

Study of the Grade 4 Mathematics Assessment

Final Report



Dr. Terry Bergeson
State Superintendent of
Public Instruction

September 2000

Study of the Grade 4 Mathematics Assessment **Final Report**

Dr. Terry Bergeson
State Superintendent of Public Instruction

Cheryl L. Mayo, Deputy Superintendent
Learning and Teaching

Rosemary Fitton, Assistant Superintendent
Assessment, Research, and Curriculum

Pete Bylsma, Director
Research and Evaluation

September 2000

About This Document

This document can be obtained by placing an order on our Web site (www.k12.wa.us); by writing the Resource Center, Office of Superintendent of Public Instruction, PO Box 47200, Olympia, WA 98504-7200; or by calling the Resource Center toll-free at (888) 595-3276. If requesting more than one copy, contact the Resource Center to determine if printing and shipping charges apply. The contents of this document can be reproduced without permission.

This material is available in alternative format upon request. Contact the Resource Center at (888) 595-3276, TTY (360) 664-3631, or e-mail erickson@ospi.wednet.edu. The Office of Superintendent of Public Instruction complies with all federal and state rules and regulations and does not discriminate on the basis of race, color, national origin, sex, disability, age, or marital status.

For more information about the contents of this document, please contact:

Pete Bylsma, Director
Research and Evaluation
Office of Superintendent of Public Instruction
PO BOX 47200
Olympia, WA 98504-7200
E-mail: pbylsma@ospi.wednet.edu

Acknowledgements

This study was conducted under the supervision of Pete Bylsma, Director of Research and Evaluation. Other staff at the Office of Superintendent of Public Instruction (OSPI) who helped conduct the study were Debora Merle, Lisa Ireland, Bev Neitzel, Mary Ann Stine, Lesley Thompson, and the members of the assessment unit. Catherine Taylor conducted various analyses of the assessment results and prepared technical reports for the 1998 and 1999 assessments. Teresa Baldwin and members of the Washington Mathematics Helping Corps assisted with site visits to selected schools administering the assessment. Staff at 10 schools provided feedback on the assessment and provided access to their classrooms so students could be observed. Staff at 38 schools completed a survey on factors contributing to improvement on the assessment. Finally, many educators contributed valuable comments about the assessment.

Dean Arrasmith, Thomas Tinkler, and Svetlana Beltyukova from the Northwest Regional Educational Laboratory analyzed issues related to the assessment's developmental level. Four independent mathematics experts provided technical assistance and other analyses. These experts were Verna Adams (Washington State University), Stanley Pogrow (University of Arizona), Cindy Walker (University of Washington), and John Woodward (University of Puget Sound). Members of OSPI's national and state technical advisory committees also provided helpful comments related to the analyses.

CONTENTS

Executive Summary	i
Chapter	
1 Introduction	1
Recent Education Reform in Washington	
Overview of State Assessment System	
2 Developing and Scoring the Assessment	5
Designing Test and Item Specifications	
Draft Item Development	
Pilot Test Scoring and Item Analysis	
Final Item Selection	
Setting Standards	
Calculating Scale Scores	
Conclusion	
3 Frequency of Test Items	17
Computation and the Use of Technology	
Analysis of Mathematics Skills Assessed	
4 Assessing the Test’s Difficulty	21
Analyzing Test Item Difficulty	
Results of the NWREL Analysis	
Alignment with the Essential Learning Requirements	
Implications	
5 Performance of Grade 4 Students on the Mathematics WASL	33
Statewide Improvement	
Analysis of Schools Showing the Most Improvement	
Analysis of Schools Receiving Math Helping Corps Assistance	
6 Other Analyses of the Test	44
Estimated Testing Time Per Session	
Test Administration	
Additional Research Needed	
7 Summary and Next Steps	47

Appendix

A	4th Grade Mathematics EALRs, Strands, and Learning Targets	49
B	Detailed Analysis of Test Items	56
C	Types of State Assessments	60
D	Guidelines for Designing 4th Grade Test Items	62
E	Scoring Open-Ended Items	66
F	Statistical Analyses of Test Items	77
G	Further Evidence of Validity	80
H	NWREL Analysis Methods	93
I	Experts Providing Assistance	100
J	Legislative Mandate	102

Abbreviations

CSL	Commission on Student Learning
CTBS	Comprehensive Test of Basic Skills
EALRs	Essential Academic Learning Requirements
ITBS	Iowa Test of Basic Skills
ITED	Iowa Test of Educational Development
LAP	Learning Assistance Program
NCTM	National Council of Teachers of Mathematics
NWREL	Northwest Regional Educational Laboratory
OSPI	Office of Superintendent of Public Instruction
SEM	Standard Error of Measurement
WASL	Washington Assessment of Student Learning

EXECUTIVE SUMMARY

Legislation passed in the early 1990s led to the creation of essential academic learning requirements (EALRs) in mathematics and other subjects. A state test, the Washington Assessment of Student Learning (WASL), was created to measure student progress towards achieving these learning requirements. Legislation required that performance standards on the tests be set at internationally competitive levels. Concerns have surfaced about the difficulty of the mathematics WASL given to 4th grade students. Having a test that is well designed, age-appropriate, and aligned with the EALRs will help educators, policymakers, and parents assess student progress toward meeting the state's academic standards.

The Legislature mandated the Superintendent of Public Instruction to conduct an objective analysis of the 4th grade mathematics WASL. The analysis was to include the percentage of items that (1) require students to use computational skills with and without the use of technology, and (2) measure mathematics communications, problem-solving, and other skills included in the assessment. In addition, the study was to determine the student developmental level required to achieve the 4th grade standard successfully.

The Office of Superintendent of Public Instruction (OSPI) took this opportunity to study the test in even greater detail to identify any aspect of the test that needed to be changed. OSPI conducted various analyses of the 4th grade mathematics WASL administered in 1998, 1999, and 2000. In addition, staff reviewed critiques about the test and visited 10 schools around the state to observe students taking the test and interview teachers and principals about the test. To help ensure a balanced and objective study, OSPI contracted with the Northwest Regional Educational Laboratory (NWREL) to examine the process used to develop the assessment, determine the developmental level required to meet the standard, and identify any aspect of the test's design that needed to be changed. Various independent experts conducted additional analyses and assisted OSPI and NWREL in conducting the study, as required by legislation. Finally, OSPI identified and surveyed schools that showed the greatest level of improvement on the 4th grade mathematics WASL from 1997 to 1999 to determine factors that contributed to the improvement.

SUMMARY OF FINDINGS

The study found that 55 percent of the test items required computation skills. Students could use calculators on half of these items. The percentage of items assessing the skills of the various mathematical strands was evenly distributed.

Various analyses concluded that the level needed to meet the standard was within the developmental capability of well taught, hard working 4th grade students. A few test items were beyond the developmental level of 4th grade students or were not aligned with the EALRs. However, students had ample opportunity to meet the standard by doing well on other items. Analyses of student test results and of selected schools that had dramatic improvement on the test found that the performance of all student groups has improved over time, providing evidence that

all students can meet high standards when given exposure to appropriate mathematics curriculum and instruction.

