacing the response.

By carrying forward validity responses that had been used on previous administrations, by the previous vendor, DRC ensured that the scoring standards stayed the same.

DRC used the same training material for all Spring operational anchor items that had been used previously. The only exception to this was if materials were not available. If materials were not available, the item went through rangefinding using the same rubric that was used for scoring during the previous administration. Using the same materials helped ensure that scoring remained consistent with previous vendors and administrations.

The rangefinding process was the same as the previous vendors. DRC had several meetings about the entire rangefinding process so that the process could be as seamless as possible. Our endeavor was to make as few changes as necessary and still keep the integrity of the rangefinding and scoring. In addition to DRC and OSPI working together in rangefinding, OSPI was also on hand during the training process and the initial days of scoring. DRC implemented, in Spring 2009, higher qualifying standards for Reading, Mathematics, and Science than the previous vendor had (two-point items had a 90% qualification rate; four-point items had an 80% qualification rate). During the scoring windows, we provided quality assurance reports to OSPI that were the same type of customized reports they were used to seeing with the previous vendor. Examples of these reports include the above/below standard writing report,
attribute scoring points converted to holistic scores report, historical data writing report, and cumulative intrerrater and score point distribution by day writing report.

One additional step DRC took to make sure that consistency was maintained during the transition of the scoring vendors was reviewing anchor items’ score point distribution from previous administrations. If we saw that there was a change in score point distribution from the previous administration, we utilized all quality tools (quality reports, validity, read-behinds, and re-scores when necessary) to determine if the change was valid or if there were potential issues that needed to be addressed.

By implementing validity (making sure images looked virtually the same as with the previous vendor), using the same training materials whenever possible, following the same range-finding process, providing score point distribution for all items (pilot and operational), and replicating many of the same customized reports OSPI was accustomed to using, DRC made the first year of the scoring contract successful.

Tables 26 through 28 below show the exact, exact + adjacent agreement rates for the operational constructed responses of the Reading, Writing, and Mathematics items for Grade 4.

<table>
<thead>
<tr>
<th>Item</th>
<th>Points Possible</th>
<th>Number of Papers Scored</th>
<th>% Exact Agreement</th>
<th>% Adjacent + Exact Agreement</th>
<th>% Non-Adjacent Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>8,338</td>
<td>95</td>
<td>100</td>
<td>0</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>8,112</td>
<td>93</td>
<td>100</td>
<td>0</td>
<td>0.88</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>8,198</td>
<td>95</td>
<td>100</td>
<td>0</td>
<td>0.92</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>8,096</td>
<td>97</td>
<td>100</td>
<td>0</td>
<td>0.94</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>7,957</td>
<td>93</td>
<td>100</td>
<td>0</td>
<td>0.87</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>8,056</td>
<td>96</td>
<td>100</td>
<td>0</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 27. 2009 Grade 4 Writing – Interrater Percent Agreement

<table>
<thead>
<tr>
<th>Item</th>
<th>Points Possible</th>
<th>Number of Papers Scored</th>
<th>% Exact Agreement</th>
<th>% Adjacent + Exact Agreement</th>
<th>% Non-Adjacent Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expository Content (COS)</td>
<td>4 (1,2,3,4)</td>
<td>9,078</td>
<td>86</td>
<td>100</td>
<td>0</td>
<td>0.77</td>
</tr>
<tr>
<td>Expository Mechanics (CONV)</td>
<td>3 (0,1,2)</td>
<td>9,078</td>
<td>90</td>
<td>100</td>
<td>0</td>
<td>0.77</td>
</tr>
<tr>
<td>Persuasive Content (COS)</td>
<td>4 (1,2,3,4)</td>
<td>8,287</td>
<td>91</td>
<td>100</td>
<td>0</td>
<td>0.86</td>
</tr>
<tr>
<td>Persuasive Mechanics (CONV)</td>
<td>3 (0,1,2)</td>
<td>8,287</td>
<td>91</td>
<td>100</td>
<td>0</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 28. 2009 Grade 4 Mathematics – Interrater Percent Agreement

<table>
<thead>
<tr>
<th>Item</th>
<th>Points Possible</th>
<th>Number of Papers Scored</th>
<th>% Exact Agreement</th>
<th>% Adjacent + Exact Agreement</th>
<th>% Non-Adjacent Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>7,813</td>
<td>94</td>
<td>100</td>
<td>0</td>
<td>0.91</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>7,788</td>
<td>93</td>
<td>100</td>
<td>0</td>
<td>0.88</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>7,786</td>
<td>97</td>
<td>100</td>
<td>0</td>
<td>0.94</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>7,805</td>
<td>97</td>
<td>100</td>
<td>0</td>
<td>0.94</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>7,797</td>
<td>96</td>
<td>100</td>
<td>0</td>
<td>0.94</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>7,796</td>
<td>95</td>
<td>100</td>
<td>0</td>
<td>0.91</td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>7,778</td>
<td>96</td>
<td>100</td>
<td>0</td>
<td>0.93</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>7,888</td>
<td>96</td>
<td>100</td>
<td>0</td>
<td>0.92</td>
</tr>
</tbody>
</table>
PART 8: PERFORMANCE OF 2009 GRADE 4 STUDENTS

The summary data presented in Tables 29 to 48 are descriptive of Grade 4 student performance on the 2009 WASL. Included are raw score means and standard deviations for strand scores and the Writing test, scaled score means and standard deviations for other Grade 4 WASL tests, and numbers of Grade 4 students tested and disaggregated by a variety of groups. The results in this section of the report are derived from the statewide student data file received from Data Recognition Corporation on July 20, 2009. Students were included in these tables based on the follow criteria:

a) Students whose reporting grade is 4;
b) Student’s test type is either WABA or WASL;
c) Student’s attempt value is TS (Tested);
d) Student not missing test pages; and

e) For Mathematics only, student did not take the Braille or the Large Print version.

Means and standard deviations were calculated relative to the number of students tested, rather than number of students in the population. Table 29 provides the statewide mean scores for Grade 4 students who took the WASL tests in Spring 2009. The column “Minimum Scaled Score” lists the lowest observed scaled score and “Maximum Scaled Score” lists the highest observed scaled score for each of the 2009 tests. The next two columns contain the mean scaled score and scaled score standard deviations for students tested statewide. Table 30 lists the 2009 Grade 4 statewide summary statistics for content strands in each WASL test on a raw score metric.

Table 29. 2009 Grade 4 Means & Standard Deviations (SD) Test Scores

<table>
<thead>
<tr>
<th>Test</th>
<th>Number Tested</th>
<th>Minimum Scaled Score or Raw Score † or Raw Score *</th>
<th>Maximum Scaled Score or Raw Score † or Raw Score *</th>
<th>Mean Scaled Score † or Raw Score *</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>75,823</td>
<td>311</td>
<td>475</td>
<td>411.5</td>
<td>22.6</td>
</tr>
<tr>
<td>Writing</td>
<td>74,358</td>
<td>2</td>
<td>12</td>
<td>8.8</td>
<td>2.1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>76,083</td>
<td>200</td>
<td>550</td>
<td>400.7</td>
<td>42.4</td>
</tr>
</tbody>
</table>

†Scaled Scores computed and reported for Reading and Mathematics tests.
*The Writing test is reported on the raw score metric. No Scaled Scores are computed or reported for this test.

