FINAL

Washington Assessment of Student Learning

Grade 4

2006

Technical Report

Prepared by
Pearson Educational Measurement

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

April 18, 2007

FINAL

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND GLOSSARY

Abbreviation or Term	Meaning
AS	Algebraic Sense (content) Mathematics strand
AS²	Application of Science strand
CONV	Writing Conventions strand
COS	Content, Organization, & Style Writing strand
CU	Communicates Understanding (process) Mathematics strand
EALR	Essential Academic Learning Requirements
form	Operational items and imbedded pilot items that uniquely define a (test) form
GLE	Grade Level Equivalents
GS	Geometric Sense (content) Mathematics strand
IA	Informational Analysis Reading strand
IC	Informational Comprehension Reading strand
IEP	Individual Education Program
IRT	Item Response Theory
IS	Inquiry in Science strand
IT	Informational Thinking Critically Reading strand
LA	Literary Analysis Reading strand
LC	Literary Comprehension Reading strand
LT	Literary Thinking Critically Reading strand
MC	Makes Connections (process) Mathematics strand
ME	Measurement (content) Mathematics strand
NS	Number Sense (content) Mathematics strand
OSPI	Office of the Superintendent of Public Instruction
PCM	Partial Credit Model
Pearson	Pearson Educational Measurement
PS	Probability and Statistics (content) Mathematics strand
PSC	Performance Scoring Center
SD	Standard Deviation
s.e.m.	Standard Error of Measurement
SR	Solves Problems & Reasons Logically (process) Mathematics strand
SS	Systems of Science strand
test	Operational test items in a testing booklet that contribute to reported student scores
WAAS	Washington Alternate Assessment System
WASL	Washington Assessment of Student Learning

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 technical reports previously produced by The Riverside Publishing Company, and is one component of a suite of reports that documents the properties and characteristics of the 2006 *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 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 become 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 WASL.

This report is limited to the results of the students in Grade 4 who took the assessments.

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, those with Section 504 plans, 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 from Pearson Educational Measurement, 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 maintain the same performance level standards 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. The *Tool Kits* also provide content frameworks to assist teachers at all grade levels to relate their classroom learning goals and instruction to the EALRs. (See http://www.k12.wa.us/assessment/toolkits/default.aspx for links to the *Tool Kits*.)

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 will be required to earn a Certificate of Academic Achievement to obtain a 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 assessments.

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, 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 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-MO are newly introduced alternatives to regular WASL administration for eligible students. 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. When a standard is defined on a criterion-referenced test, examinee scores are compared to the standard 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, WASL performance is reported at the individual, school, district, and state levels. The information in these reports 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 2006 TESTS

The Grade 4 2006 *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 2006 test.

Table 1. 2006 Grade 4 Reading Items - Content Classification

Type of Reading Passage	Test Strand	Number of Items
Literary ‡	Comprehension †	9
	Analysis †	8
Informational ‡	Comprehension †	5
	Analysis †	7
Total Number of Items		29

* 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. 2006 Grade 4 Writing Prompts - Content Classification

Task	Purposes [1]	Process [2]	Number of Prompts	Scores [3]
Extended Piece	Narrative	• prewrite • first draft • revise • edit • final draft	1	• Content, Organization & Style • Writing Conventions
Extended Piece	Inform	• prewrite • first draft • revise • edit • final draft	1	• Content, Organization & Style • Writing Conventions
Total Number of Prompts			2	

[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 3. 2006 Grade 4 Mathematics Items - Content Classification

Process Strand	Concept Strand	Number of Items
Concepts & Procedures	Number Sense [1]	5
	Measurement [1]	5
	Geometric Sense [1]	4
	Probability and Statistics [1]	5
	Algebraic Sense [1]	5
Solves Problems [2] & Reasons Logically [3]		5
Communicates Understanding [4]		3
Making Connections [5]		3
Total Number of Items		35

[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; Mathematics EALR 3 The student uses mathematical reasoning.

[3] Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[4] Mathematics EALR 5: The student makes mathematical connections.

SCHEDULE FOR TESTING – 4th GRADE - SPRING 2006

Grade 4 Reading, Writing, and Mathematics tests were administered within the April 17 – May 5 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 three 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. 2006 Grade 4 State Standardized Testing Schedule

Subject	Testing Window	Schedule
Reading	April 17 – May 5	Approved locally
Writing	April 17 – May 5	Approved locally
Mathematics	April 17 – May 5	Approved locally

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 2006 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.

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 Pearson Educational Measurement (Pearson) 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, Mathematics, and Science tests: multiple choice, short answer, and extended response. For each multiple-choice item, students 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 two points (scored 0, 1, or 2) and extended-response items are worth 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.[1] The most recent versions of these specifications are available through the web site 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] 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

Text types/Strands	No. of Reading Selections	No. of Words Per Passage	No. of Multiple-Choice Items	No. of Short Answer Items	No. of Extended Response Items
Literary ‡	3	50-800	*10*	*2-4*	*1*
Comprehension †			5	1-2	0
Analysis †			5	1-2	1
Informational	3	150-700	*10*	*2-4*	*1*
Comprehension †			5	1-2	0
Analysis †			5	1-2	1
Total	6	1800-2200	20	7	2

* 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

Strands	Scored 0-2 Points	Scored 0-4 Points
Narrative		
COnent & Style		1
CONVentions & Mechanics	1	
Expository		
COnent & Style		1
CONVentions & Mechanics	1	
Total Points	4	8

[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 7. Grade 4 Mathematics Test Design

Strands	Multiple Choice	Short Answer	Extended Response
Number Sense [1]	2-4	1-2	0
Measurement Concepts [1]	2-4	1-2	0
Geometric Sense [1]	2-4	1-2	0
Probability and Statistics Procedures [1]	2-4	1-2	0
Algebraic Sense [1]	2-4	1-2	0
Solves Problems & Reasons Logically [2]	0-1	1-2	2
Communicates Understanding [3]	0	2-3	1
Making Connections [4]	1-2	1-2	0
Maximum Number of Items	21	11	3
Maximum Number of Points	21	22	12

[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 and Mathematics EALR 3 The student uses mathematical reasoning.

[3] Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[4] 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 Pearson 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 Pearson's full-time staff of Content Specialist professionals who have, on average, 14 years of classroom and pedagogical experience. All Pearson item writers receive in-depth training before actively working on a Pearson 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 are also imbedded 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 be reasonable estimates to the data 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 2003. Newly developed pilot prompts will be administered to a volunteer sample in a stand-alone pilot program in Fall 2006 to replenish the prompt bank.

For each pilot form, at least 1200 student responses are scored. Of the 1200 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*.

CALIBRATION, SCALING, AND ITEM ANALYSIS

After pilot 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 statistics, based on classical test theory, include item means and item-test correlations. The Rasch Partial Credit Model is one class of mathematical models based on modern item response theory, used to estimate item location and item fit statistics. A generalized Cochran Mantel-Haenszel chi-square and a generalized Mantel-Haenszel alpha odds ratio are computed for each pilot item to evaluate the presence and directionality of differential item functioning or "item bias" for each pilot item. Differential item functioning is observed when examinees from different demographic groups with the same ability perform differently on the same item.