In addition to these findings, the study found that the processes used to develop the test and set the standard were sound. Moreover, the study identified factors that contributed to higher test scores. The study also found that some items had difficult or unusual formats that may mask students' true cognitive and mathematics abilities. Finally, the study found that the test, as currently constructed, is too long for 4th grade students and is not administered under the same conditions.

OSPI will take various steps to improve the quality of future tests.

BACKGROUND

New tests for mathematics and other subjects were first developed for the 4th grade and were first administered on a voluntary basis in the spring of 1997. Participation was mandatory in 1998. The test uses multiple-choice, short-answer, and extended-response items to allow students to demonstrate their knowledge, skills, and understanding in each part of the EALRs. Although untimed, the tests are considered “standardized,” that is, all students are to respond to the same items, under the same conditions, and during the same three-week period. Guidelines for providing accommodations to students with special needs have been developed to encourage the participation of as many students as possible.

To help teachers, students, and parents understand the assessments, an Example Test and other materials were created. These materials include samples of test items, scoring criteria for the items, and examples of student responses that have been scored. In addition, professional development opportunities have been provided to help school and district staff improve their understanding of the EALRs and effective instructional strategies that will help students reach the standards.

TEST DEVELOPMENT AND SETTING STANDARDS

Washington engaged in an extensive test development process. Several committees comprised of diverse populations from across the state helped create the EALRs for the various mathematical content areas (“strands”), design the content of the test, draft items given on a pilot test, ensure that items were free of potentially offensive and biased content, and review the results of the pilot test.

After the test was administered in 1997, a diverse standard setting committee was created to set the level that students needed to achieve in order to meet the state standard. Most members of the committee either had direct experience with 4th grade students or with curriculum materials relevant for 4th graders. Setting the level required extensive analysis and consultation. In determining the level, the committee was guided by what they believed a “well taught, hard working student” should be able to do in the spring of the 4th grade. This thorough “expert judgment” process ensured that the standards set for proficiency would have careful scrutiny from a broad range of constituents of education. Committee members had significant input from their peers and ample opportunities to discuss their diverse opinions on the standards.

After the committee set a “cut score” that students had to achieve in order to meet the standard, they set other levels to indicate partial achievement of the standard and achievement above the standard. Statistical equating methods are used to ensure that these levels are held constant over time.

Independent experts found the test development and standard setting processes were sound, well documented, and met the standards set forth in the *Standards for Educational and Psychological Testing*. The test development process is used across the country, and the standard setting process is used to set standards in other states.

FREQUENCY OF TEST ITEMS

Each mathematics WASL is divided into two parts. The first part has 20 items and allows the use of tools (e.g., calculators, rulers), although these tools are not needed to complete every item. The second part also has 20 items but does not allow these tools. Each of the 40 items on the test assesses proficiency on one “strand” of the 4th grade mathematics EALRs. However, concepts from other strands are almost always included in each item.

Analyses of the items on the tests administered in 1998, 1999, and 2000 found that on average, 55 percent of all the items required computation skills, and calculators were allowed on half (27.5%) of these. Other analyses found that the percentage of items that assessed communications, problem solving, and other mathematics strands was well balanced across the nine strands and conformed to the test specifications. Table 1 shows the extent to which each strand was assessed.

Table 1: Analysis of Items Assessing Mathematics Strands

Essential Learning Strands	<i>Number of items assessing strand</i>			<i>Percent of items assessing strand</i>		
	1998	1999	2000	1998	1999	2000
Concepts and Procedures	26	26	25	65%	65%	62.5%
Number sense	6	6	5	15.0%	15.0%	12.5%
Measurement	5	5	5	12.5%	12.5%	12.5%
Geometric sense	5	5	5	12.5%	12.5%	12.5%
Probability/statistics	5	5	5	12.5%	12.5%	12.5%
Algebraic sense	5	5	5	12.5%	12.5%	12.5%
Processes	14	14	15	35%	35%	37.5%
Solving problems	4	3	3	10.0%	7.5%	7.5%
Reasoning logically	4	4	3	10.0%	10.0%	7.5%
Communicating	3	3	4	7.5%	7.5%	10.0%
Making connections	3	4	5	7.5%	10.0%	12.5%

Nearly every item on the test also contained concepts from at least one other strand, a practice found in similar tests. Some items included concepts related to four or more strands. Concepts related to “number sense” and “communications” were found most frequently, while some strands (i.e., algebraic sense, geometric sense, and making connections) were not present very often.

ASSESSING THE TEST’S DIFFICULTY

The standard setting committee established a standard based on what a well taught, hard working student should be able to do in the last trimester of grade 4. NWREL and various mathematics experts analyzed the individual test items and the results of the 4th grade mathematics WASL to determine if the items and the test as a whole were inappropriate in any way.

These independent sources came to the same conclusion—the level needed to meet the standard was within the developmental capability of well taught, hard working 4th grade students. Analyses of schools that have shown dramatic improvement on the mathematics WASL found that students from all ethnic/racial and socioeconomic groups were making substantial progress toward the standard, and that the average student score for some groups in these schools had met the standard. Such progress is evidence that the test is developmentally appropriate, that all 4th grade students are capable of meeting high standards when given exposure to appropriate curriculum and instruction.

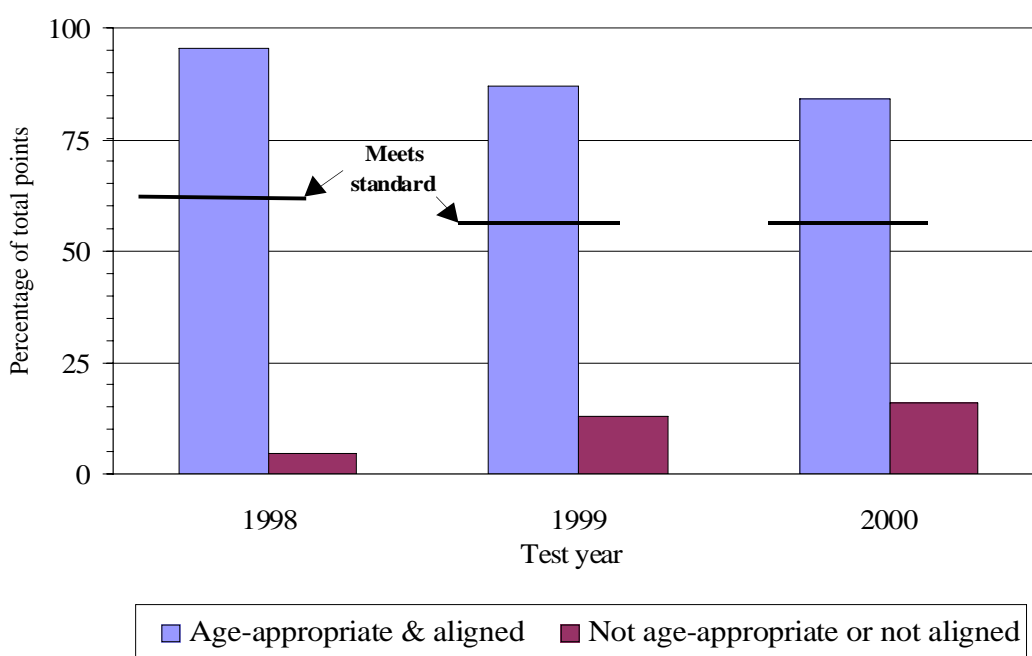
- On the other hand, the analyses by NWREL and the experts most familiar with the EALRs and WASL found that a few items on each test were beyond the developmental level of 4th grade students. The number of such items depended on the type of analysis. NWREL analyzed the level of format, cognitive, and mathematical complexity of each item as well as the scores of individual students for each item on the 1998 and 1999 tests. The experts analyzed the items’ alignment with the 4th grade EALRs.
- NWREL’s analysis found that five of the 80 items (6 percent) were very difficult and potentially inappropriate. NWREL also found a few items that were very easy. Items that are either very difficult or very easy provide little information about a student’s ability to meet the standard. However, the test is intended to provide information about student ability across a range of item difficulty, so having some very easy and very difficult items is appropriate. NWREL also found some items had formats that were unusual or difficult that could mask students’ cognitive and mathematical abilities.¹
- The experts found that 10 of the 120 items (8 percent) on the tests from 1998, 1999, and 2000 were not aligned with the 4th grade EALRs. Seven of these 10 items were not aligned with one EALR benchmark—they required students to “create a plan” rather than “follow a plan.” The remaining three items that were not aligned were found to be above the developmental level of 4th grade students. An analysis of items on the Example Test also found some items were not aligned with 4th grade benchmarks.