9 During the scanning of the test booklets, it was noted that twelve students were missing one or more pages from their test booklets. These students are excluded from this report.
10 Some operational items were excluded from the Mathematics Braille and Large Print forms, thus special raw score to scaled score relationship tables were created for these forms. All items were included in the Reading and Science Braille and Large Print forms, therefore no special scoring tables were produced for Reading or Science.
Table 30. 2009 Grade 4 Raw Test Score Summaries, Percent Students with Strength in Strand

<table>
<thead>
<tr>
<th>Strand</th>
<th>Number Tested</th>
<th>Points Possible</th>
<th>Raw Score Mean</th>
<th>SD</th>
<th>Percent with Strength in Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>75,823</td>
<td>36</td>
<td>25.7</td>
<td>6.5</td>
<td></td>
</tr>
<tr>
<td>ALY</td>
<td>75,823</td>
<td>18</td>
<td>13.0</td>
<td>3.3</td>
<td>71.3</td>
</tr>
<tr>
<td>CMP</td>
<td>75,823</td>
<td>18</td>
<td>12.7</td>
<td>3.6</td>
<td>67.7</td>
</tr>
<tr>
<td>Writing</td>
<td>74,358</td>
<td>12</td>
<td>8.8</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>COS</td>
<td>74,358</td>
<td>8</td>
<td>5.5</td>
<td>1.4</td>
<td>53.3</td>
</tr>
<tr>
<td>CONV</td>
<td>74,358</td>
<td>4</td>
<td>3.3</td>
<td>1.0</td>
<td>78.5</td>
</tr>
<tr>
<td>Mathematics</td>
<td>76,083</td>
<td>38</td>
<td>24.0</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>76,083</td>
<td>28</td>
<td>18.0</td>
<td>5.6</td>
<td>51.2</td>
</tr>
<tr>
<td>PC</td>
<td>76,083</td>
<td>10</td>
<td>6.0</td>
<td>2.3</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Tables 31 through 39 summarize the number of students tested, the mean scaled score, and scaled score standard deviation by various demographic and categorical programs for each WASL test.

Table 31. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>38,763</td>
<td>408.8</td>
<td>22.4</td>
</tr>
<tr>
<td>Females</td>
<td>36,821</td>
<td>414.4</td>
<td>22.3</td>
</tr>
</tbody>
</table>

Note. Unknown gender case are not included in the table.

Table 32. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska Native/Native American</td>
<td>2,038</td>
<td>402.4</td>
<td>21.3</td>
</tr>
<tr>
<td>Asian</td>
<td>5,941</td>
<td>417.2</td>
<td>22.5</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,329</td>
<td>402.5</td>
<td>21.1</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,489</td>
<td>400.9</td>
<td>20.6</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,798</td>
<td>414.9</td>
<td>22.0</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>533</td>
<td>404.8</td>
<td>21.8</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,104</td>
<td>413.1</td>
<td>22.2</td>
</tr>
</tbody>
</table>

Note. Unknown ethnicity cases are not included in the table.
Table 33. 2009 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAP Reading</td>
<td>4,489</td>
<td>400.2</td>
<td>19.3</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,545</td>
<td>400.5</td>
<td>20.2</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,323</td>
<td>404.6</td>
<td>21.5</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>12,944</td>
<td>405.6</td>
<td>21.7</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,412</td>
<td>436.5</td>
<td>18.0</td>
</tr>
<tr>
<td>Special Ed</td>
<td>9,862</td>
<td>390.1</td>
<td>22.3</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,555</td>
<td>397.3</td>
<td>19.5</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,511</td>
<td>389.6</td>
<td>16.8</td>
</tr>
</tbody>
</table>

Table 34. 2009 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>37,774</td>
<td>8.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Females</td>
<td>36,356</td>
<td>9.3</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Note. Unknown gender case are not included in the table.

Table 35. 2009 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska Native/Native American</td>
<td>1,949</td>
<td>7.8</td>
<td>2.3</td>
</tr>
<tr>
<td>Asian</td>
<td>5,830</td>
<td>9.5</td>
<td>1.8</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,181</td>
<td>8.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,120</td>
<td>8.0</td>
<td>2.2</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,114</td>
<td>9.0</td>
<td>2.1</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>518</td>
<td>8.6</td>
<td>2.0</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,075</td>
<td>8.9</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Note. Unknown ethnicity cases are not included in the table.
Table 36. 2009 Grade 4 Writing – Raw Score Means & Standard Deviations (SD) by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAP Reading</td>
<td>4,378</td>
<td>7.8</td>
<td>2.1</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,445</td>
<td>7.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>15,917</td>
<td>8.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>12,623</td>
<td>8.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,402</td>
<td>10.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Special Ed</td>
<td>9,302</td>
<td>6.7</td>
<td>2.4</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,493</td>
<td>7.8</td>
<td>2.1</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,180</td>
<td>7.1</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Table 37. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>38,927</td>
<td>400.1</td>
<td>43.2</td>
</tr>
<tr>
<td>Females</td>
<td>36,931</td>
<td>401.5</td>
<td>41.5</td>
</tr>
</tbody>
</table>

Note. Unknown gender case are not included in the table.

Table 38. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska Native/Native American</td>
<td>2,043</td>
<td>381.7</td>
<td>39.6</td>
</tr>
<tr>
<td>Asian</td>
<td>6,039</td>
<td>415.2</td>
<td>44.9</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,354</td>
<td>378.6</td>
<td>39.1</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,584</td>
<td>378.7</td>
<td>38.2</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,832</td>
<td>407.7</td>
<td>40.4</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>538</td>
<td>384.8</td>
<td>40.9</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,110</td>
<td>401.9</td>
<td>41.1</td>
</tr>
</tbody>
</table>

Note. Unknown ethnicity cases are not included in the table.
Table 39. 2009 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number Tested</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAP Reading</td>
<td>4,514</td>
<td>381.3</td>
<td>39.0</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,560</td>
<td>379.8</td>
<td>40.7</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,393</td>
<td>388.7</td>
<td>41.1</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>13,006</td>
<td>390.7</td>
<td>41.6</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,410</td>
<td>454.5</td>
<td>36.3</td>
</tr>
<tr>
<td>Special Ed</td>
<td>9,886</td>
<td>365.1</td>
<td>41.9</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,584</td>
<td>372.3</td>
<td>37.8</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,781</td>
<td>361.9</td>
<td>34.3</td>
</tr>
</tbody>
</table>

PERCENT MEETING STANDARD

Tables 40 through 48 list the percent of students in each gender, ethnic, and categorical program group who did or did not meet standard for each content area.

Following are general descriptions of the performance level standards for the WASL.

Level 4 “Advanced”: This level represents superior performance, notably above that required for meeting the standard at Grade 4.

Level 3 “Proficient”: This level represents solid academic performance for Grade 4. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 “Basic”: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at Grade 4.