IRT Analysis

The Rasch Partial Credit Model is a class of Item Response Theory (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:

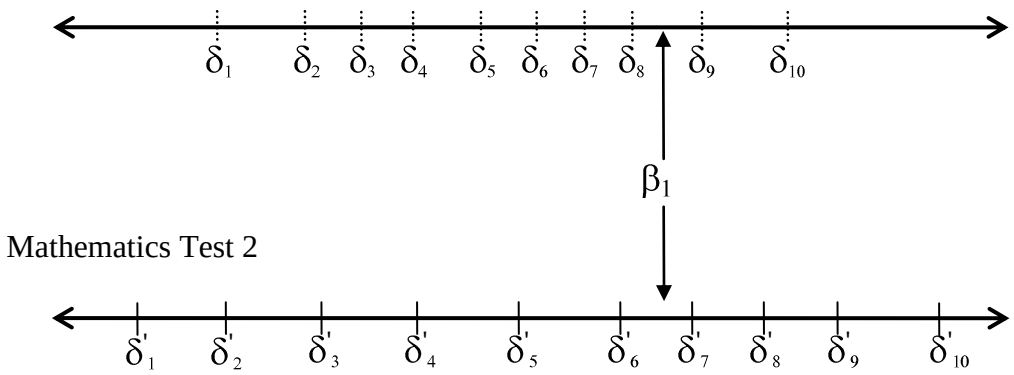
$$P_{xni} = \frac{\exp\sum_{j=0}^{x}(B_n - D_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(B_n - D_{ij})} \qquad \text{(Equation 1)}$$

where $x = 0, 1, 2, \ldots, m - 1$;
B_n = person parameter;
D_{ij} = item-category parameter; and

$$\sum_{j=0}^{m-1}(B_n - D_{ij}) = 0 \qquad \text{(Equation 2)}$$

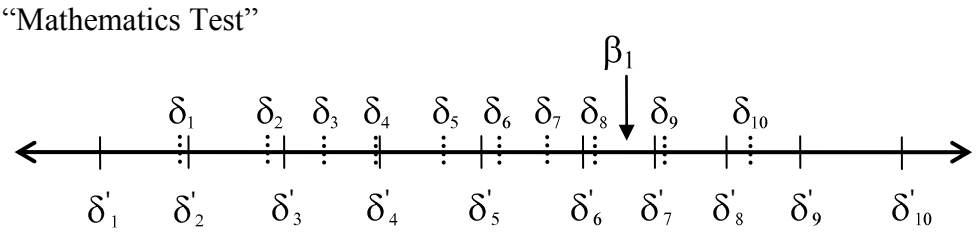
Item difficulties and examinee abilities can be estimated for a test using this mathematical model. The item difficulty is the location on the ability scale where examinees have a 50/50 chance of answering an item correctly. Figure 1 illustrates the relationship between examinee ability and item difficulty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

Figure 1. Location of examinee β_1 on two tests with different items



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 (β_1, β_2, β_3, . . .,β'_{10}) are placed on the same scale, the examinee's ability can be reported relative to an underlying, common scale – a value between δ_8 (from Test 1) and δ'_7 (from Test 2).

Figure 2. Location of examinee β_1 on the same "Mathematics Test" scale



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:

$$P_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)} \qquad \text{(Equation 3)}$$

where B_n = 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, and Mathematics 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 operational items. For Reading and Mathematics tests, the same set of operational items appeared in all test forms, but different sets of pilot items were imbedded in or appended to the operational sections. Pilot items were then calibrated and scaled to the grade level/content area scale through the common operational items.

Traditional Item Analysis

For multiple-choice items, item means or p-values and item-test correlations or point-biserials are computed. 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 0.0 to 1.0. 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.

Bias Analysis

The Mantel-Haenszel statistic is a chi-square (χ^2) statistic. Examinees are separated into relevant subgroups based on ethnicity or gender. Examinees in each subgroup are ranked relative to their total test score. 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. Multiple 2×2 contingency tables are created for each item by each total test score and for every demographic contrast. The 2×2 contingency tables represent the number of examinees at a specific total test score in each subgroup who correctly answered the item and the number of examinees in each group who answered incorrectly. Table 8 is an example of a 2×2 table of performance on hypothetical multiple-choice "Item X" for males and females with Total Test Score Y_i for a gender contrast. Among these 200 examinees with total test score Y_i, the item appears to be more difficult for females than for males, and fewer examinees overall correctly answered the item.

Table 8. Scores on "Item X" for Examinees with Total Test Score Y_i by Gender

Item X	Number Responding Correctly	Number Responding Incorrectly
Males (N = 100)	50	50
Females (N = 100)	30	70

Examinees with Total Test Score = Y_i

To compute the Mantel-Haenszel statistic, similar 2×2 tables are created at every total test score. A χ^2 statistic is computed for each 2×2 table and the sum of all χ^2 statistics yields the total bias statistic for a single item. A generalized Mantel-Haenszel statistic is computed for polytomous items using all item score points. Items that demonstrate a high $\Sigma\chi^2$ are flagged for potential bias. Generally, a certain percentage of items in any given pool of items will be flagged for item bias by chance alone. Careful review of items can help to identify whether some characteristic of an item may cause the bias (e.g., the content or language is unfamiliar to girls) or whether the bias flag is likely a result of statistical error.

Mantel-Haenszel item 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. Mantel-

Haenszel item statistics are not computed on operational items. Table 9 summarizes the percentage of items with statistically significant Mantel-Haenszel statistics from the 2006 pilot. The 2006 operational tests are comprised of items that were piloted in years prior to 2006, which were reviewed and approved by Content Review, Bias and Fairness Review, and Data Review Committees.

Table 9. Percent of Pilot 2006 Items with Statistically Significant Mantel-Haenszel Statistics

	MC pilot items					SA pilot items					ER pilot items				
	White-Asian	White-Black	White-Hispanic	White-Native American	Male-Female	White-Asian	White-Black	White-Hispanic	White-Native American	Male-Female	White-Asian	White-Black	White-Hispanic	White-Native American	Male-Female
Grade 4 Reading	10.0%	15.0%	25.05%	6.3%	22.5%	22.2%	0.0%	0.0%	11.1%	11.1%	13.0%	8.7%	17.4%	4.3%	21.7%
# Pilot Items	80					9					23				
Grade 4 Writing	NA										NA				100%
# Pilot Items											11				
Grade 4 Mathematics	15.9%	12.7%	23.8%	6.3%	19.0%	50.0%	0.0%	50.0%	0.0%	50.0%	9.1%	12.1%	12.1%	9.1%	36.4%
# Pilot Items	63					2					33				

The Grade 4 Writing Pilot was last administered in Fall 2006 to develop the current bank of operational prompt pairs. Further details about the pilot design, analysis procedures, and pilot results are provided in *Summary Report of the 2006 Fall Grade 5 Writing Pilot for the WASL Grade 4 Writing Assessment*.

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. 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, Pearson content specialists, and Pearson psychometricians 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.

Data Review meetings are usually conducted in late autumn and early winter to evaluate items piloted during the previous spring. The summary results from Data Review meetings are not available until late winter or early spring of the following year. In 2005, Data Review meetings were convened for Reading and Mathematics content areas at Grade 4.

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

New operational forms are created for each test administration, usually sometime in the spring after Data Review. Building an operational form is a complex puzzle. OSPI content specialists, Pearson 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, represent an overall test with balanced context. 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, 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 and Mathematics 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. Each test has target statistical characteristics and criteria. The better the match to these criteria, the better the equating accuracy of test scores between different test administrations. The test developer's objective is to construct a new, parallel operational test form for each administration.

The weighted mean Rasch difficulty is used to construct an operational test form of the same level of difficulty from administration to administration. 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 tend to shift. Table 10 lists the empirical weighted mean Rasch values from 2000 through 2006 and the predicted values for 2006 Reading and Mathematics tests.

Table 10. Empirical Weighted Mean Rasch of 2000 ~ 2006 Grade 4 Reading & Mathematics Tests

		Empirical							Predicted	
Subject		2000	2001	2002	2003	2004	2005	2006	2005	2006
Reading	Mean Rasch	0.43	0.62	-0.01	0.04	0.44	0.07	0.27	-0.05	0.13
	Cut Score	28 out of 43	24 out of 40	30 out of 40	28 out of 40	27 out of 40	30 out of 40	31 out of 42	28 out of 40	31 out of 42
	% Correct	65.1%	60.0%	75.0%	70.0%	67.5%	75.0%	73.8%	70.0%	73.8%
Mathematics	Mean Rasch	0.37	0.23	0.00	0.05	-0.02	0.21	0.11	0.13	0.05
	Cut Score	35 out of 62	33 out of 56	36 out of 55	34 out of 54	35 out of 55	33 out of 55	35 out of 55	33 out of 55	35 out of 55
	% Correct	56.5%	58.9%	67.3%	63.0%	63.6%	60.0%	63.6%	60.0%	63.6%

PART 3: VALIDITY

An important issue in test development is the degree to which the achievement test elicits the conceptual understanding and skills it is intended to measure. If students must use logical reasoning skills to respond to an item, for example, we need evidence that the item elicits logical reasoning in students' responses rather than memorization. Validity is an evaluative judgment about the degree to which test scores represent the intended construct. There are several different strategies to obtain evidence for the validity of test scores (Messick, 1989):

1. We can look at the content of the test in relation to the content of the domain of reference;
2. We can probe the ways in which individuals respond to the items or tasks;
3. We can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses;
4. We can survey relationships of test scores with other measures and background variables, that is, the test's external structure;
5. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions;
6. Finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

Validity is a judgment about the relationships between a test score and its context (including the instructional practices and the examinee), the knowledge and skills it represents, the intended interpretations and uses, and the consequences of its interpretation and use. Messick stated that multiple sources of evidence are needed to investigate the validity of assessments. The following sections provide descriptions about available validity evidence for the Grade 4 WASL, pertaining to types of validity evidence 1~3 above. Concurrent, predictive, and consequential validity evidence are not relevant to the intended uses of the criterion-referenced WASL tests. The evidence includes correlations among scores and strands within the WASL and factor analysis studies to examine the construct validity of WASL.