¹ Format complexity includes sources of test item difficulty resulting from the layout of the items (e.g., directions, question and problem formats, response spaces and options, how information is presented in tables, graphs, and illustrations).

The “create a plan” items were among the more difficult items on the test, although over half the 4th grade students still received partial or full credit on some of these items. This suggests that “create a plan” items are within the developmental capability of 4th grade students. Nevertheless, if teachers do not prepare their students to create a plan because such a task is not one of the benchmarks, their students are not likely to do well on such items.

When taken together, these findings suggest that more than 90 percent of the items were appropriate for inclusion on the test. Since only a few items on each test were beyond the developmental level of a well taught, hard working student, there was still ample opportunity for students to meet the standard by doing well on the other items (see Figure 1).

Figure 1: Students Could Meet Standard By Doing Well on Other Items



In light of these findings, OSPI has begun work with its contractors and technical advisors to ensure that future assessments and the materials designed to help students prepare for the WASL contain only items that are within the developmental level of 4th grade students and aligned with the EALR benchmarks. In addition, OSPI will revise items with format complexity problems.

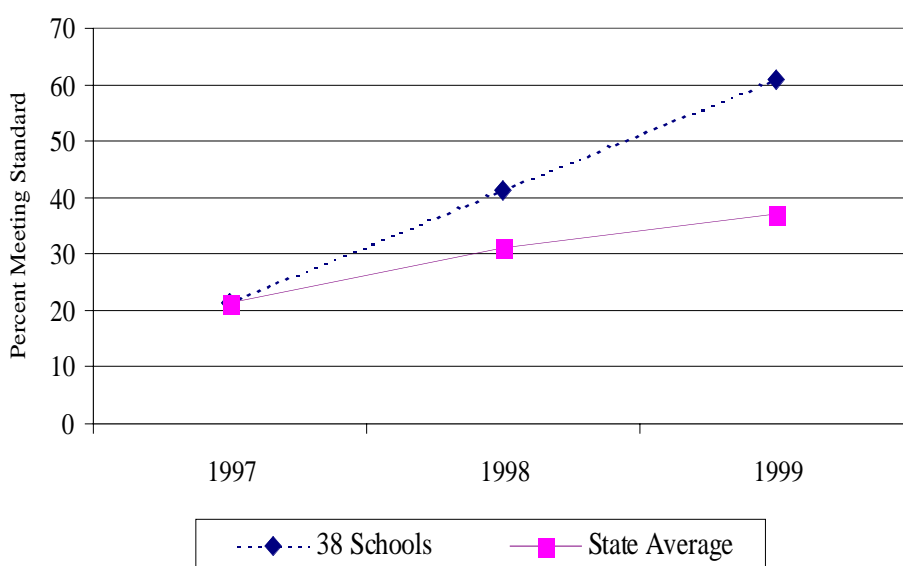
PERFORMANCE OF 4TH GRADE STUDENTS

The performance of 4th grade students on the mathematics WASL has improved over time. This provides evidence that the developmental level needed to meet the standard is appropriate to 4th graders.² Analyses of test results found that improvement has occurred statewide in most mathematical strands and among students in each gender, ethnic/racial group, and program.

² If more 4th grade students are able to correctly respond to “hard” items over time and meet the standard, there is evidence that the test is sensitive to changes in curriculum, instruction, and student motivation and is therefore appropriate for a “well taught, hard working” student.

Some schools have shown dramatic improvement. The percentage of students meeting the standard has nearly tripled among the 38 schools that had the most improvement over the 1997–1999 period (see Figure 2). These schools represented the full range of socioeconomic status and are found in different parts of the state. Staff in these schools reported many reasons for their improved performance. The most important reasons were increased and improved mathematics instruction. Students in the 13 elementary schools receiving technical assistance from the Math Helping Corps performed much better than previous groups of 4th grade students in those schools and students in similar schools. Thus, students can improve mathematics performance when exposed to better mathematics curriculum and instruction.

Figure 2: Results of Grade 4 Mathematics WASL, State and 38 Most Improved Schools



OTHER ANALYSES OF THE TEST

OSPI investigated several more issues that have been raised about the test—the amount of time it takes for students to complete the test and how the test is administered.

Testing Time

The parts of the various WASL tests are not timed. Students have as much time as they need to work on the tests, and professional judgment should determine when a student is no longer “productively engaged.” The estimated testing time required for most students to complete each part of the test is provided to school staff for planning purposes.

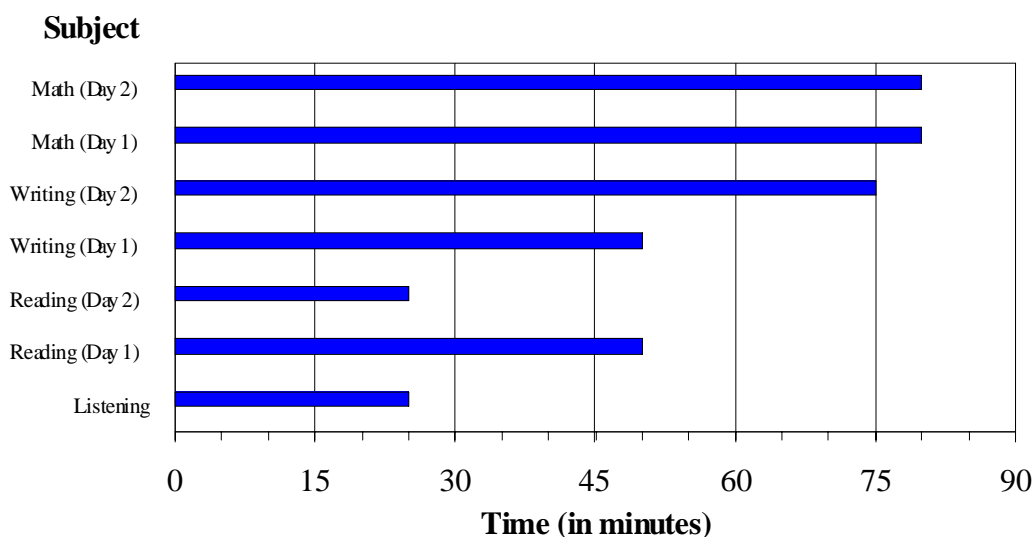
Figure 3 shows that the estimated time to complete the mathematics sessions of the WASL is the longest of all the tested subjects (two 80-minute sessions).³ Some schools impose a time limit on