Level 1 “Below Basic”: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at Grade 4.

* In all content areas, “Proficient” reflects what a well taught, hard working student should know and be able to do.

For all WASL tests, “Meets Standard” is defined by Level 3 “Proficient” and Level 4 “Advanced.” Level 1 “Below Basic” and Level 2 “Basic” do not meet standard.

As noted in Tables 40 to 48, the percentage entries are based on the number of students within a particular subgroup or program category. All students with a reporting Grade of 4 in the July file provided by DRC are included in these tables. Students are then classified into one of five categories.

Performance Level 1 “Below Basic” in these tables includes students who;

a) Attempted the WASL but received no score for an unexcused absence;
b) Are missing booklet;
c) Have an incomplete record;
d) Refusal to test;
e) Have an invalidated test; or
f) Tested with an out of grade level test.

“Not tested” consists of students excluded from testing on the basis of
a) Limited English proficiency (LEP);
b) Medical condition;
c) Excused absence;
d) Partial enrollment during the testing window; or
e) Exemption due to participating in the alternate assessment portfolio (WAAS) or in the Developmentally Appropriate WASL (DAW).

“Percent Exempt” is a subset of “Percent Not Tested,” and reflects the percent of total grade level enrollment that participated in the WAAS or DAW programs. Within each row of the following tables, “Meets Standard,” “Does Not Meet Standard,” and “Percent Not Tested” percentages sum to 100%.

Table 40. 2009 Grade 4 Reading – Percent Meeting Standards by Gender

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent Level 3</td>
<td>Percent Level 2</td>
<td>Percent Level 1</td>
</tr>
<tr>
<td>All Students</td>
<td>77,530</td>
<td>33.3</td>
<td>37.7</td>
<td>22.2</td>
<td>5.7</td>
</tr>
<tr>
<td>Females</td>
<td>37,504</td>
<td>37.9</td>
<td>37.7</td>
<td>19.4</td>
<td>4.0</td>
</tr>
<tr>
<td>Males</td>
<td>39,620</td>
<td>29.2</td>
<td>37.9</td>
<td>25.0</td>
<td>6.9</td>
</tr>
</tbody>
</table>
### Table 41. 2009 Grade 4 Reading – Percent Meeting Standards by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska Native/Native American</td>
<td>2,097</td>
<td>19.2 37.7</td>
<td>31.6 10.2</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Asian</td>
<td>6,098</td>
<td>42.4 35.9</td>
<td>16.6 3.1</td>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,432</td>
<td>18.6 37.7</td>
<td>33.6 8.7</td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,773</td>
<td>16.0 36.9</td>
<td>36.2 9.2</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>48,637</td>
<td>38.9 38.3</td>
<td>17.9 4.3</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>556</td>
<td>21.9 35.8</td>
<td>30.6 8.5</td>
<td>3.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,158</td>
<td>34.9 38.7</td>
<td>20.4 4.9</td>
<td>1.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### Table 42. 2009 Grade 4 Reading – Percent Meeting Standards by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAP Reading</td>
<td>4,549</td>
<td>13.8 37.6</td>
<td>40.4 7.3</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td>LAP Math</td>
<td>3,595</td>
<td>15.0 36.7</td>
<td>39.3 8.2</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,603</td>
<td>21.9 38.0</td>
<td>31.8 7.3</td>
<td>1.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Title I Math</td>
<td>13,174</td>
<td>23.6 38.4</td>
<td>29.9 7.1</td>
<td>1.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,422</td>
<td>83.0 15.7</td>
<td>1.0 0.2</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Special Ed</td>
<td>10,189</td>
<td>9.5 22.1</td>
<td>42.4 25.2</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,610</td>
<td>11.2 33.7</td>
<td>41.5 10.8</td>
<td>2.9</td>
<td>0.0</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,907</td>
<td>3.1 24.7</td>
<td>51.3 16.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Table 43. 2009 Grade 4 Writing – Percent Meeting Standards by Gender

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Level 4</td>
<td>Level 3</td>
<td>Level 2</td>
<td>Level 1</td>
<td>Not Tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Students</td>
<td>76,467</td>
<td>21.8</td>
<td>37.8</td>
<td>23.2</td>
<td>16.0</td>
<td>1.2</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>37,148</td>
<td>29.5</td>
<td>40.2</td>
<td>19.4</td>
<td>9.8</td>
<td>1.2</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>38,925</td>
<td>14.6</td>
<td>35.8</td>
<td>26.9</td>
<td>21.5</td>
<td>1.2</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 44. 2009 Grade 4 Writing – Percent Meeting Standards by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Level 4</td>
<td>Level 3</td>
<td>Level 2</td>
<td>Level 1</td>
<td>Not Tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alaska Native/Native American</td>
<td>2,042</td>
<td>9.5</td>
<td>31.9</td>
<td>28.5</td>
<td>28.2</td>
<td>2.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>6,014</td>
<td>31.1</td>
<td>42.1</td>
<td>16.9</td>
<td>7.7</td>
<td>2.2</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,318</td>
<td>12.6</td>
<td>36.0</td>
<td>28.0</td>
<td>21.8</td>
<td>1.6</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,480</td>
<td>11.2</td>
<td>33.5</td>
<td>29.9</td>
<td>23.4</td>
<td>1.9</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>48,167</td>
<td>24.9</td>
<td>38.9</td>
<td>21.6</td>
<td>13.9</td>
<td>0.8</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>545</td>
<td>15.0</td>
<td>41.5</td>
<td>25.1</td>
<td>14.5</td>
<td>3.9</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,138</td>
<td>22.7</td>
<td>38.9</td>
<td>23.4</td>
<td>13.7</td>
<td>1.2</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>
Table 45. 2009 Grade 4 Writing – Percent Meeting Standards by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent Level 3</td>
<td>Percent Level 2</td>
<td>Percent Level 1</td>
</tr>
<tr>
<td>LAP Reading</td>
<td>4,463</td>
<td>7.8</td>
<td>31.4</td>
<td>32.0</td>
<td>27.8</td>
</tr>
<tr>
<td>LAP Math</td>
<td>3,512</td>
<td>9.7</td>
<td>30.1</td>
<td>30.5</td>
<td>28.7</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,304</td>
<td>13.8</td>
<td>35.0</td>
<td>28.4</td>
<td>21.6</td>
</tr>
<tr>
<td>Title I Math</td>
<td>12,935</td>
<td>14.4</td>
<td>35.1</td>
<td>27.9</td>
<td>21.3</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,416</td>
<td>55.1</td>
<td>36.8</td>
<td>6.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Special Ed</td>
<td>9,728</td>
<td>5.4</td>
<td>18.2</td>
<td>27.0</td>
<td>48.4</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,558</td>
<td>8.7</td>
<td>31.5</td>
<td>30.6</td>
<td>26.3</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,624</td>
<td>3.9</td>
<td>22.2</td>
<td>32.9</td>
<td>35.6</td>
</tr>
</tbody>
</table>