CONTENT VALIDITY

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.

CONSTRUCT VALIDITY

Content representation and item quality are important aspects of a test, but they do not ensure the valid interpretation of test scores. To evaluate test score validity, it is important to determine whether the internal structure of the test is consistent, and whether subsets of items that purport to measure a particular construct do so consistently and in concert. This type of evidence represents the construct validity of test scores.

Studies were previously conducted to gather construct validity evidence for the Grade 4 WASL Reading, Writing, and Mathematics tests. The *WASL Technical Reports for Grade 4* from 1997 to 2002 provide construct validity information for the 1998 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, the internal structure of WASL was evaluated through correlational analysis between strand scores on WASL content area tests.

Correlations Among WASL Strand Scores

Table 11 lists the intercorrelations of strand scores between different 2006 WASL content area tests. These intercorrelations were completed only using the 70,755 cases for which all three Grade 4 WASL content area scores were available for analysis. Scores for Reading strands (Literary Comprehension, Literary Analysis, Informational Comprehension, Informational Analysis) exhibit correlations between 0.489 and 0.594. The Writing Content, Organization, & Style strand score correlated 0.507 with the Writing Conventions strand score. Intercorrelations of Mathematics concepts strand scores (Number Sense, Measurement, Geometric Sense, Probability and Statistics, and Algebraic Sense) range from 0.405 to 0.533. Shavelson, Baxter, & Gao (1993) showed that students perform differently on mathematical tasks that tap different mathematics skills. Intercorrelations between the Mathematics process scores (Solves Problems and Reasons Logically, Communicates Understanding, and Makes Connections) are modest but slightly higher than intercorrelations between the Mathematics concept scores (0.481 to 0.578) suggesting that the skills required in these strands share a common construct. Most mathematics items in the process strands, in fact, involve short-answer and extended response item formats.

Intercorrelations between Mathematics content strand scores and Mathematics process strand scores are informative. Process strand scores for Solves Problems and Reasons Logically, Communicates Understanding, and Makes Connections are moderately well correlated with strand scores for all other content area strand scores (0.405 to 0.597).

Table 11. 2006 Grade 4 WASL Strand Score Intercorrelations

Strands	LC	LA	IC	IA	COS	CONV	NS	ME	GS	PS	AS	SR	CU	MC
LA	0.511	1												
IC	0.551	0.489	1											
IA	0.566	0.587	0.594	1										
COS	0.409	0.442	0.417	0.493	1									
CONV	0.444	0.431	0.435	0.520	0.507	1								
NS	0.422	0.368	0.450	0.473	0.350	0.401	1							
ME	0.379	0.343	0.391	0.416	0.335	0.365	0.488	1						
GS	0.397	0.367	0.416	0.447	0.331	0.376	0.469	0.405	1					
PS	0.487	0.417	0.484	0.506	0.371	0.416	0.505	0.423	0.475	1				
AS	0.477	0.416	0.480	0.503	0.365	0.419	0.533	0.456	0.475	0.526	1			
SR	0.543	0.506	0.547	0.599	0.462	0.507	0.566	0.493	0.513	0.582	0.597	1		
CU	0.436	0.436	0.465	0.514	0.419	0.447	0.488	0.426	0.483	0.496	0.490	0.578	1	
MC	0.382	0.367	0.412	0.447	0.357	0.374	0.517	0.441	0.435	0.462	0.480	0.546	0.481	1

LC-Literary Comprehension
LA- Literary Analysis
IC-Informational Comprehension
IA-Informational Analysis
COS-Content, Organization, & Style
CONV-Writing Conventions

NS-Number Sense
ME-Measurement
GS-Geometric Sense
PS-Probability and Statistics
AS-Algebraic Sense
SR-Solves Problems & Reasons Logically
CU-Communicates Understanding
MC-Makes Connections

Intercorrelations between Reading strand scores and Mathematics content strand scores are low to moderate (0.343 to 0.506). The intercorrelations between Reading strand scores and Mathematics process strand scores are slightly higher, but also modest (0.367 to 0.599). Intercorrelations between Writing strand scores and Mathematics strand scores are modest (0.331 to 0.507). Writing strand scores share higher correlations with Reading strand scores (0.409 to 0.520) than with Mathematics strand scores. These intercorrelations suggest that, for Reading, Writing, and Mathematics tests, writing skill, critical thinking, and synthesis are moderately related to performance. To further investigate the relationships between Reading, Writing, and Mathematics, an exploratory factor analysis was completed on the content area strand scores.

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 percent of the variance is explained. The eigenvalues suggested a three-factor solution that explained 62 percent of the total variance. Table 12 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 57% of the total variance. Kaiser's measure of sampling adequacy, a summary of how much smaller the partial correlations are than the original correlations, all exhibited values higher than 0.95.

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. The residual correlations are low, with the largest value of 0.094 between Writing COS and Writing CONV strand scores. Table 12 shows the rotated factor pattern based on three retained factors.

Table 12. 2006 Grade 4 Rotated Factor Pattern on WASL Tests for Three-Factor Solution

Variables	Factor 1	Factor 2	Factor 3
Number Sense (Math)	0.739*	0.236	0.150
Makes Connections (Math)	0.706	0.158	0.230
Measurement (Math)	0.677	0.122^	0.239
Algebraic Sense (Math)	0.644	0.411	0.108
Geometric Sense (Math)	0.643	0.285	0.123
Solves Problems & Reasons Logically (Math)	0.615	0.466	0.287
Probability & Statistics (Math)	0.591	0.467	0.097^
Communicates Understanding (Math)	0.582	0.311	0.339
Literary Comprehension (Reading)	0.280	0.754*	0.165
Informational Comprehension (Reading)	0.334	0.707	0.176
Literary Analysis (Reading)	0.172^	0.698	0.339
Informational Analysis (Reading)	0.338	0.662	0.371
Content, Organization, & Style (Writing)	0.202	0.254	0.816*
Conventions (Writing)	0.296	0.289	0.711

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

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, Section 504 Programs, 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, with consistently higher proportions of tested students that meet standard in all content area tests. Mean scaled scores are higher than the Level 3 "Proficient" cut score only in Reading for both males and females and only for females in Writing. Mean scaled scores are near but below the Level 3 "Proficient" cut score in Mathematics for both males and females.

SUMMARY

The results of these analyses provide evidence of validity based on test content and content area constructs of the 2006 Grade 4 WASL. Although achievement in one subject area is generally related to achievement in other subject areas, an examination of WASL strand scores suggest that Reading, Writing, and Mathematics comprise different underlying dimensions of academic achievement and performance on the WASL tests.

Reference

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

Shavelson, R. J., Baxter, G. P., Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

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. How consistently does each examinee perform on all of the items within a test? Internal consistency can be estimated by Cronbach's coefficient alpha. When a test is composed entirely of dichotomously scored multiple-choice items, a modification of Cronbach's alpha can be used (KR-20). When a test includes polytomously scored items, the internal consistency estimate is computed 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:

$$r_{xx'} = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum s_i^2}{s_x^2}\right) \qquad \text{(Equation 4)}$$

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 2006 Grade 4 WASL tests are listed in Table 13. The 2006 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.