³ Additional time is required to distribute and collect materials and to cover the directions for taking the test.

the students, even though the test is not to be timed. Observations and anecdotal reports found that the actual time needed by many students is longer than the estimated times.

Nearly all the experts OSPI consulted about the assessment believed that the mathematics test, as currently designed, is too long for 4th grade students. Thus, OSPI and its contractors will consider ways to make the testing time more appropriate for 4th graders, including breaking the test into smaller periods of time (e.g., having three or four shorter sessions rather than two 80-minute sessions) and possibly placing limits on the amount of time students are given to take the test.

Figure 3: Estimated Testing Time Per Session, Grade 4 WASL Subjects



Test Administration

In addition to variations in the amount of time given to take the test, other issues have been raised that suggest that the test may not be administered under “standardized” conditions. Some educators said that they lacked clear guidance about testing conditions, such as the types of materials that can be left on the walls of the classroom and the kinds of guidance they could provide students during the test. Some teachers provide breaks and snacks, while others do not. Some teachers allow students to leave the room when they are finished, while others require all students to stay until all students are finished. Moreover, schools have the discretion to administer the tests in whatever order they wish (e.g., mathematics can come before or after the reading WASL) and may administer more than one test in a single day. OSPI will improve the guidelines for administering the test in an age-appropriate and a more standardized manner.

ADDITIONAL RESEARCH NEEDED

Time and resources did not permit analyses of other issues that have been raised about the test. Research needs to be conducted in several areas.

- More information is needed to determine the amount of time most students need to complete the mathematics WASL and the tests of other subjects. When this information is available, adjustments can be made in the test administration procedures.
- OSPI did not study how individual students approach and respond to individual test items. Having students “think aloud” while taking the test provides insights into the cognitive processes students use when approaching a test item.
- More analysis needs to be conducted on the readability of the mathematics test. Teachers on the test development committees reviewed the vocabulary and overall reading level of test items to ensure that they were below grade level. However, some have expressed concerns that the reading level of the test is too difficult, particularly for students whose primary language is not English. If the reading level is too high, the test may reflect reading ability rather than mathematical ability. Preliminary analyses suggest that the reading level is at grade 4, although this level may be inflated because current readability models are not well suited to analyzing mathematics items. When better models are available, more accurate readability analyses can be conducted.

CONCLUSION

This study found that the developmental level of the grade 4 mathematics WASL is appropriate for Washington students—the standard set for the test is attainable for hard working and well taught students in the 4th grade. The study also identified some changes that need to be made to improve the test. OSPI is committed to providing state assessments that help determine the extent to which students meet the essential academic learning requirements. Thus, OSPI will take the necessary steps to improve the assessment to enhance the validity of the WASL scores.

Chapter 1

INTRODUCTION

Recent education reform efforts in Washington State have their roots in legislation passed in the early 1990s. The reforms included the creation of essential academic learning requirements (EALRs) in various subjects and a series of state assessments, known as the Washington Assessment of Student Learning (WASL). These assessments measure student progress towards achieving the requirements. Legislation required that performance standards on the assessment be set at internationally competitive levels. Thus, students have to perform at a high level on these assessments to meet these high standards.

Some believe the assessments may be too difficult and may not be aligned with the EALRs. The mathematics WASL given to students in grade 4 has been singled out by some educators and others as a particularly difficult and possibly inappropriate test. Having a test that is well designed, age-appropriate, and aligned with the EALRs will help educators, policymakers, and parents assess student progress toward meeting the state's academic standards.

The Legislature mandated the Superintendent of Public Instruction to conduct an objective analysis of the 4th grade mathematics WASL (see Appendix I). The analysis was to include the percentage of items that (1) require students to use computational skills with and without the use of technology, (2) measure mathematics communications and problem-solving skills, and (3) measure other skills included in the assessment. In addition, the study was to determine the student developmental level required to achieve the 4th grade standard successfully.

The Office of Superintendent of Public Instruction (OSPI) took this opportunity to study the test in even greater detail to identify any aspect of the test that needs to be changed. OSPI conducted various analyses of the 4th grade mathematics WASL administered in 1998, 1999, and 2000. In addition, staff reviewed critiques about the test and visited 10 schools around the state to observe students taking the test and interview teachers and principals about the test. To help ensure a balanced and objective study, OSPI contracted with the Northwest Regional Educational Laboratory (NWREL) to examine the process used to develop the assessment, determine the developmental level required to meet the standard, and identify any aspect of the test's design that needed to be changed. Various independent experts conducted additional analyses and assisted OSPI and NWREL in conducting the study, as required by legislation.⁴ Finally, OSPI identified

⁴ OSPI consulted with independent mathematics experts when conducting the study. In addition to the experts working for NWREL, four experts assisted OSPI and NWREL in conducting the study: Verna Adams (Washington State University), Stanley Pogrow (University of Arizona), Cindy Walker (University of Washington), and John Woodward (University of Puget Sound). The members of OSPI's national Technical Advisory Committee were also consulted. These advisors are Peter Behuniak (Connecticut Department of Education), Robert Linn (University of Colorado),

and surveyed schools that showed the greatest level of improvement on the mathematics WASL from 1997 to 1999 to determine factors that contributed to the improvement.

RECENT EDUCATION REFORMS IN WASHINGTON

Washington began a comprehensive effort to improve teaching and learning in the early 1990s. The Legislature passed Engrossed Substitute House Bill 1209 in 1993, noting that “student achievement in Washington must be improved to keep pace with societal changes, changes in the workplace, and an increasingly competitive international economy.” The Legislature created the Commission on Student Learning (CSL)⁵ and required the state to establish academic standards at an internationally competitive level.

The CSL had three major tasks:

- Establish 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 a state 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 CSL adopted EALRs for mathematics, reading, writing, and communications in 1995 and revised them in 1997. Performance “standards” were established at internationally competitive levels for grades 4, 7, and 10, as required by legislation. EALRs for science, social studies, health/fitness, and the arts were adopted in 1996 and revised in 1997.⁶ (See Appendix A for the EALRs and related benchmarks for 4th grade mathematics.)