Table 46. 2009 Grade 4 Mathematics – Percent Meeting Standards by Gender

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent Level 3</td>
<td>Percent Level 2</td>
<td>Percent Level 1</td>
</tr>
<tr>
<td>All Students</td>
<td>77,489</td>
<td>31.5</td>
<td>19.8</td>
<td>23.7</td>
<td>24.2</td>
</tr>
<tr>
<td>Females</td>
<td>37,489</td>
<td>31.7</td>
<td>20.5</td>
<td>24.4</td>
<td>22.8</td>
</tr>
<tr>
<td>Males</td>
<td>39,596</td>
<td>31.6</td>
<td>19.3</td>
<td>23.2</td>
<td>25.2</td>
</tr>
</tbody>
</table>
### Table 47. 2009 Grade 4 Mathematics – Percent Meeting Standards by Ethnic Group

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent Level 3</td>
<td>Percent Level 2</td>
<td>Percent Level 1</td>
</tr>
<tr>
<td>Alaska Native/Native American</td>
<td>2,096</td>
<td>16.1</td>
<td>16.7</td>
<td>25.7</td>
<td>40.1</td>
</tr>
<tr>
<td>Asian</td>
<td>6,095</td>
<td>44.5</td>
<td>20.0</td>
<td>19.6</td>
<td>15.5</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,432</td>
<td>14.1</td>
<td>15.1</td>
<td>27.3</td>
<td>42.4</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,769</td>
<td>13.5</td>
<td>15.1</td>
<td>28.1</td>
<td>42.2</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>48,606</td>
<td>37.2</td>
<td>21.7</td>
<td>22.6</td>
<td>17.9</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>556</td>
<td>18.0</td>
<td>18.5</td>
<td>25.7</td>
<td>35.4</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,158</td>
<td>31.8</td>
<td>20.0</td>
<td>25.0</td>
<td>22.2</td>
</tr>
</tbody>
</table>

### Table 48. 2009 Grade 4 Mathematics – Percent Meeting Standards by Categorical Program

<table>
<thead>
<tr>
<th>Categorical Program</th>
<th>Number of Students</th>
<th>Meets Standard</th>
<th>Does Not Meet Standard</th>
<th>Percent Not Tested</th>
<th>Percent Exempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percent Level 4</td>
<td>Percent Level 3</td>
<td>Percent Level 2</td>
<td>Percent Level 1</td>
</tr>
<tr>
<td>LAP Reading</td>
<td>4,549</td>
<td>15.3</td>
<td>13.9</td>
<td>29.1</td>
<td>41.3</td>
</tr>
<tr>
<td>LAP Math</td>
<td>3,596</td>
<td>15.5</td>
<td>12.6</td>
<td>26.9</td>
<td>44.3</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,599</td>
<td>21.6</td>
<td>17.1</td>
<td>26.5</td>
<td>34.1</td>
</tr>
<tr>
<td>Title I Math</td>
<td>13,171</td>
<td>23.5</td>
<td>17.5</td>
<td>25.8</td>
<td>32.5</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,420</td>
<td>84.4</td>
<td>10.7</td>
<td>3.9</td>
<td>0.8</td>
</tr>
<tr>
<td>Special Ed</td>
<td>10,159</td>
<td>10.5</td>
<td>9.8</td>
<td>19.9</td>
<td>58.9</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,609</td>
<td>10.1</td>
<td>12.2</td>
<td>26.7</td>
<td>50.0</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,903</td>
<td>4.8</td>
<td>8.5</td>
<td>24.1</td>
<td>61.5</td>
</tr>
</tbody>
</table>

Table 49 and Figure 9 illustrate the trend in student performance from 1996-97 to 2008-09 in each content area. These data are based on information from published statewide score reports.
Table 49. Grade 4 Percentage of Students Meeting Standard from 1996-97 through 2008-09

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>47.9%</td>
<td>55.6%</td>
<td>59.1%</td>
<td>65.8%</td>
<td>66.1%</td>
<td>65.6%</td>
<td>66.7%</td>
<td>74.4%</td>
<td>79.2%</td>
<td>81.2%</td>
<td>76.2%</td>
<td>72.3%</td>
<td>73.60%</td>
</tr>
<tr>
<td>Writing</td>
<td>42.8%</td>
<td>36.7%</td>
<td>32.6%</td>
<td>39.4%</td>
<td>43.3%</td>
<td>49.5%</td>
<td>53.6%</td>
<td>55.8%</td>
<td>57.5%</td>
<td>60.6%</td>
<td>60.1%</td>
<td>62.1%</td>
<td>60.40%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>21.4%</td>
<td>31.2%</td>
<td>37.3%</td>
<td>41.8%</td>
<td>43.4%</td>
<td>51.8%</td>
<td>55.2%</td>
<td>59.9%</td>
<td>60.6%</td>
<td>59.0%</td>
<td>57.9%</td>
<td>53.3%</td>
<td>52.30%</td>
</tr>
</tbody>
</table>

Note. The numbers provided in this Table for 1996-97 through 2007-08 were provided in the previous 2008 Technical Report. The numbers for 2008-09 are from the OSPI website.

Figure 9. Grade 4 Results for 1996-97 through 2008-09 by Content Area
MEAN ITEM PERFORMANCE AND ITEM-TEST CORRELATIONS

Tables 50 through 52 provide summary statistics for the 2009 Grade 4 WASL operational items and are based on the equating files provided by DRC\textsuperscript{11}. Students were included in the equating analyses, and therefore the production of the raw score to scaled score tables, if they met the following criteria:

a) Student’s reporting grade is 4;
   b) Student’s test type was either WABA or WASL;
   c) Student’s test attempt value was TS (Tested);
   d) Student responded to two or more items per test section; and
   e) For Mathematics only, student did not take the Braille or Large Print version.\textsuperscript{12}

Table 50 contains the item difficulties and point-biserial correlations for the Reading and Mathematics tests. Table 51 lists those items with an omit rate greater than 5%. Table 52 contains summary statistics for the Rasch item difficulties by item type. The analyses performed on the items is described in Part 2. The data listed in Tables 53 through 55 indicate the number of points possible for each operational item, the item means, the item-test score correlations, and the Rasch item difficulties for each of the items in the Reading, Writing, and Mathematics tests.