Table 13. 2006 Grade 4 WASL Test & Content Strand Reliability Estimates

Strand	Alpha Coefficient	Raw Score Standard Error of Measurement
Reading	0.84	2.72
LC	0.61	1.08
LA	0.58	1.71
IC	0.52	0.98
IA	0.61	1.56
Writing	0.74	1.00
COS	0.63	0.77
CONV	0.71	0.53
Mathematics	0.88	3.74
NS	0.54	1.15
ME	0.39	1.09
GS	0.40	1.09
PS	0.53	1.02
AS	0.49	1.05
SR	0.63	1.89
CU	0.47	1.58
MC	0.44	1.11

LC-Literary Comprehension
LA- Literary Analysis
IC-Informational Comprehension
IA-Informational Analysis
COS-Content, Organization, & Style
CONV-Writing Conventions

NS-Number Sense
ME-Measurement
GS-Geometric Sense
PS-Probability and Statistics
AS-Algebraic Sense
SR-Solves Problems & Reasons Logically
CU-Communicates Understanding
MC-Makes Connections

STANDARD ERROR OF MEASUREMENT

One way to interpret the reliability of test scores is with the conditional standard error of measurement (s.e.m.). The s.e.m. is an estimate of the standardized distribution of error around a particular score. An observed score bounded by one s.e.m. represents a 68 percent probability that, over repeated observations, an examinee's true score estimate falls within the band. A two-s.e.m. 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 s.e.m. from:

$$\text{s.e.m.} = s_x\sqrt{1-r_{xx'}} \qquad \text{(Equation 5)}$$

where: s_x is the observed score standard deviation, and

$r_{xx'}$ is the reliability estimate or alpha coefficient.

Tables 17 and 18 list the 2006 Grade 4 conditional standard errors of measurement for the WASL Reading and Mathematics tests on the scaled score metric. Table 19 also includes the 2006 Grade 4 conditional standard errors of measurement for the WASL Writing test on the raw score metric.

INTERJUDGE AGREEMENT

Part 7 describes aspects about polytomous item scoring. Because constructed response items are scored by trained human readers, inter-rater agreement is another important facet of the consistent application of scoring standards and the subsequent reliability of test scores. 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 with increasing the numbers of different items, judges, students' responses, and ranges of item score points. The quality of inter-rater 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
3. kappa coefficient.

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.

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 inter-rater reliability that incorporates a correction for the rate of chance agreement. Kappa is computed by:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \qquad \text{(Equation 6)}$$

where p_a = overall proportion of exact agreement

p_e = overall proportion of chance agreement = $\sum_{i=1}^{m} p_{i\bullet} p_{\bullet i}$, for item score points i to m.

(At the time of this report preparation, the necessary data file components were not available for analysis.)

SUMMARY

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. 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.

PART 5: SCALING AND EQUATING

The 2006 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 "Meets Standard" (Level 3) and a scaled score of 375 is the cut score for "Proficient" (Level 2).

SCALED SCORE DEVELOPMENT

Scores on the WASL are reported as scaled scores. Tables 17 and 18 provide the 2006 Grade 4 number correct to scaled scores conversions for each test. 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 yard-stick, 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).[2]

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 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:

$$\mathbf{a}*\theta + \mathbf{b} = \text{scaled score} \qquad \text{(Equation 7)}$$

where **a** is the slope and **b** is the intercept of the linear transformation to scaled scores.

---

[2] 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.

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*(θ associated with Level 3 "Proficient") + **b** = 400
a*(θ associated with Level 2 "Basic") + **b** = 375 (Equation 8)

Table 14 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 14. Theta to Scaled Score Linear Transformation Equations

Content Area	θ at Level 2 "Basic" (Scaled Score 375)	θ at Level 3 "Proficient" (Scaled Score 400)	θ to Scaled Score Equation
Reading	-0.331	0.952	Scaled Score = 19.4856*θ + 381.4497
Writing †	NA		
Mathematics	-0.090	0.572	Scaled Score = 37.76435*θ + 378.3988

† 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 15 below.

Table 15. Scaled Score Ranges for Performance Level Categories

Content Area	Level 1 "Below Basic"	Level 2 "Basic"	Level 3 "Proficient"	Level 4 "Advanced"
Reading	275-374	375-399	400-423	424-475
Writing †	0-6	7-8	9-10	11-12
Mathematics	200-374	375-399	400-426	427-550

† 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 detail about Steps 1-3 above are described in the annual equating report, *WASL Grades 4/5/7/8 2006 Equating Study Technical Report.]*

4. Item parameter estimates resulting from Step 3, above, 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.

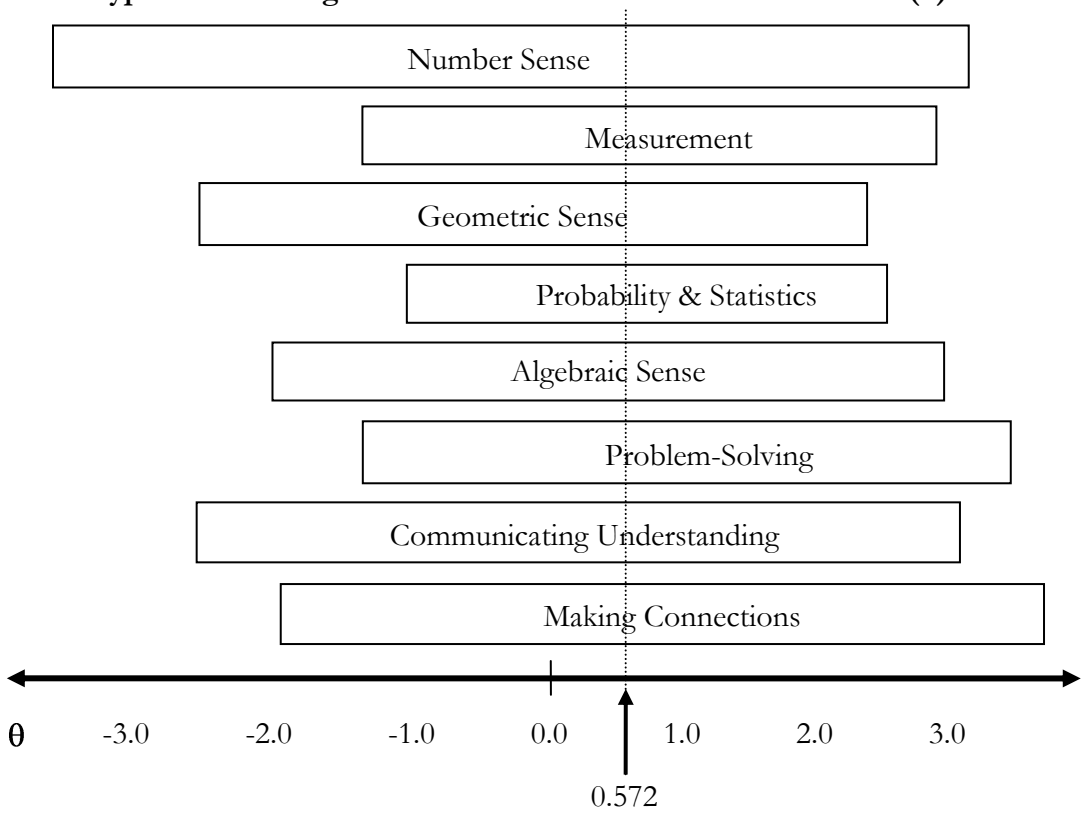
Table 16 lists the strand score and strand θ ranges, and the raw cut points that operationalize the "+/–" content strand cut point. The Writing test is not equated from year to year on a scale score metric, and therefore have no corresponding "+/–" content strand cut points.

Table 16. Content Strand Cut-Points

	Strand	θ Range	Max Raw Strand Score	"+" Strand	"-" Strand
Reading	LC	-4.291 ~ 3.077	10	0 – 7	8 – 10
	LA	-4.483 ~ 4.855	14	0 – 7	8 – 14
	IC	-2.610 ~ 3.287	6	0 – 3	4 – 6
	IA	-3.296 ~ 4.658	12	0 – 6	7 - 12
Mathematics	NS	-2.980 ~ 2.988	6	0 – 3	4 – 6
	ME	-2.860 ~ 3.870	6	0 – 3	4 – 6
	GS	-2.894 ~ 2.323	5	0 – 3	4 – 5
	PS	-3.426 ~ 2.730	6	0 – 4	5 – 6
	AS	-4.321 ~ 2.173	6	0 – 4	5 – 6
	SR	-2.927 ~ 3.531	13	0 – 9	10 – 13
	CU	-1.672 ~ 3.873	8	0 – 3	4 – 8
	MC	-1.857 ~ 3.341	5	0 – 2	3 – 5

Figure 3 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.