OVERVIEW OF STATE ASSESSMENT SYSTEM

To determine the extent to which students achieved the knowledge and skills defined by the EALRs, the CSL developed the WASL to be administered statewide.⁷ The WASL is a set of **criterion-referenced assessments**. Students need to achieve a certain score on these assessments to “meet the standard.” In Washington, the score required to meet the standard is a reflection of what a “well taught, hard working student” should know and be able to do.

Assessments for mathematics, reading, writing, and listening were developed first for grade 4 and were administered on a voluntary basis for the first time in the spring of 1997, with mandatory

William Mehrens (Michigan State University), Joseph Ryan (Arizona State University), and Kenneth Sirotnick (University of Washington).

⁵ The duties of the Commission were transferred to OSPI and the Academic Achievement and Accountability Commission in July 1999. See SSB 5418 (1999) and RCW 28A.655.

⁶ Performance “benchmarks” for science were established at grades 5, 8, and 10.

⁷ The state also administers “norm-referenced” assessments. See Appendix C for a description of the different types of tests administered by the state.

participation in 1998. Tests for these subjects were then developed for grades 7 and 10.⁸ Other subjects will be tested in the future in other grades.

These tests require students to both select and create answers to demonstrate their knowledge, skills, and understanding in each of the EALRs. The tests use multiple-choice, short-answer, and extended-response items. The tests are considered “standardized”—all students are to respond to the same items, under the same conditions, and during the same three-week period in the spring. The tests are also untimed (i.e., students 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 need students include those in special education programs and with Section 504 plans, English language learners (ESL/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.⁹

To help teachers, students, and parents understand the assessments, an Example Test and an Assessment Sampler were created for each of the grade 4, 7, and 10 assessments. These materials include samples of test items, scoring criteria for the items, and examples of student responses that have been scored. OSPI also provides an interactive CD-ROM system called NCS Mentor for Washington for teachers and students as another means to review the EALRs and practice scoring student responses to items like those contained on the actual tests.

The WASL is not the only state-administered test. **Norm-referenced tests** are administered in the grades just prior to the grades in which the WASL is administered. The Iowa Test of Basic Skills (ITBS) is given in grades 3 and 6 and the Iowa Test of Educational Development (ITED) is given in grade 9. The primary purpose of norm-referenced tests is to make comparisons between students, schools, and districts. Items on such tests vary in difficulty so that even the most gifted students may find that some of the items are challenging and students who have difficulty in school may respond correctly to some items. Scores can be compared to the performances of a norm-group (i.e., a group of students of similar age and grade), either locally or nationally.

Table 1-1 shows the assessment schedule of both WASL and norm-referenced tests for Washington state students.

⁸ Assessments for grade 7 were administered on a voluntary basis in the spring of 1998. The grade 10 assessments were pilot-tested at this time and were administered in the spring of 1999. Participation in the grade 7 and 10 assessments is voluntary until 2001.

⁹ See *Guidelines for Participation and Testing Accommodations for Special Populations on the Washington Assessment of Student Learning (WASL)*, Superintendent of Public Instruction, June 2000. See OSPI’s website at www.k12.wa.us/specialed/document.asp.

Table 1-1: Planned Assessment Schedule, 1997–2007

Subjects	Test	Grade Level	Available for Voluntary Use	Required
Reading, Writing, Listening, Mathematics	WASL	Grade 4	Spring 1997	Spring 1998
	WASL	Grade 7	Spring 1998	Spring 2001
	WASL	Grade 10	Spring 1999	Spring 2001
Reading, Mathematics	ITBS	Grade 3	—	Spring 1999
	ITBS	Grade 6	—	Spring 1999
Reading, Mathematics, Language Arts	ITED	Grade 9	—	Spring 1999
Science	WASL	Grade 5	Spring 2002	Spring 2005
	WASL	Grade 8	Spring 2000	Spring 2001
	WASL	Grade 10	Spring 2000	Spring 2001
Social Studies	WASL	Grade 5	Spring 2003	Spring 2006
	WASL	Grade 8	Spring 2003	Spring 2006
	WASL	Grade 10	Spring 2003	Spring 2006
Arts	WASL	Grade 5	Spring 2004	Spring 2008
	WASL	Grade 8	Spring 2004	Spring 2007
	WASL	Grade 10	Spring 2004	Spring 2007
Health and Fitness	WASL	Grade 5	Spring 2004	Spring 2008
	WASL	Grade 8	Spring 2004	Spring 2007
	WASL	Grade 10	Spring 2004	Spring 2007

Statewide tests are not the only part of the assessment system. **Classroom-based assessments** provide information from oral interviews and presentations, work products, experiments and projects, or exhibitions 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, middle, and transition 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, sample rating scales, and generic protocols for oral communications and personal interviews. The tool kits also provide content frameworks to assist teachers, at grades K–10, in relating their classroom learning goals and instruction to the EALRs.

Professional development is still another component of the state assessment system. For students to meet the EALRs and do well on the WASL, teachers and administrators need ongoing and comprehensive support and professional training to improve their understanding of the EALRs, the characteristics of sound assessments, and effective instructional strategies that will help students reach the standards. CSL established 15 Learning and Assessment Centers across the state, most of which are managed through Washington’s nine Educational Service Districts, to provide professional development and support to assist school and district staff. OSPI provides a host of other training activities, including summer institutes for teachers and administrators.

Chapter 2

DEVELOPING AND SCORING THE ASSESSMENT

Until recently, Washington has relied mainly on norm-referenced assessments to determine how well students were learning. Such tests used multiple-choice items that are easy to score, resulting in lower costs and faster feedback. They also made comparisons between states relatively easy. The education reform movement created a demand for a different type of test—criterion-referenced assessments that were related to standards.

Creating these assessments is a complicated process because it first requires criteria to be established to define what students should know and be able to do. In Washington, academic standards that were set at an internationally competitive level were developed by the Commission on Student Learning (CSL) through an interactive and collaborative process involving teachers and other educators, parents, and community leaders from around the state. The Commission adopted EALRs for mathematics in 1995 and revised them in 1997.

Once academic standards were in place, a test was designed to assess the extent to which students meet the standards. Open-ended items were needed in addition to multiple choice items so students can show what they know and what they are able to do. Such items require more time and resources to score. This chapter provides a brief overview of the process used to develop the mathematics WASL for grade 4, how the performance levels were established, and how the assessment is scored.

DESIGNING TEST AND ITEM SPECIFICATIONS

To develop an assessment related to the content of the EALRs, classroom teachers and curriculum specialists from across Washington were selected to be on a “content committee.” Most of the committee members had teaching experience at or near the tested grades and in the content areas that were to be assessed (e.g., mathematics). In late 1995, this committee worked with CSL staff and with content and assessment specialists from the Riverside Publishing Company (one of CSL’s test development contractors) to define test and item specifications consistent with the EALRs. This required coming to an agreement about the meaning and interpretation of the EALRs as well as what aspects of the EALRs could be assessed on a state-level test. These test and item specifications are also useful for teachers in developing instructional practices and for administrators in reviewing instructional programs.¹⁰

¹⁰ The test and item specifications can be obtained through OSPI’s website (www.k12.wa.us).