\textsuperscript{11} The equating files for Reading and Mathematics were received on June 16, 2009.
\textsuperscript{12} Some operational items were excluded from the Mathematics Braille and Large Print forms, thus special raw score to scaled score relationship tables were created for these forms. All items were included in the Reading and Science Braille and Large Print forms, therefore no special scoring tables were produced for Reading or Science.
Table 50.  2009 Grade 4 Summary Statistics – Operational Items

<table>
<thead>
<tr>
<th>Item Statistics</th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item Difficulties: P-Values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 0.90</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.80 - 0.89</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>0.70 - 0.79</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>0.60 - 0.69</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>0.50 - 0.59</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>0.25 - 0.49</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>&lt; 0.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.74</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Std. Dev.</strong></td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Item Discrimination: Point-Biserial Correlations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 0.60</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.50 – 0.59</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>0.40 - 0.49</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>0.30 - 0.39</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>0.20 - 0.29</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>&lt;0.20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.42</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Std. Dev.</strong></td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Total Number of Items</strong></td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 51.  2009 Grade 4 Omit Rate – Operational Item Statistics

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Percent Omitting</th>
<th>Item Number</th>
<th>Percent Omitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>-</td>
<td>N/A</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note.* – indicates no item has more than 5% missing responses.
Table 52. 2009 Grade 4 Rasch Item Difficulty Summary Statistics – Operational Items

<table>
<thead>
<tr>
<th>Response Type</th>
<th>Item Statistics</th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>Mean 0.03</td>
<td>-0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum -1.90</td>
<td>-2.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum 1.29</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 0.69</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N 24</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>Mean 0.75</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum 0.02</td>
<td>-0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum 1.41</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 0.64</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N 6</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Table 53. 2009 Grade 4 Writing – Operational Item Statistics

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Score Type</th>
<th>Score Points Possible</th>
<th>Score Mean</th>
<th>Score-Total Test Correlation¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Narrative Content, Organization &amp; Style</td>
<td>4</td>
<td>2.876</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td>Narrative Writing Conventions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Expository Content, Organization &amp; Style</td>
<td>4</td>
<td>2.641</td>
<td>0.564</td>
</tr>
<tr>
<td></td>
<td>Expository Writing Conventions</td>
<td></td>
<td>1.660</td>
<td>0.580</td>
</tr>
</tbody>
</table>

Note ¹. Score itself is not included in the total when score-total test correlation is calculated. This follows the 2008 technical report procedure.
<table>
<thead>
<tr>
<th>Item Number in Test Booklet</th>
<th>Points Possible</th>
<th>Item Mean</th>
<th>Item-Test Correlation</th>
<th>Rasch Item Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.75</td>
<td>0.41</td>
<td>0.179</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.52</td>
<td>0.58</td>
<td>0.020</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.87</td>
<td>0.34</td>
<td>-0.725</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.78</td>
<td>0.37</td>
<td>-0.012</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.95</td>
<td>0.43</td>
<td>-1.897</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.80</td>
<td>0.42</td>
<td>-0.143</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.82</td>
<td>0.40</td>
<td>-0.379</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1.17</td>
<td>0.59</td>
<td>1.365</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.54</td>
<td>0.42</td>
<td>1.285</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1.19</td>
<td>0.61</td>
<td>1.168</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.77</td>
<td>0.56</td>
<td>0.103</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.72</td>
<td>0.36</td>
<td>0.337</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.84</td>
<td>0.43</td>
<td>-0.517</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>1.06</td>
<td>0.59</td>
<td>1.413</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.71</td>
<td>0.29</td>
<td>0.469</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0.87</td>
<td>0.50</td>
<td>-0.721</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0.80</td>
<td>0.44</td>
<td>-0.294</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0.75</td>
<td>0.40</td>
<td>0.233</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0.68</td>
<td>0.44</td>
<td>0.614</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.66</td>
<td>0.29</td>
<td>0.749</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0.76</td>
<td>0.37</td>
<td>0.169</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.67</td>
<td>0.37</td>
<td>0.681</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>1.38</td>
<td>0.53</td>
<td>0.417</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0.61</td>
<td>0.42</td>
<td>1.001</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0.69</td>
<td>0.44</td>
<td>0.543</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>0.74</td>
<td>0.39</td>
<td>0.066</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>1.56</td>
<td>0.64</td>
<td>0.125</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>0.82</td>
<td>0.38</td>
<td>-0.070</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>0.79</td>
<td>0.37</td>
<td>-0.087</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>0.89</td>
<td>0.47</td>
<td>-0.912</td>
</tr>
<tr>
<td>Item Number in Test Booklet</td>
<td>Points Possible</td>
<td>Item Mean</td>
<td>Item-Test Correlation</td>
<td>Rasch Item Difficulty</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.79</td>
<td>0.44</td>
<td>-1.058</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.46</td>
<td>0.33</td>
<td>0.836</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.57</td>
<td>0.47</td>
<td>0.186</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.58</td>
<td>0.19</td>
<td>0.247</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.78</td>
<td>0.48</td>
<td>-0.901</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.78</td>
<td>0.38</td>
<td>-0.993</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.59</td>
<td>0.45</td>
<td>0.181</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1.11</td>
<td>0.58</td>
<td>0.399</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.61</td>
<td>0.51</td>
<td>0.199</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1.15</td>
<td>0.50</td>
<td>-0.019</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.82</td>
<td>0.51</td>
<td>-1.150</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1.01</td>
<td>0.46</td>
<td>0.610</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>1.16</td>
<td>0.61</td>
<td>0.398</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.70</td>
<td>0.56</td>
<td>-0.383</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.68</td>
<td>0.29</td>
<td>-0.261</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0.78</td>
<td>0.42</td>
<td>-0.740</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.93</td>
<td>0.41</td>
<td>-2.334</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.92</td>
<td>0.53</td>
<td>0.691</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.51</td>
<td>0.45</td>
<td>0.552</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0.59</td>
<td>0.56</td>
<td>0.133</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0.72</td>
<td>0.44</td>
<td>-0.527</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>1.42</td>
<td>0.64</td>
<td>-0.130</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>0.72</td>
<td>0.51</td>
<td>-0.533</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0.90</td>
<td>0.30</td>
<td>-1.927</td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>0.90</td>
<td>0.58</td>
<td>1.086</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>0.79</td>
<td>0.45</td>
<td>-0.949</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>0.50</td>
<td>0.45</td>
<td>0.634</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>0.76</td>
<td>0.34</td>
<td>-0.731</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>1.59</td>
<td>0.45</td>
<td>-0.812</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>0.70</td>
<td>0.48</td>
<td>-0.371</td>
</tr>
</tbody>
</table>
APPENDIX A: RELIABILITY BY SUBGROUP
<table>
<thead>
<tr>
<th>Subgroup</th>
<th>N</th>
<th>Maximum Possible Raw Score</th>
<th>Alpha Coefficient</th>
<th>Raw Score Standard Error of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Students</td>
<td>75,823</td>
<td>36</td>
<td>0.86</td>
<td>2.48</td>
</tr>
<tr>
<td>Females</td>
<td>36,821</td>
<td>36</td>
<td>0.85</td>
<td>2.44</td>
</tr>
<tr>
<td>Males</td>
<td>38,763</td>
<td>36</td>
<td>0.86</td>
<td>2.51</td>
</tr>
<tr>
<td>Alaska Native/Native American</td>
<td>2,038</td>
<td>36</td>
<td>0.85</td>
<td>2.63</td>
</tr>
<tr>
<td>Asian</td>
<td>5,941</td>
<td>36</td>
<td>0.84</td>
<td>2.38</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,329</td>
<td>36</td>
<td>0.85</td>
<td>2.63</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,489</td>
<td>36</td>
<td>0.84</td>
<td>2.66</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,798</td>
<td>36</td>
<td>0.84</td>
<td>2.41</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>533</td>
<td>36</td>
<td>0.85</td>
<td>2.60</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,104</td>
<td>36</td>
<td>0.85</td>
<td>2.46</td>
</tr>
<tr>
<td>LAP Reading</td>
<td>4,489</td>
<td>36</td>
<td>0.82</td>
<td>2.67</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,545</td>
<td>36</td>
<td>0.83</td>
<td>2.67</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,323</td>
<td>36</td>
<td>0.85</td>
<td>2.60</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>12,944</td>
<td>36</td>
<td>0.85</td>
<td>2.58</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,412</td>
<td>36</td>
<td>0.61</td>
<td>1.85</td>
</tr>
<tr>
<td>Special Education</td>
<td>9,862</td>
<td>36</td>
<td>0.87</td>
<td>2.72</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,555</td>
<td>36</td>
<td>0.83</td>
<td>2.72</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,511</td>
<td>36</td>
<td>0.78</td>
<td>2.80</td>
</tr>
<tr>
<td>Subgroup</td>
<td>N</td>
<td>Maximum Possible Raw Score</td>
<td>Alpha Coefficient</td>
<td>Raw Score Standard Error of Measurement</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>All Students</td>
<td>74,358</td>
<td>12</td>
<td>0.77</td>
<td>1.02</td>
</tr>
<tr>
<td>Females</td>
<td>36,356</td>
<td>12</td>
<td>0.73</td>
<td>1.00</td>
</tr>
<tr>
<td>Males</td>
<td>37,774</td>
<td>12</td>
<td>0.77</td>
<td>1.04</td>
</tr>
<tr>
<td>Alaska Native/Native American</td>
<td>1,949</td>
<td>12</td>
<td>0.78</td>
<td>1.07</td>
</tr>
<tr>
<td>Asian</td>
<td>5,830</td>
<td>12</td>
<td>0.72</td>
<td>0.98</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,181</td>
<td>12</td>
<td>0.76</td>
<td>1.05</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,120</td>
<td>12</td>
<td>0.77</td>
<td>1.04</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,114</td>
<td>12</td>
<td>0.76</td>
<td>1.02</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>518</td>
<td>12</td>
<td>0.77</td>
<td>0.97</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,075</td>
<td>12</td>
<td>0.74</td>
<td>1.02</td>
</tr>
<tr>
<td>LAP Reading</td>
<td>4,378</td>
<td>12</td>
<td>0.75</td>
<td>1.05</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,445</td>
<td>12</td>
<td>0.76</td>
<td>1.05</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>15,917</td>
<td>12</td>
<td>0.76</td>
<td>1.04</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>12,623</td>
<td>12</td>
<td>0.77</td>
<td>1.04</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,402</td>
<td>12</td>
<td>0.53</td>
<td>0.90</td>
</tr>
<tr>
<td>Special Education</td>
<td>9,302</td>
<td>12</td>
<td>0.80</td>
<td>1.09</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,493</td>
<td>12</td>
<td>0.76</td>
<td>1.05</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,180</td>
<td>12</td>
<td>0.75</td>
<td>1.06</td>
</tr>
</tbody>
</table>
Table 58. 2009 Grade 4 Mathematics Test Reliability Estimates