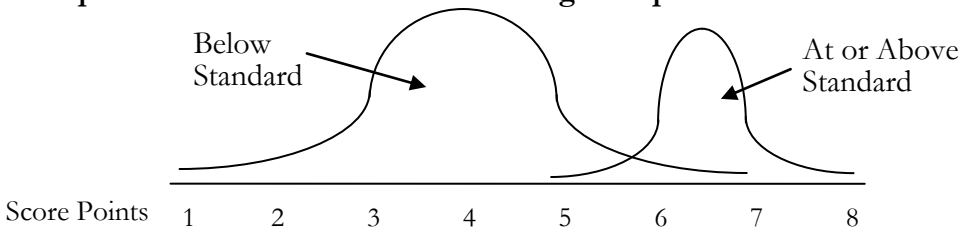
Figure 3. Hypothetical Range of Mathematics Strand Item Difficulties (θ)



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 4).

Figure 4. Sample Score Distribution of Contrasting Groups – COS Strand



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 Pearson 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.

Further details are described in the *WASL 2006 Grade 4/5/7/8 Equating Study Technical Report* and previous annual equating study technical reports.

Equating the Writing Test

For Writing, writing prompts were selected for the 2006 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 2003 to develop the current bank of operational prompt pairs.

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 17 and 18 include the raw score (Raw) to scaled score (SS) relationship for the 2006 Grade 4 Reading and Mathematics tests. Because the Writing test is already "scaled" on the raw score metric, there is no raw score to SS relationship. Table 19 lists the conditional standard errors of measurement at each Writing raw score point.

Table 17. 2006 Grade 4 Reading Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (s.e.m.)

Raw	Reading SS	Conditional s.e.m.
0	278	35.990
1	303	20.226
2	318	14.790
3	327	12.432
4	334	11.048
5	340	10.133
6	345	9.451
7	349	8.944
8	353	8.535
9	357	8.184
10	360	7.911
11	363	7.677
12	366	7.463
13	369	7.288
14	371	7.151
15	375	*7.015*
16	377	6.917
17	379	6.820
18	381	6.742
19	384	6.684
20	386	6.645
21	388	6.606
22	390	6.586
23	393	6.586
24	395	6.606
25	397	6.645
26	400	6.684
27	402	6.742
28	404	6.839
29	406	6.937
30	409	7.073
31	412	7.249
32	414	7.443
33	417	7.697
34	420	8.009
35	424	8.398
36	428	8.905
37	432	9.587
38	437	10.522
39	444	11.945
40	452	14.361
41	467	19.934
42	491	35.834

Table 18. 2006 Grade 4 Mathematics Raw Score (Raw) to Scaled Scores (SS) with Conditional Standard Errors of Measurement (s.e.m.)

Raw	Mathematics SS	Conditional s.e.m.	Raw	Mathematics SS	Conditional s.e.m.
0	184	69.411	28	383	9.668
1	231	38.595	29	385	9.743
2	259	27.795	30	388	9.819
3	276	23.074	31	391	9.932
4	288	20.279	32	393	10.008
5	298	18.353	33	396	10.159
6	306	16.881	34	400	10.272
7	313	15.748	35	401	10.423
8	319	14.804	36	404	10.574
9	325	14.011	37	407	10.763
10	330	13.331	38	411	10.952
11	334	12.727	39	414	11.178
12	338	12.198	40	417	11.443
13	342	11.745	41	421	11.707
14	346	11.329	42	424	12.009
15	349	10.989	43	428	12.387
16	352	10.687	44	433	12.764
17	355	10.423	45	437	13.255
18	358	10.196	46	442	13.822
19	361	10.045	47	447	14.502
20	363	9.894	48	453	15.332
21	366	9.781	49	460	16.352
22	368	9.705	50	467	17.749
23	371	9.630	51	477	19.637
24	373	9.592	52	488	22.432
25	376	9.592	53	504	27.228
26	378	9.592	54	531	38.142
27	380	9.630	55	577	69.184

Table 19. 2006 Grade 4 Writing Raw Scores (Raw) with Conditional Standard Errors of Measurement (s.e.m.)

Raw	Conditional s.e.m.
0	1.525
1	0.878
2	0.687
3	0.645
4	0.675
5	0.784
6	0.930
7	0.865
8	0.819
9	0.925
10	1.311
11	1.199
12	1.664

Reference

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

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.

QUALIFICATIONS OF SCORERS

Highly-qualified, experienced readers (scorers) are essential to achieving and maintaining consistency and reliability when scoring student-constructed (open-ended) responses. Readers selected to score the WASL tests are required to possess the following qualifications.

- A minimum of a bachelor's degree in an appropriate academic discipline with priority to English, English Education, Math, Math Education, Science, Science Education, or related fields.
- Demonstrable ability in performance assessment scoring.
- Teaching experience, especially at the elementary or secondary level, is preferred.

In 2006, Washington teachers were involved in the scoring of the open-ended responses. Teachers who wish to score are required to meet the same standards for selection and training criteria as professionally trained scorers hired by the test contractor. Table 20 provides the number of Washington teachers involved in scoring Grade 4 subjects. Involvement of teachers in the scoring of the WASL assessments is seen as a means to increase the knowledge of Washington teachers in the assessment of students. Some special education teachers are involved in the scoring as well. The number of teachers involved in scoring continues to increase each year.

Table 20. 2006 Grade 4 Washington Teacher Participation

Subject	Number of WA teachers involved in scoring	Item Type or Trait
Reading	NA	
Writing	32	Expository Writing
Mathematics	36	2 Short Answer, 1 Extended Response

Team leaders or scoring supervisors and table leaders, responsible for supervising small groups of scorers, are selected on the basis of demonstrated expertise in all facets of the scoring process including strong organizational abilities, leadership, and interpersonal communication skills.

RANGE-FINDING AND ANCHOR PAPERS

The thoughtful selection of papers for range-finding and the subsequent compilation of anchor papers and other training materials are the essential first steps to ensure that scoring is conducted consistently, reliably, and equitably.

In the range-finding process, OSPI facilitators, performance assessment and curriculum specialists working with team and table leaders and teachers from Washington, all become thoroughly familiar with and reach consensus on the scoring rules (rubrics) approved by the Content Committees for each open-ended item. The Performance Scoring Center (PSC) staff is responsible for preparing all training materials in consultation with and subject to approval from OSPI. These range-finding teams begin work with random selections of student responses for each item. They review these responses, select an appropriate range of responses, and placed them into packets, numbered for easy reference. The packets of responses are read independently by members of a team of the most experienced scorers. Following these independent readings and tentative ratings of the papers, the range finding group discusses both the common and divergent scores. From this work, they assemble tentative sets of example responses for each prompt.

The primary task of the range-finding committee then is the identification of anchor papers—exemplars that clearly and unambiguously represented the solid center of a score point as described in the scoring rubric. Those exemplary anchor papers form the basis, not only of scorer training, but of subsequent range-finding discussions as well.

Discussion is ongoing with the goal of identifying a sufficient pool of additional student responses for which consensus scores can be achieved and that illustrated the full range of student performance in response to the prompt or item. This pool of responses includes borderline responses – ones that appeared to straddle adjacent score points which therefore can present decision-making problems that trained scorers need to be able to resolve.

TRAINING MATERIALS

Following the range-finding sessions, the performance assessment specialists and team leaders finalize the anchor sets and other training materials, as identified in the range-finding meetings. The final anchor papers are chosen for their clarity in exemplifying the criteria defined in the scoring rubrics.

The anchor set for each 4-point question consists of a minimum of thirteen papers, three examples for each of the four score points and one example of a non-scorable paper. The anchor set for each 2-point question consists of a minimum of seven papers, three examples of each of each score point and one example of a non-scorable paper. Score point exemplars consist of one low, one solid mid-range, and one high example at each score point.

Additional training sets and qualifying sets of responses are selected to be used in scorer training. One training set consists of responses that are clear-cut examples of each score point; the second set consists of responses closer to the borderline between two score points. The training sets give scorers an introduction to the variety of responses they will encounter while scoring, as well as allowing them to develop their decision-making capability for scoring responses that do not fall clearly into one of the scoring levels. Calibration/validity papers to be circulated during scoring are also identified at this time, as are scorer qualifying sets.

After all training materials have been compiled, OSPI content specialists and assigned Pearson representatives document approval of all training materials to be used during the current year's scoring process.

Washington teachers and Pearson's professional scorers must be able to apply scoring standards to which they are trained in a consistent manner in order to qualify for scoring. Table 21 summarizes the 2006 scorer qualification rates between Washington teachers and Pearson professional scorers for those items scored by both groups. The remainder of constructed response items in each test that are not designated for teacher scoring are scored by Pearson professional scorers.