The *test specifications* define how the test is to be constructed. This includes the kinds and number of items on the assessment, the blueprint or physical layout of the assessment, the estimated amount of time to be provided, and the scores to be generated once the test is administered. In addition, the test specifications are the basis for developing equivalent test forms in subsequent years as well as creating new items to supplement the item pool. The 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

Once the test specifications were developed, the next step was to develop *item specifications*. Item specifications provide sufficient detail, including sample items, to direct item writers in the development of appropriate test items for each assessment strand. Separate specifications were produced for the three different item types—multiple-choice, short-answer and extended-response. (See Appendix D for more information about how items are designed.)

Organization of the Test

The test specifications create a framework for constructing the entire test. For the 4th grade mathematics WASL, each test contains 40 items worth a total of 62 points (see Table 2-1). The test is designed to be administered in two sessions. Each session contains 20 items and has a mixture of multiple-choice, short-answer, and extended-response items.

- **Multiple-choice items** Students have three to four possible responses to choose from (the correct answer and at least two distracters). The test contains 24 multiple-choice items, each worth one point. These items are machine scored.
- **Short-answer items (including enhanced multiple-choice¹¹)** Students construct a short response. For example, a student may be asked to write a sentence or equation, complete a chart, draw a picture, or perform a calculation. The test contains 13 short-answer items worth two points each. Short-answer items are hand-scored by well-trained professional scorers.
- **Extended-response items** Students construct a longer response than is required for short-answer items. For example, students may be required to create a graph showing the appropriate data, with labeled axes and a title; create and/or extend tables, diagrams, or pictures; or provide a lengthy written explanation or a written explanation with number sentences pictures, and/or diagrams. The test has three extended-response items worth four points each. Like short-answer items, extended-response items are hand-scored by well-trained professional scorers.

¹¹ Enhanced multiple-choice items require student to select from a list of possible responses and explain their reason(s) for choosing that response.

Table 2-1: Test Item Types and Score Point Values for Mathematics Test Forms

Item Type	Number of Items Per Test Form	Total Point Value	Percent of Total Score Points
Multiple-choice (1 point)	24	24	39
Short-answer (2 points)	13	26	42
Extended-response (4 points)	3	12	19
Total	40	62	100

Each test form contains a variety of items so that all strands are addressed (see Table 2-2). Thus, each of the two parts of the test consists of a mix of items addressing content and process strands. Each strand includes at least one short-answer or one extended-response item. The two parts of the test are constructed to separate the items on which tools (such as rulers or calculators) must *not* be used from the items for which tools are encouraged or possibly required.

Table 2-2: Item Distribution by Strand and Item Type

Strands	Multiple-choice	Short-answer	Extended-response	Number of points
Concepts and Procedures				
Number sense	3-6	1-2	0	5-9
Measurement concepts	3-6	1-2	0	5-8
Geometric sense	3-6	1-2	0	4-7
Probability statistics	3-6	1-2	0	5-8
Algebraic sense	3-6	1-2	0	4-7
Process Strands				
Solving problems	0-2	1-2	1-2	6-12
Reasoning logically	0-2	1-4	0-1	6-12
Communicating understanding	0-2	1-4	0-1	6-12
Making connections	0-2	1-4	0	4-12
Total Number of Items	24	13	3	40
Total Number of Points	24	26	12	62

DRAFT ITEM DEVELOPMENT

Once the test and item specifications were completed and reviewed by the content committee, the contractor's item writers prepared sample items and scoring criteria to these specifications. The content committee then reviewed the items and scoring criteria to assure that the item writers had followed the specifications. When necessary, items were revised to ensure that they measured the EALRs both accurately and comprehensively. This process occurred in late 1995.

When the content committee was satisfied that the sample items and scoring criteria were appropriate, the item writers then produced items for pilot testing. Each test item was coded by content according to particular EALR strands and item types (multiple-choice, short-answer, extended-response). Draft test items were presented to the committee for final review in the exact

format they were to appear on the pilot test forms (including graphics, art work, and location on pages).

The committee reviewed each draft item in early 1996, focusing on its fit to the item specifications, the EALRs, and the appropriateness of item content. For all short-answer and extended-response items, the proposed scoring guidelines (rubrics) were also reviewed. The committee had three options with each item: approve the item (and scoring guidelines) as presented, recommend changes or actually edit the item (or scoring guidelines) to improve the item’s “fit” to the EALRs and specifications, or eliminate the item from use in the assessment.

A separate fairness review committee, comprised of individuals reflective of Washington’s diversity, also reviewed each item. This committee was to identify language or content that might be inappropriate or offensive to students, parents, or communities, items that might contain “stereotypic” or biased references to gender, ethnicity, or culture, and items that might provide an advantage to a particular group of students. This committee reviewed each item and accepted, edited, or rejected it for use on the pilot assessment.

Thus, to be included on the pilot assessment, every item was reviewed and approved by both the content committee and the fairness review committee. Approved items were to:

- Be appropriate measures of the intended content.
- Be appropriate in difficulty for the grade level of the examinees.
- Have only one correct or best answer for each multiple-choice item.
- Have appropriate and complete scoring guidelines for the open response items.
- Be free from content that might disadvantage some students for reasons unrelated to the concept or skill being tested.

PILOT TEST SCORING AND ITEM ANALYSIS

Items approved by the content and fairness committees were then assembled into various pilot test forms and administered to carefully selected representative samples of students across the state. All schools in the state of Washington were invited to participate in the pilot test, and 85 percent of the 4th graders took part in the pilot test in May 1996. Test forms were randomly distributed in order to ensure that each test form was administered in districts with high populations of ethnic minority students. Each test form was given to at least 1,000 students.

Scoring criteria for all open-ended items focus on the clear communication of mathematical ideas, information, and solutions. The conventions of writing (sentence structure, word choice and usage, grammar, spelling, and mechanics) are disregarded, as long as they do not interfere with communicating the response. (See Appendix E for more information about the scoring of open-ended items.)

The items that were developed and pilot-tested created a “pool” of hundreds of items. This allowed the creation of new forms of the assessment each year by sampling from the pool. Having a large pool of items provides the opportunity to vary the kinds of items from year to year so that a particular item format (e.g. multiple-choice, short-answer, or extended-response) is not always associated with the same EALRs. Statistical “equating” procedures are used to maintain the same

performance standard from year to year and to allow for longitudinal comparisons across years even though different items are used.

Following the pilot assessment, student responses were scored by applying the scoring criteria approved by the content committee. A variety of statistical analyses were then employed to determine the effectiveness of the items and check for item bias that may have been missed by the earlier reviews (see Appendix F).

FINAL ITEM SELECTION

After the statistical analyses were completed for the WASL, the content and fairness committees reviewed the results and made the final determination about item quality and appropriateness based on the pilot test data. Items and scoring rules were reviewed again for fit to the EALRs. The fairness committee reviewed the bias data to determine whether content or language might have resulted in large bias statistics. During these reviews, items were either accepted or rejected for the final pool of items.