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>N</th>
<th>Maximum Possible Raw Score</th>
<th>Alpha Coefficient</th>
<th>Raw Score Standard Error of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Students</td>
<td>76,083</td>
<td>38</td>
<td>0.87</td>
<td>2.69</td>
</tr>
<tr>
<td>Females</td>
<td>36,931</td>
<td>38</td>
<td>0.86</td>
<td>2.68</td>
</tr>
<tr>
<td>Males</td>
<td>38,927</td>
<td>38</td>
<td>0.87</td>
<td>2.70</td>
</tr>
<tr>
<td>Alaska Native/Native American</td>
<td>2,043</td>
<td>38</td>
<td>0.86</td>
<td>2.79</td>
</tr>
<tr>
<td>Asian</td>
<td>6,039</td>
<td>38</td>
<td>0.87</td>
<td>2.56</td>
</tr>
<tr>
<td>African American/Black</td>
<td>4,354</td>
<td>38</td>
<td>0.86</td>
<td>2.79</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>12,584</td>
<td>38</td>
<td>0.85</td>
<td>2.78</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>47,832</td>
<td>38</td>
<td>0.85</td>
<td>2.65</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>538</td>
<td>38</td>
<td>0.87</td>
<td>2.76</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2,110</td>
<td>38</td>
<td>0.86</td>
<td>2.68</td>
</tr>
<tr>
<td>LAP Reading</td>
<td>4,514</td>
<td>38</td>
<td>0.85</td>
<td>2.80</td>
</tr>
<tr>
<td>LAP Mathematics</td>
<td>3,560</td>
<td>38</td>
<td>0.86</td>
<td>2.79</td>
</tr>
<tr>
<td>Title I Reading</td>
<td>16,393</td>
<td>38</td>
<td>0.87</td>
<td>2.76</td>
</tr>
<tr>
<td>Title I Mathematics</td>
<td>13,006</td>
<td>38</td>
<td>0.87</td>
<td>2.75</td>
</tr>
<tr>
<td>Gifted</td>
<td>3,410</td>
<td>38</td>
<td>0.71</td>
<td>2.10</td>
</tr>
<tr>
<td>Special Education</td>
<td>9,886</td>
<td>38</td>
<td>0.87</td>
<td>2.79</td>
</tr>
<tr>
<td>Migrant</td>
<td>1,584</td>
<td>38</td>
<td>0.85</td>
<td>2.77</td>
</tr>
<tr>
<td>ELL/Bilingual</td>
<td>6,781</td>
<td>38</td>
<td>0.82</td>
<td>2.80</td>
</tr>
</tbody>
</table>
APPENDIX B: SUMMARY STATISTICS FOR EMBEDDED PILOT ITEMS
The 2009 Grade 4 Reading and Mathematics WASL tests included embedded pilot items. These items are administered in order to obtain item statistics (p-values, point-biserials and Rasch item difficulty values). These items and their associated statistics are reviewed by committee members to determine whether an item should be included in the item bank and possible use on future operational tests.

The summary statistics for the 2009 Grade 4 WASL embedded pilot items are based on the files provided by DRC\textsuperscript{13}. Students were included in the equating analyses, and therefore the production of the raw score to scaled score tables, if they met the following criteria:

a) Student’s reporting grade is 4;
b) Student’s test type was either WABA or WASL;
c) Student’s test attempt value was TS (Tested);
d) Student responded to two or more items per test section; and
e) For Mathematics only, student did not take the Braille or Large Print version.

The Traditional item analyses, IRT analyses, DIF analyses are described in Part 2. Table 59 provides summary statistics for the p-values and item-test correlations or point-biserials, Table 60 lists the items with an omit rate greater than 5%, and Table 61 provides summary statistics of the Rasch item difficulties. Tables 62 and 63 contain the results from the DIF analyses. Table 64 provides a summary of the items that were flagged during the analyses for the following:

a) Low average item score (or p-value);
b) High average item score;
c) Low item-test correlation (Pt Biserial);
d) Distractor having a positive item-test correlation;
e) Omit rate greater than 5%;
f) C-level DIF;
g) Poor fit to the Rasch IRT model; and
h) No student response to a distractor or score level.