Table 21. 2006 Grade 4 WASL Scorer Qualification

Subject	Item Type or Trait	Percentage of WA Teachers Qualified for Scoring	Percentage of Pearson Scorers Qualified for Scoring
Reading	NA		
Writing	Expository Writing	100	100
Mathematics	Item # 07 SA Item # 20 ER Item # 36 SA	100 100 100	100 100 100

INTER-RATER RELIABILITY AND RATER CONSISTENCY

Scorer training for each prompt is led by performance assessment specialists and team leaders. The primary purpose of the training is to help the scorers understand the decisions made by the range-finding committee. Training also helps scorers internalize the scoring rubrics, so that they can effectively and consistently apply them.

Scorer training sessions include an introduction to the assessment itself. Scorers are informed of the parameters or context within which the students' performance was elicited. This gives scorers a better understanding of what types of responses can be expected, given such parameters as grade level, instruction or time limitations. Scorers next receive a description of the scoring rules that apply to the responses for each item.

The scoring rubrics are always presented in conjunction with the anchor papers. After presentation and discussion of the anchor papers, each scorer is given a training set consisting of ten papers. The scorers score the papers independently. When all scorers have scored the training set, their preliminary scores are collected for reference.

Group discussion of the scores assigned is the next step, allowing the scorers to raise questions about the application of the scoring rubric and giving them a context for those questions. The purpose of the discussion among the scorers in training is to establish a consensus to ensure consistency of scores between scorers. Even after scorers qualify for the scoring, training continues throughout the scoring of all responses to maintain high inter- and intra-rater reliability. Therefore, training is a continuous process and scorers are consistently given feedback as they score.

Frequent reliability checks are used to closely monitor the consistency of each scorer's performance over time. The primary method of monitoring scorers' performances is by a process called "back-reading." In back-reading, each table leader rereads and checks scores on an average of five to ten percent of each scorer's work each day, with a higher percentage early in the scoring. If a scorer is consistently assigning scores other than those the table leader would assign, the team leader and performance assessment specialist, together, retrain that scorer, using the original anchor papers and training materials. This continuous, on-the-spot checking provides an effective guard against "rater drift," (beginning to score higher or lower than the anchor paper scores). Scorers are replaced if they are unable to score consistently with the rubric and the anchor papers after significant training.

Tables 22 through 25 provide the rater agreement information for the open-ended items in the 2006 Grade 4 WASL. Two types of rater agreement were calculated from approximately 5 percent of the examinees randomly selected from the students' response booklets: score agreement for individual items and score agreement across the total score for the open-ended item set for each content area. For item-by-item interjudge agreement in Reading, the range of exact agreement was 77% to 95% and the range of exact and adjacent agreement was 99% to 100%. For interjudge agreement in Writing, the range of exact agreement was 84.2% to 89.1% and the exact and adjacent agreement was between 99.9% and 100%. For item-by-item interjudge agreement in Mathematics, the range of exact agreement was 82% to 97% and the range of exact and the range of exact and adjacent agreement was 97% to 100%.

Table 22. 2006 Grade 4 Reading – Interrater Percent Agreement

Item	Points Possible	Number of Papers Scored	% Exact Agreement	% Adjacent + Exact Agreement	% Non-Adjacent Agreement
4	2	8042	91	99	0
6	2	8028	93	100	0
8	2	8038	95	99	0
15	2	8028	94	100	0
18	4	8012	85	99	0
22	2	8046	95	99	0
24	2	8038	88	100	0
28	4	7891	77	99	0
29	2	8030	91	99	0

Table 23. 2006 Grade 4 Writing – Interrater Percent Agreement

Item	Points Possible	Number of Papers Scored	% Exact Agreement	% Adjacent + Exact Agreement	% Non-Adjacent Agreement
Expository Content (COS)	4	15329	84.2	100.0	0.0
Expository Mechanics (CONV)	2	15329	85.2	100.0	0.0
Persuasive Content (COS)	4	15750	87.3	100.0	0.0
Persuasive Mechanics (CONV)	2	15750	89.1	100.0	0.0

Table 24. 2006 Grade 4 Mathematics – Interrater Percent Agreement

Item	Points Possible	Number of Papers Scored	% Exact Agreement	% Adjacent + Exact Agreement	% Non-Adjacent Agreement
2	2	8132	95	100	0
5	2	8130	86	98	2
7	2	8128	91	99	0
9	4	8130	93	99	2
11	2	8130	94	98	2
16	2	8128	85	97	2
18	2	8130	97	99	0
20	4	8132	82	98	2
23	2	8132	91	99	0
28	2	8132	88	100	0
30	2	8132	97	99	0
33	4	8130	85	99	2
35	2	8132	89	99	0
36	2	8132	95	99	0

Table 25. 2006 Grade 4 Validity Paper Agreements – Washington Teacher Scorers & Pearson Scorers

	Item Type or Trait	Validity for Washington Teachers	Validity for Pearson Scorers
Reading	NA		
Writing	COS CONV	COS 79 CONV 96	COS 84 CONV 96
Mathematics	Item # 07 SA Item # 20 ER Item # 36 SA	93 94 97	91 94 95

ADDITIONAL CONDITIONS FOR SCORING WRITING

Although *training* to score Writing is the same as described above, various approaches can be used to evaluate the quality of Writing. For the WASL, a "focused holistic" approach was selected. Focused holistic scoring, or general impression scoring, assesses relative writing fluency and measures the degree to which a writer has connected to the scorer of a paper. When a paper is scored holistically, a scorer considers the overall effectiveness of the piece of writing and assigns a score that reflects the scorer's impression of the paper's overall quality. In a focused holistic approach, the scorer also takes into account all of the elements that make up a successful piece of writing, for example content, organization, style, and mechanics. In the WASL Writing test, Content, Organization, and Style are scored together on a 4-point scale (score points 1-4) and Writing Mechanics are scored on a 3-point scale (score points 0-2).

PART 8: PERFORMANCE OF 2006 GRADE 4 STUDENTS

The summary data presented in Tables 26 to 46 are descriptive of Grade 4 student performance on the 2006 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. Means and standard deviations were calculated relative to the number of students tested, rather than number of students in the population. Table 26 provides the statewide mean scores for Grade 4 students who took the WASL tests in Spring 2006. The column "Maximum Scaled Score" lists the highest reported scaled score points for each of the 2006 tests. Actual calculated maximum scaled score point values are listed in Tables 17 and 18 in Part 5 of this report. The next two columns contain the mean scaled score and scaled score standard deviations for students tested statewide. Table 27 lists the 2006 Grade 4 statewide summary statistics for content strands in each WASL test on a raw score metric.

Table 26. 2006 Grade 4 Means & Standard Deviations (SD) Test Scores

Test	Number Tested	Maximum Scaled Score † or Raw Score *	Mean Scaled Score † or Raw Score *	SD
Reading †	70748	475	414.6	20.5
Writing *	70837	12	8.8	2.0
Mathematics †	70974	550	406.8	36.8

†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.