Once these reviews were completed, the final pool of items was used to develop an “operational” WASL test form. These are the tests administered each year to monitor progress of schools and districts in helping students achieve the EALRs. Each operational form is developed by selecting items from the large pool of items tested in the 1996 item tryouts and approved by the content and fairness committees.

SETTING STANDARDS

Following the administration of the first operational grade 4 assessment in the spring of 1997, the tests were scored for all participating students. A standard setting committee was then created during the summer of 1997 to establish the performance levels related to student achievement of the EALRs. The standard setting committee was composed of teachers, curriculum specialists in the relevant subject area, school administrators, parents, and community members. All standard setting committee members had direct experience with 4th graders or with the curriculum materials relevant for 4th graders.

This committee determined the level of performance on the assessments that would be required for students to “meet the standard” on the EALRs. In determining the level, *the committee was guided by what they believed a “well taught, hard working student” should be able to do in the spring of the 4th grade.* In addition, “progress categories” above and below the standard were established to show growth over time as well as to give students and parents an indication of how close a student’s performance is from the standard. School and district performance on the assessments is reported in terms of the percentage of students meeting the standard and the percentage of students in each of the progress categories.

The committee used standard setting procedure described below to set the performance standards on the 4th grade mathematics assessments. The procedure has been applied successfully in other large-scale assessment programs and was reviewed and approved by CSL and a national technical advisory committee composed of recognized measurement professionals. Essentially, setting standards for student performance on the WASL was a systematic, judgmental process aimed at

establishing a consensus, among knowledgeable people, about what 4th grade students should know and be able to do.

Standard Setting Procedures

Each of the “judges” on the standard setting committee was required to take the operational test just as the students experienced it. The judges also reviewed scoring guides for the open-ended items and examples of student responses anchoring each item’s score points.

Next, each standard setting judge received a complete set of the items ordered by difficulty from easiest to hardest, rather than in the order they appeared in the students’ test booklets. Data from the spring 1997 operational assessment was used to establish item difficulties. The first item in the judges’ ordered booklet was the easiest item on the test (i.e., the one the highest number of students answered correctly). The last item in the judges’ ordered booklet was the hardest item on the test (i.e., the one the fewest number of students answered correctly). Multiple-choice items appeared only once in the ordered booklet. Two- and four-point items appeared two or four times, according to the difficulty of achieving each score point.

In small groups, the judges examined the items in the ordered booklet one at a time until all items (and their scoring rubrics) were examined. As judges examined each item, they were asked to consider what each item was measuring and what made each item more difficult than the items that preceded it. As judges proceeded through the ordered item booklets, trained table leaders encouraged them to observe the increase in the complexity of the items and to note the increase in knowledge, skills, and abilities required to answer the items.

At the conclusion of this first round of review of the ordered booklets, judges were asked to make an individual decision about where to place a “flag” at “meets standard.” Each flag was placed in the ordered item booklet according to the individual judge’s expectation of what students who are performing at standard should know and be able to do. For example, each judge placed his or her “meets standard” flag at a location in the booklet such that if a student is able to respond correctly to the items that precede the flag (with at least 2/3 likelihood of success), then the student has demonstrated sufficient knowledge, skills, and abilities to infer that the student is meeting the standard.¹² Judges were asked to insert two additional flags: one at “exceeds standard” and one between “near standard” (partially proficient) and “low” (minimal proficiency). In this way, progress toward or beyond standards could also be identified.

Because not all judges set their flags in the same locations, a second round of review involved each judge sharing and discussing the locations at which his or her flags were placed. In small groups, the judges discussed differences in expectations as indicated by their different flag placements. When productive discussion of these items was completed, judges were then asked to reevaluate their own initial flag locations in light of the small group discussion. Judges could decide to agree on a common flag placement during this second round. After judges had made their second round flag placements for “meets standard,” the process was repeated for the other two cut-points—the below standard and the above standard locations.

¹² For multiple-choice items, this means the student who “meets standard” should be likely to know the correct response. For short-answer or extended-response items, this means the student who “meets standard” should be likely to achieve at least that score point.

A third round of review brought the small groups back together as a large group to share and discuss each small group’s flag placements. Large group discussion now focused on the items between the first and last flags for each performance level. Following the large group discussion, judges were asked to make a new (or reconfirm their former) flag placements. A fourth and final round of review consisted of sharing with the large group these small group results. Individual judges were then asked to make their final post-it flag placements, which were compiled to establish the final standard and other performance levels for each content area.

The following are descriptions of the performance levels established for the WASL.

- Level 4 Above Standard** This level represents superior performance, notably above that required for meeting the standard at grade 4.
- Level 3 MEETS STANDARD** 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. *“Meets Standard” reflects what a well taught, hard working student should know and be able to do.*
- Level 2 Below Standard** This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 4.
- Level 1 Well Below Standard** This level denotes little or no demonstration of the required knowledge and skills that are fundamental for meeting the standard at grade 4.

This thorough “expert judgment” process ensured that the standards set for proficiency on the WASL would have careful scrutiny from a broad range of constituents of education. The judges had significant input from their peers and sufficient opportunities for discussion about their diverse opinions on the standards.

CALCULATING SCALE SCORES

Following the standard setting process, a linear conversion was used to transform raw test scores (i.e., the number correct) to whole number “scale scores.” For all tests, the raw score identified as “meets standard” was converted to a WASL scale score of 400. The rest of the raw scores were converted to the whole number scale using the linear conversion equations and Levels 1, 2, and 4 were assigned.

- _ Scale scores below 375 are in the “Below Standard, Level 1” category.
- _ Scale scores of 375 to 399 are in the “Below Standard, Level 2” category.
- _ Scale scores of 400 to 421 are in the “Meets Standard, Level 3” category.
- _ Scale scores of 422 or higher are in the “Above Standard, Level 4” category.

Because scaled scores are on an equal interval scale, it is possible to compare score performance at different points along the scale. Much like a yardstick, differences are constant at different measurement points. For example, a difference of two inches between 12 and 14 inches is the same as a difference of two inches between 30 and 32 inches. Similarly, for equal interval

achievement scales, a difference of 40 scaled score points between 360 and 380 means the same difference in achievement as between 400 and 420.

The score scales established in 1997 will stay in place for all subsequent years and test forms. Although new test forms are developed each year, scale scores are “equated” using “anchor” items that were used in the previous years and which have been equated to the base operational year (1997). This maintains the same scale score (400) for meeting the standard. Although the raw score to scale score relationship will change, the level of difficulty associated with meeting the standard in each tested content area will remain statistically equivalent over time.

Each year WASL tests may have a different raw score to scale score relationship, although the underlying scale remains the same from year to year. Tables 2-3 through 2-5 show the raw scores and their equivalent scale scores for the tests given in 1998, 1999, and 2000.¹³ A scale score of 400 is the “cut point” where the standard is met. The points and scores indicating partial achievement of the standard (Level 2) and achievement above the standard (Level 4) are also highlighted.

Students receive a single, comprehensive scale score based on the performance standards established by the standard setting committee. The level of performance (Level 1–4) is also provided for each student. In addition, the five content and four process strands are reported as strengths or weaknesses. Individual student data are also aggregated at the school and district levels.