\textsuperscript{13} The file for Reading and Mathematics the file was received on August 12, 2009 and the file for Science was received on September 23, 2009.
Table 59. 2009 Grade 4 Summary Statistics for Embedded Pilot Items

<table>
<thead>
<tr>
<th>Item Statistics</th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item Difficulties: P-Values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 0.90</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.80 - 0.89</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0.70 - 0.79</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>0.60 - 0.69</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>0.50 - 0.59</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>0.25 - 0.49</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>&lt; 0.25</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.68</td>
<td>0.50</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.65</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Std. Dev.</strong></td>
<td>0.09</td>
<td>0.19</td>
</tr>
</tbody>
</table>

| **Item Discrimination: Point-Biserial Correlations** |         |             |
|                                                      |         |             |
| ≥ 0.60                                               | 0       | 0           |
| 0.50 – 0.59                                         | 2       | 7           |
| 0.40 - 0.49                                         | 4       | 15          |
| 0.30 - 0.39                                         | 3       | 18          |
| 0.20 - 0.29                                         | 2       | 13          |
| <0.20                                               | 0       | 3           |
| **Mean**                                           | 0.40    | 0.36        |
| **Median**                                         | 0.42    | 0.36        |
| **Std. Dev.**                                      | 0.10    | 0.12        |
| **Total Number of Items**                           | 11      | 56          |

Table 60. 2009 Grade 4 Omit Rate Statistics for Embedded Pilot Items

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Percent Omitting</th>
<th>Item Number</th>
<th>Percent Omitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>-</td>
<td>N/A</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note.* – indicates no item has more than 5% missing responses.
Table 61. 2009 Grade 4 Rasch Item Difficulty Summary
Statistics for Embedded Pilot Items

<table>
<thead>
<tr>
<th>Response Type</th>
<th>Item Statistics</th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>Mean</td>
<td>0.45</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>-0.55</td>
<td>-2.49</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1.44</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.63</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>SA</td>
<td>Mean</td>
<td>1.04</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0.96</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1.15</td>
<td>2.54</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.10</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>CP</td>
<td>Mean</td>
<td></td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td></td>
<td>-1.74</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td></td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td></td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>

Note 1: Beginning in Spring 2010 a new item type is added a 1-point completion items, requiring a numerical or one/two word response, were piloted for the first time in Mathematics.
Table 62. 2009 Grade 4 Reading Differential Item Functioning Summary Statistics for Embedded Pilot Items

<table>
<thead>
<tr>
<th>Item Type</th>
<th>DIF Category</th>
<th>Males vs Females</th>
<th>White vs Asian</th>
<th>White vs African American</th>
<th>White vs Hispanic</th>
<th>White vs Native American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>MC</td>
<td>C-</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B-</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>8</td>
<td>100.00</td>
<td>8</td>
<td>100.00</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>B+</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C+</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SMALL N</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>8</td>
<td>100.00</td>
<td>8</td>
<td>100.00</td>
<td>8</td>
</tr>
<tr>
<td>SA</td>
<td>C-</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B-</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>3</td>
<td>100.00</td>
<td>3</td>
<td>100.00</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B+</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C+</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SMALL N</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>3</td>
<td>100.00</td>
<td>3</td>
<td>100.00</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: 1 Sample size insufficient to conduct DIF analysis.
<table>
<thead>
<tr>
<th>Item Type</th>
<th>DIF Category</th>
<th>Males vs Females</th>
<th>White vs Asian</th>
<th>White vs African American</th>
<th>White vs Hispanic</th>
<th>White vs Native American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>1</td>
<td>2.50</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>B-</td>
<td>3</td>
<td>7.50</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>2.50</td>
</tr>
<tr>
<td>A</td>
<td>36</td>
<td>90.00</td>
<td>38</td>
<td>95.00</td>
<td>39</td>
<td>97.50</td>
</tr>
<tr>
<td>B+</td>
<td>2</td>
<td>5.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>C+</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SMALL N</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>40</td>
<td>100.00</td>
<td>40</td>
<td>100.00</td>
<td>40</td>
<td>100.00</td>
</tr>
<tr>
<td>SA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>11.11</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>B-</td>
<td>1</td>
<td>11.11</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>77.78</td>
<td>7</td>
<td>77.78</td>
<td>8</td>
<td>88.88</td>
</tr>
<tr>
<td>B+</td>
<td>1</td>
<td>11.11</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>C+</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>11.11</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SMALL N</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>9</td>
<td>100.00</td>
<td>9</td>
<td>100.00</td>
<td>9</td>
<td>100.00</td>
</tr>
<tr>
<td>CP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>B-</td>
<td>2</td>
<td>28.57</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>71.43</td>
<td>5</td>
<td>71.43</td>
<td>7</td>
<td>100.00</td>
</tr>
<tr>
<td>B+</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>14.29</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>C+</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>14.29</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SMALL N</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>7</td>
<td>100.00</td>
<td>7</td>
<td>100.01</td>
<td>7</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Note: 1 Sample size insufficient to conduct DIF analysis.
Table 64. 2009 Grade 4 Summary of Items Flagged for Embedded Pilot Items

<table>
<thead>
<tr>
<th></th>
<th>P-Value</th>
<th>P-Value</th>
<th>Point Biserial</th>
<th>Distractor Point-Biserial</th>
<th>Omit Rate</th>
<th>C-Level DIF</th>
<th>Poor IRT Fit</th>
<th>Missing Responsea</th>
<th>Total Flags</th>
<th>N Items Flagged</th>
<th>N Itemsb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>&lt; 0.30</td>
<td>&gt; 0.90</td>
<td>&lt; 0.25</td>
<td>&gt; 0</td>
<td>&gt; 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td></td>
<td>27</td>
<td>22</td>
<td>56</td>
</tr>
</tbody>
</table>

aMC response option or SA option score point with no responses
bTotal number of embedded piloted items.
APPENDIX C: EXAMPLE OF DATA REVIEW REPORTS
## Grade 4 Reading
### 2009 Data Review

### Item Information

<table>
<thead>
<tr>
<th>Grade</th>
<th>Item Code</th>
<th>Form</th>
<th>Position</th>
<th>Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>34407</td>
<td>B/C</td>
<td>31</td>
<td>MC</td>
</tr>
</tbody>
</table>

**Key**
- **B**

**Strand/Target**
- LC04

**Passage Title**
- The Covered Bridge House

### Item Information Summary

<table>
<thead>
<tr>
<th>Rasch</th>
<th>% Choosing</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.015</td>
<td>A</td>
<td>12.333</td>
</tr>
<tr>
<td>0.435</td>
<td>B</td>
<td>78.085</td>
</tr>
<tr>
<td>0.781</td>
<td>C</td>
<td>9.325</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Count</th>
<th>%A</th>
<th>%B</th>
<th>%C</th>
<th>%D</th>
<th>%Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>18316</td>
<td>12.333</td>
<td>78.085</td>
<td>9.325</td>
<td>0.257</td>
<td></td>
</tr>
</tbody>
</table>