Table 27. 2006 Grade 4 Raw Test Score Summaries, Percent Students with Strength in Strand

Strand	Number Tested	Points Possible	Raw Score Mean	SD	Percent with Strength in Strand
Reading	70748	42	31.0	6.8	82.0
Literacy Text Comprehension	70748	10	8.5	1.7	79.0
Literacy Text Analyze/Interpret	70748	14	9.5	2.6	79.1
Informational Text Comprehension	70748	6	4.6	1.4	79.0
Informational Text Analyze/Interpret	70748	12	8.3	2.5	78.0
Writing	70837	12	8.8	2.0	61.5
Writing Content, Organization Style	70837	8	5.6	1.3	56.7
Writing Conventions	70837	4	3.2	1.0	75.3
Mathematics	70974	55	35.0	10.8	59.7
Number Sense	70974	6	3.7	1.7	53.6
Measurement	70974	6	3.5	1.4	52.7
Geometric Sense	70974	5	3.4	1.4	52.3
Probability & Statistics	70974	6	4.2	1.5	50.7
Algebraic Sense	70974	6	4.5	1.5	58.3
Solves Problems/ Reasons Logically	70974	13	9.0	3.1	51.9
Communicates Understanding	70974	8	4.0	2.2	63.4
Makes Connections	70974	5	2.6	1.5	52.8

Tables 28 through 36 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 28. 2006 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Males	35961	412.0	20.5
Females	34762	417.3	20.1

Table 29. 2006 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
Alaska Native/Native American	1922	407.0	20.5
Asian	5783	418.4	19.8
African American/Black	4076	406.9	20.8
Latino/Hispanic	10053	404.8	21.6
White/Caucasian	47399	417.3	19.4
Pacific Islander	201	407.4	21.2
Multi-Racial	676	415.2	20.9

Table 30. 2006 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Males	35944	8.4	2.0
Females	34850	9.2	1.8

Table 31. 2006 Grade 4 Writing – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
Alaska Native/Native American	1942	8.0	2.0
Asian	5777	9.4	1.7
African American/Black	4097	8.2	2.1
Latino/Hispanic	10047	8.0	2.1
White/Caucasian	47432	9.0	1.9
Pacific Islander	204	8.7	1.9
Multi-Racial	668	8.9	1.8

Table 32. 2006 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Males	36113	405.7	37.0
Females	34828	407.9	36.5

Table 33. 2006 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
Alaska Native/Native American	1918	391.6	36.4
Asian	5790	417.1	38.9
African American/Black	4076	387.5	34.0
Latino/Hispanic	10107	387.4	34.1
White/Caucasian	47579	412.1	35.1
Pacific Islander	204	391.0	33.7
Multi-Racial	672	403.7	36.0

Table 34. 2006 Grade 4 Reading – Scaled Score Means & Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Read	2884	402.4	18.9
LAP Math	1955	402.9	19.8
Title I Read	9695	408.5	20.1
Title I Math	6746	409.3	20.1
Gifted	2661	433.2	16.0
Section 504	7751	396.5	23.5
Special Ed	7603	396.4	23.6
Migrant	1514	400.7	21.9
ELL/Bilingual	5851	396.2	21.1

Table 35. 2006 Grade 4 Writing – Raw Score Means & Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Read	2877	7.7	2.0
LAP Math	1968	7.7	2.0
Title I Read	9671	8.2	2.0
Title I Math	6741	8.3	2.0
Gifted	2655	10.3	1.3
Section 504	8161	7.1	2.3
Special Ed	8012	7.1	2.3
Migrant	1498	7.7	2.1
ELL/Bilingual	5837	7.4	2.1

Table 36. 2006 Grade 4 Mathematics – Scaled Score Means & Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Read	2899	384.9	31.5
LAP Math	1984	382.9	31.7
Title I Read	9730	394.8	34.7
Title I Math	6793	396.5	34.9
Gifted	2656	454.2	30.7
Section 504	8025	379.2	36.0
Special Ed	7876	379.0	36.0
Migrant	1530	381.4	33.7
ELL/Bilingual	5892	375.8	32.3

PERCENT MEETING STANDARD

Tables 37 through 45 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 each of Tables 37 to 45, the percentage entries are based on the number of students within a particular subgroup or program category. Performance Level 1 "Below Basic" in these tables includes students who attempted the WASL but received no score for unexcused absence, missing booklet, incomplete record, refusal to test, invalidated test, or testing with an out of grade level test. "Not tested" consist of students excluded from testing on the basis of limited English proficiency (LEP), medical condition, excused absence, partial enrollment during the testing window, exemptions due to previously passing tested content, or exemption due to participation in the alternate assessment portfolio (WAAS) or in the Developmentally Appropriate WASL (DAW). In the following tables, "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 37. 2006 Grade 4 Reading – Percent Meeting Standards by Gender

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
All Students	74682	33.1	44.5	14.1	3.9	4.4	3.2
Females	36328	38.2	43.3	12.2	2.9	3.5	2.4
Males	38287	28.4	45.7	15.9	4.9	5.2	4.0

Table 38. 2006 Grade 4 Reading – Percent Meeting Standards by Ethnic Group

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Ethnic Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
Alaska Native/Native American	2076	19.6	45.1	22.7	6.1	6.5	5.7
Asian	6040	39.8	43.3	11.0	2.1	3.8	1.6
African American/Black	4349	19.8	45.2	23.4	6.3	5.3	4.1
Latino/Hispanic	10848	17.8	43.9	23.2	8.0	6.7	4.4
White/Caucasian	49654	37.6	44.8	11.1	2.9	3.6	2.9
Pacific Islander	209	21.1	46.4	23.0	5.7	3.8	1.9
Multi-Racial	704	34.9	43.2	14.1	4.4	3.4	2.4

Table 39. 2006 Grade 4 Writing – Percent Meeting Standards by Gender

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
All Students	74642	16.8	41.6	24.9	12.9	3.8	2.5
Females	36313	22.3	45.0	21.3	8.5	2.9	1.7
Males	38262	11.6	38.4	28.4	17.1	4.5	3.2

Table 40. 2006 Grade 4 Writing – Percent Meeting Standards by Ethnic Group

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Ethnic Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
Alaska Native/Native American	2076	7.7	34.5	31.2	21.9	4.7	3.6
Asian	6040	25.8	46.7	17.1	6.7	3.7	1.4
African American/Black	4347	9.6	37.8	30.2	18.1	4.4	2.8
Latino/Hispanic	10812	7.4	35.1	31.7	19.8	6.0	3.4
White/Caucasian	49652	18.8	43.1	23.7	11.4	3.0	2.3
Pacific Islander	209	11.5	47.4	27.8	11.5	1.9	0.5
Multi-Racial	704	17.0	42.8	27.4	8.5	4.3	2.7

Table 41. 2006 Grade 4 Mathematics – Percent Meeting Standards by Gender

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
All Students	74682	27.2	29.5	20.9	18.3	4.1	2.8
Females	36328	28.0	30.0	21.3	17.3	3.3	2.2
Males	38287	26.4	29.0	20.5	19.3	4.7	3.5

Table 42. 2006 Grade 4 Mathematics – Percent Meeting Standards by Ethnic Group

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Ethnic Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
Alaska Native/Native American	2076	15.3	23.7	23.9	30.8	6.3	5.2
Asian	6040	37.9	28.6	17.4	12.4	3.7	1.5
African American/Black	4349	11.7	22.9	25.8	34.3	5.3	4.0
Latino/Hispanic	10848	11.3	23.4	25.2	33.8	6.2	3.7
White/Caucasian	49654	31.5	31.9	19.8	13.7	3.2	2.5
Pacific Islander	209	13.4	23.4	30.1	30.6	2.4	1.0
Multi-Racial	704	25.3	28.6	21.9	20.9	3.4	2.4

Table 43. 2006 Grade 4 Reading – Percent Meeting Standards by Categorical Program

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Categorical Program	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
LAP Read	3009	11.9	47.4	30.2	6.7	3.8	2.8
LAP Math	2049	14.0	45.0	29.9	6.8	4.2	3.4
Title I Read	10211	22.5	46.2	22.0	4.7	4.6	3.6
Title I Math	7136	23.6	46.6	20.3	4.6	5.0	3.8
Gifted	2669	75.1	23.9	0.7	0.1	0.2	0.0
Section 504	10395	9.2	28.0	25.3	13.9	23.7	22.8
Special Ed	10239	9.1	27.8	25.1	14.0	23.9	23.1
Migrant	1658	13.0	41.1	27.4	10.6	7.9	5.0
ELL/Bilingual	6556	7.8	37.1	32.1	13.0	10.1	3.6

Table 44. 2006 Grade 4 Writing – Percent Meeting Standards by Categorical Program

		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Categorical Program	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
LAP Read	3007	4.5	31.1	36.7	24.1	3.6	2.5
LAP Math	2047	5.2	30.5	35.7	25.5	3.1	2.3
Title I Read	10198	9.8	35.5	32.2	18.2	4.3	3.2
Title I Math	7127	10.5	36.5	30.7	17.9	4.4	3.1
Gifted	2669	47.0	45.0	6.9	0.9	0.2	0.0
Section 504	10385	4.5	18.3	26.8	31.7	18.7	17.8
Special Ed	10229	4.4	18.1	26.7	31.8	19.0	18.0
Migrant	1641	6.0	27.2	35.2	24.3	7.2	4.1
ELL/Bilingual	6523	3.8	26.0	33.8	27.0	9.3	2.7

Table 45. 2006 Grade 4 Mathematics – Percent Meeting Standards by Categorical Program