**Flags:**

### Option/Score Distribution by Subgroup

<table>
<thead>
<tr>
<th></th>
<th>Total Count</th>
<th>P-value</th>
<th>Item-Total Corr</th>
<th>%A</th>
<th>%B</th>
<th>%C</th>
<th>%D</th>
<th>%Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>18316</td>
<td>0.781</td>
<td>0.435</td>
<td>12.333</td>
<td>78.085</td>
<td>9.325</td>
<td>0.257</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>9403</td>
<td>0.757</td>
<td>0.441</td>
<td>13.762</td>
<td>75.742</td>
<td>10.156</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>8869</td>
<td>0.806</td>
<td>0.421</td>
<td>10.824</td>
<td>80.595</td>
<td>8.411</td>
<td>0.169</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>11639</td>
<td>0.817</td>
<td>0.42</td>
<td>10.345</td>
<td>81.691</td>
<td>7.664</td>
<td>0.301</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1566</td>
<td>0.815</td>
<td>0.44</td>
<td>10.856</td>
<td>81.481</td>
<td>7.535</td>
<td>0.128</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>1062</td>
<td>0.717</td>
<td>0.421</td>
<td>17.137</td>
<td>71.657</td>
<td>11.017</td>
<td>0.188</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>2966</td>
<td>0.655</td>
<td>0.388</td>
<td>18.678</td>
<td>65.509</td>
<td>15.644</td>
<td>0.169</td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td>456</td>
<td>0.708</td>
<td>0.398</td>
<td>16.667</td>
<td>70.833</td>
<td>12.281</td>
<td>0.219</td>
<td></td>
</tr>
</tbody>
</table>

### DIF Summary

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DIF Statistic</th>
<th>DIF Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male vs. Female</td>
<td>0.132</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Asian</td>
<td>-0.23</td>
<td>A</td>
</tr>
<tr>
<td>White vs. African American</td>
<td>-0.112</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Hispanic</td>
<td>-0.696</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Native American</td>
<td>-0.173</td>
<td>A</td>
</tr>
</tbody>
</table>
### Item Information

<table>
<thead>
<tr>
<th>Grade</th>
<th>Item Code</th>
<th>Form</th>
<th>Position</th>
<th>Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>WAM10408</td>
<td>A</td>
<td>20</td>
<td>SA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PE</td>
<td>4.1.I</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Item Information Summary

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
<th>%</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch</td>
<td>0.556</td>
<td>0</td>
<td>40.181</td>
</tr>
<tr>
<td>Item-Total Corr</td>
<td>0.594</td>
<td>1</td>
<td>16.566</td>
</tr>
<tr>
<td>Item Mean</td>
<td>1.025</td>
<td>2</td>
<td>42.651</td>
</tr>
<tr>
<td>Total Count</td>
<td>1660</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Option/Score Distribution by Subgroup

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Total Count</th>
<th>Item Mean</th>
<th>Item-Total Corr</th>
<th>% 0 Pt</th>
<th>% 1 Pt</th>
<th>% 2 Pt</th>
<th>%Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>1660</td>
<td>1.025</td>
<td>0.594</td>
<td>40.181</td>
<td>16.566</td>
<td>42.651</td>
<td>0.602</td>
</tr>
<tr>
<td>Male</td>
<td>883</td>
<td>1.017</td>
<td>0.607</td>
<td>40.657</td>
<td>16.082</td>
<td>42.356</td>
<td>0.906</td>
</tr>
<tr>
<td>Female</td>
<td>776</td>
<td>1.032</td>
<td>0.582</td>
<td>39.691</td>
<td>17.139</td>
<td>42.912</td>
<td>0.258</td>
</tr>
<tr>
<td>White</td>
<td>1059</td>
<td>1.103</td>
<td>0.583</td>
<td>34.939</td>
<td>19.263</td>
<td>45.137</td>
<td>0.661</td>
</tr>
<tr>
<td>Asian</td>
<td>142</td>
<td>1.296</td>
<td>0.56</td>
<td>31.69</td>
<td>7.042</td>
<td>61.268</td>
<td>0</td>
</tr>
<tr>
<td>African American</td>
<td>106</td>
<td>0.667</td>
<td>0.671</td>
<td>61.321</td>
<td>9.434</td>
<td>28.302</td>
<td>0.943</td>
</tr>
<tr>
<td>Hispanic</td>
<td>254</td>
<td>0.719</td>
<td>0.534</td>
<td>57.48</td>
<td>12.598</td>
<td>29.528</td>
<td>0.394</td>
</tr>
<tr>
<td>Native American</td>
<td>49</td>
<td>0.875</td>
<td>0.55</td>
<td>42.857</td>
<td>24.49</td>
<td>30.612</td>
<td>2.041</td>
</tr>
</tbody>
</table>

### DIF Summary

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DIF Statistic</th>
<th>DIF Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male vs. Female</td>
<td>0.046</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Asian</td>
<td>0.048</td>
<td>A</td>
</tr>
<tr>
<td>White vs. African American</td>
<td>-0.102</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Hispanic</td>
<td>-0.005</td>
<td>A</td>
</tr>
<tr>
<td>White vs. Native American</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D: WASHINGTON ASSESSMENT OF STUDENT LEARNING ADVISORY MEMBERS
National Technical Advisory Committee Members

Patricia Almond, University of Oregon
Peter Behuniak, University of Connecticut
Richard Duran, Professor, University of California – Santa Barbara
George Engelhard, Professor, Emory University
Robert Linn, Professor Emeritus, University of Colorado and UCLA/CRESST
William Mehrens, Professor Emeritus, Michigan State University
James Popham, Professor Emeritus, University of California, Los Angeles
Joseph Ryan, Professor Emeritus, Arizona State University
Catherine Taylor, Associate Professor, University of Washington

Washington State Assessment Advisory Team

Richard Basnaw, Director of Curriculum, Instruction & Assessment, Chief Leschi Tribal Schools
Anne DeHaven, Director of Assessment and Student Services, Lake Stevens School District
Phil Dommes, Director of Assessment, North Thurston Public Schools
Tersea Easley, Student Assessment Coordinator, Tacoma School District
Linda Elman, Director of Research and Evaluation, Central Kitsap School District
Barbara Gilbert, Director of Student Services, Highland School District #509
LaVonne Grimes, Special Services Director, Chimacum School District #049
Bev Henderson, Assessment and Staff Development Coordinator, Kennewick School District
Feng-Yi Hung, Director of Assessment and Program Evaluation, Clover Park School District
Mike Jacobson, Assessment and Curriculum Director, White River School District
Nancy Katims, Director of Assessment, Research, and Evaluation, Edmonds School District
Debbie Lahue, Director of Humanities Center for Instructional Services
June Lee, Testing Coordinator & School Counselor, Soap Lake School District
Barbara Lomas, Director, School Improvement and Professional Development
Allan Miedema, Information Systems Manager, Northshore School District
Brian Rick, Assessment and Evaluation Specialist, Bellingham Public Schools
Lorna Spear, Executive Director of Teaching and Learning, Spokane Public Schools
Nancy Steers, District Assessment Coordinator, Seattle Public Schools
Dawn Wakeley, Associate Director Teaching and Learning, Tahoma School District