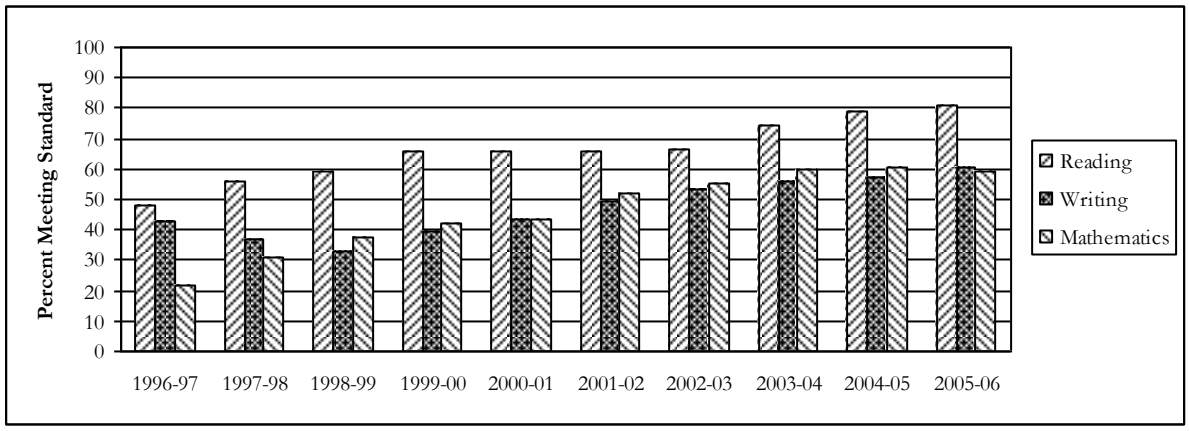
		Meets Standard		Does Not Meet Standard		Percent Not Tested	
Categorical Program	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		Percent Exempt
LAP Read	3009	9.2	21.8	29.6	36.1	3.3	2.1
LAP Math	2049	8.4	20.4	29.9	38.5	2.8	2.0
Title I Read	10211	16.3	26.6	25.6	27.2	4.3	3.3
Title I Math	7136	18.2	27.5	25.2	26.2	4.3	3.2
Gifted	2669	83.2	14.2	1.9	0.3	0.3	0.0
Section 504	10395	7.9	14.4	18.6	38.2	20.9	20.0
Special Ed	10239	7.8	14.4	18.5	38.1	21.2	20.3
Migrant	1658	8.6	20.9	23.2	40.3	7.1	4.3
ELL/Bilingual	6556	5.0	16.6	24.2	44.5	9.7	3.2

Table 46 and Figure 5 illustrate the trend in student performance from 1996-97 to 2005-06 in each content area. These data are based on information from published statewide score reports.

Table 46. Grade 4 Percentage of Students Meeting Standard from 1996-97 through 2005-06

	Administration Year									
	1996-97	**1997-98**	**1998-99**	**1999-00**	**2000-01**	**2001-02**	**2002-03**	**2003-04**	**2004-05**	**2005-06**
Reading	47.9%	55.6%	59.1%	65.8%	66.1%	65.6%	66.7%	74.4%	79.2%	81.2%
Writing	42.8%	36.7%	32.6%	39.4%	43.3%	49.5%	53.6%	55.8%	57.5%	60.6%
Mathematics	21.4%	31.2%	37.3%	41.8%	43.4%	51.8%	55.2%	59.9%	60.6%	59.0%

Figure 5. Grade 4 Results for 1996-97 through 2005-06 by Content Area



MEAN ITEM PERFORMANCE AND ITEM-TEST CORRELATIONS

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 2006 tests.

The data listed in Tables 47 through 49 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.

Table 47. 2006 Grade 4 Writing – Operational Item Statistics

Prompt	Score Type	Score Points Possible	Score Mean	Score-Total Test Correlation
1	Narrative Content, Organization & Style	4	2.9	0.57
	Narrative Writing Conventions	2	1.6	0.56
2	Expository Content, Organization & Style	4	2.7	0.49
	Expository Writing Conventions	2	1.6	0.54

Table 48. 2006 Grade 4 Reading – Operational Item Statistics

Item Number in Test Booklet	Points Possible	Item Mean	Item-Test Correlation	Rasch Item Difficulty
1	1	0.955	0.38	-1.711
2	1	0.901	0.33	-0.937
3	1	0.927	0.30	-1.319
4	2	1.558	0.47	0.470
5	1	0.966	0.27	-2.175
6	2	1.515	0.45	0.609
7	1	0.792	0.37	0.066
8	2	1.597	0.55	0.125
9	1	0.809	0.31	-0.070
10	1	0.815	0.27	-0.087
11	1	0.898	0.36	-0.912
12	1	0.887	0.37	-0.773
13	1	0.895	0.36	-0.821
14	1	0.897	0.42	-0.917
15	2	0.964	0.53	1.690
16	1	0.756	0.33	0.325
17	1	0.812	0.27	-0.008
18	4	2.044	0.57	1.425
19	1	0.970	0.28	-2.273
20	1	0.884	0.32	-0.679
21	1	0.749	0.40	0.331
22	2	1.358	0.47	1.008
23	1	0.755	0.30	0.130
24	2	1.476	0.54	0.606
25	1	0.759	0.45	0.485
26	1	0.768	0.41	0.061
27	1	0.763	0.44	0.153
28	4	2.329	0.61	1.315
29	2	1.167	0.62	1.256

Table 49. 2006 Grade 4 Mathematics – Operational Item Statistics

Item Number in Test Booklet	Points Possible	Item Mean	Item-Test Correlation	Rasch Item Difficulty
1	1	0.910	0.34	-1.851
2	2	1.482	0.49	-0.433
3	1	0.744	0.26	-0.450
4	1	0.762	0.30	-0.574
5	2	0.610	0.48	1.402
6	1	0.783	0.41	-0.800
7	2	1.042	0.57	0.561
8	1	0.692	0.44	-0.314
9	4	3.237	0.57	-0.368
10	1	0.644	0.35	0.052
11	2	1.181	0.55	0.398
12	1	0.518	0.36	0.736
13	1	0.761	0.33	-0.620
14	1	0.656	0.38	0.139
15	1	0.881	0.37	-1.574
16	2	0.585	0.58	1.597
17	1	0.600	0.17	0.247
18	2	0.912	0.62	0.902
19	1	0.767	0.44	-0.706
20	4	2.833	0.57	0.000
21	1	0.783	0.38	-0.736
22	1	0.693	0.37	-0.100
23	2	0.881	0.53	1.086
24	1	0.876	0.43	-1.493
27	1	0.599	0.30	0.446
28	2	0.963	0.59	0.872
29	1	0.578	0.41	0.453
30	2	1.040	0.49	0.691
31	1	0.800	0.36	-0.841
32	1	0.623	0.49	0.133
33	4	2.952	0.58	-0.120
34	1	0.694	0.31	-0.232
35	2	0.911	0.51	0.847
36	2	1.224	0.57	0.349
37	1	0.759	0.45	-0.631

APPENDIX: 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

Edys Quellmalz, Stanford Research Institute

Joseph Ryan, Professor Emeritus, Arizona State University

Catherine Taylor, Associate Professor, University of Washington

Washington State Assessment Advisory Team

Jan Baxter, Director of Assessment, Kelso School District

Charisse Berner, Director of Curriculum Director and Assessment Coordination, Oak Harbor School District

Phil Dommes, Director of Assessment and Evaluation, North Thurston Public Schools

Linda Elman, Director of Research and Evaluation, Central Kitsap School District

Tersea Easley, Assistant Director of Assessment, Tacoma School District

Bev Henderson, Director of Assessment and Staff Development, Kennewick School District

Peter Hendrickson, Assessment Specialist, Everett Public Schools

Feng-Yi Hung, Director of Assessment and Program Evaluation, Clover Park School District

Nancy Katims, Director of Assessment, Research, and Evaluation, Edmonds School District

June Lee, District Assessment Coordinator, Soap Lake School District

Allen Miedema, Information Systems Manager, Northshore School District

Michael Power, Director of Instruction and Assessment, Mercer Island School District

Nancy Skerritt, Assistant Superintendent for Curriculum and Assessment, Tahoma School District

Robert Silverman, Executive Director, Assessment and Accountability, Puyallup School District

Nancy Steers, District Assessment Coordinator, Seattle Public Schools