The student-level test results were analyzed in various ways to determine the validity of the scores, that is, that the scores measure what they are supposed to measure. Technical reports that contain these analyses were prepared in accordance with professional testing standards.¹⁴ (See Appendix G) for more information about validity-related analyses.)

¹³ The range of scale scores each year differs. For example, the range of scale scores in 1998 was from 195 to 552 and the range in 1999 was from 194 to 573.

¹⁴ See *Standards for Educational and Psychological Testing*, a document prepared by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 1999.

Table 2-3: 1998 Grade 4 Mathematics Raw Scores to Scale Scores

Raw Score	Mathematics Scale Score	Raw Score	Mathematics Scale Score
0	195	32	386
1	222	33	388
2	249	34	391
3	265	35	393
4	277	36	395
5	286	37	398
6	294	38	400
7	301	39	402
8	306	40	405
9	312	41	407
10	317	42	410
11	321	43	412
12	325	44	414
13	330	45	417
14	333	46	420
15	337	47	422
16	341	48	425
17	344	49	428
18	347	50	431
19	351	51	435
20	354	52	439
21	357	53	443
22	360	54	447
23	362	55	452
24	365	56	458
25	368	57	465
26	371	58	473
27	373	59	484
28	376	60	500
29	379	61	526
30	381	62	552
31	384		

Table 2-4: 1999 Grade 4 Mathematics Raw Scores to Scale Scores

Raw Score	Mathematics Scale Score	Raw Score	Mathematics Scale Score
0	194	32	392
1	220	33	394
2	248	34	397
3	264	35	400
4	276	36	402
5	286	37	404
6	294	38	407
7	301	39	410
8	307	40	412
9	313	41	415
10	318	42	418
11	323	43	421
12	327	44	423
13	332	45	426
14	336	46	429
15	340	47	433
16	344	48	436
17	347	49	439
18	351	50	443
19	354	51	447
20	357	52	451
21	361	53	456
22	364	54	461
23	367	55	467
24	370	56	473
25	373	57	481
26	375	58	490
27	378	59	502
28	381	60	518
29	384	61	546
30	386	62	573
31	389		

Table 2-5: 2000 Grade 4 Mathematics Raw Scores to Scale Scores

Raw Score	Mathematics Scale Score	Raw Score	Mathematics Scale Score
0	190	32	394
1	218	33	396
2	246	34	399
3	264	35	401
4	277	36	403
5	288	37	406
6	297	38	408
7	305	39	410
8	312	40	413
9	318	41	415
10	323	42	418
11	329	43	420
12	333	44	423
13	338	45	426
14	342	46	429
15	346	47	432
16	350	48	435
17	353	49	439
18	357	50	442
19	360	51	446
20	363	52	450
21	366	53	455
22	369	54	460
23	372	55	466
24	374	56	472
25	377	57	480
26	380	58	489
27	382	59	502
28	385	60	519
29	387	61	549
30	389	62	577
31	392		

CONCLUSION

Washington engaged in an extensive test development process, summarized in Table 2-6. This process was in accordance with the 1999 professional standards. The various committees were comprised of a wide range of people from across the state. The experts who have reviewed the test development and standard setting processes have concluded that the “expert judgment” process used to set the standard in Washington was sound. Validity studies have also been conducted to help ensure the WASL measures what it intends to measure. Nevertheless, concerns have been raised about various aspects of the 4th grade test. The rest of this report provides information related to these issues.

Table 2-6: Summary of the Test Development Process (Grade 4)

Action	Dates
Essential Academic Learning Requirements	March 1995
Test and Item Specifications	Sept. – Oct. 1995
Item Development	Oct. – Dec. 1995
Item Review (Content and Fairness)	January 1996
Pilot Testing	May 1996
Item Review (Content and Fairness)	August 1996
Item Bank	September 1996
Operational Tests Created	Oct. – Dec. 1996
Published Example Test Assessment Sampler	February 1997
First Operational Test Administered	April – May 1997
Standard Setting	June 1997
Score Reports Designed	September 1997

Chapter 3

FREQUENCY OF TEST ITEMS

The Legislature required OSPI to analyze the percentage of items on the 4th grade mathematics WASL that (1) require students to use computational skills with and without the use of technology, (2) measure mathematics communications and problem-solving skills, and (3) measure other skills included in the assessment. OSPI staff and an independent expert familiar with the mathematics EALRs and WASL conducted these analyses for the mathematics assessments administered in 1998, 1999, and 2000.

To understand our analysis of the assessment, a brief review of the test's structure is needed. The two sections of the 4th grade mathematics WASL have similar formats. First, there is a mixture of multiple-choice, short-answer, and extended-response items. Multiple-choice items are worth one point, short-answer items are worth two points, and extended-response items are worth four points. Second, 20 items are given during one session, and 20 more items are given in a second session (typically given on the next day), for a total of 40 items. Students can score up to a maximum of 62 points on the 40 items. Third, each item assesses student performance on only one of the nine mathematics strands, even though an item may have material related to other strands. For example, an item assessing "measurement" might include geometric figures, which relates to the "geometric sense" strand.

COMPUTATION AND THE USE OF TECHNOLOGY

Some items on the assessment require computation, while others do not. An example of an item that does not require computation is one that asks students to identify parallel lines from among other lines. Some students may use computational skills when approaching a particular item, while others may not. This is due to the fact that students approach items in different ways. Our approach counted items that required computational understanding, even though actual computations may not have been required to complete the item.

The test specifications allow the use of calculators during the first session of the assessment. Since there are 20 items on this portion of the assessment, calculators can be used on half the items on the assessment. However, roughly half the items during the first session do not require any type of computation skills, so having a calculator is not useful on these items.

On average, 55 percent of all the items may require computation skills, and calculators were allowed (but not required) on half of these items.¹⁵ Table 3-1 shows the results of this analysis for the assessments given in 1998, 1999, and 2000 as well as for the combined 3-year period.

¹⁵ If a more conservative approach were used to determine the percentage of items requiring computational skills (i.e., items that required actual calculations in any part of the item were included, but items that required only an

Table 3-1: Analysis of Items Allowing Use of Calculators and Requiring Computation

	Year of Assessment			3-Year Total
	1998	1999	2000	
Total number of items allowing use of calculators	20	20	20	60
Total number of items on the assessment	40	40	40	120
<i>Percent of items that allows use of calculators</i>	50%	50%	50%	50%
1 st session items that may utilize computation skills	11	10	12	33
2 nd session items that may utilize computation skills	9	10	14	33
Total number of items that may utilize computation skills	20	20	26	66
<i>Percent of all items that may utilize computation skills</i>	50%	50%	65%	55%
<i>Percent of all items that may utilize computation skills and allow the use of calculators</i>	27.5%	25%	30%	27.5%

ANALYSIS OF MATHEMATICS SKILLS ASSESSED

The Legislature required OSPI to analyze the percentage of items on the 4th grade mathematics WASL that measure mathematics communications skills, problem-solving skills, and other skills included in the assessment. Communications and problem-solving are just two of the strands included in the nine EALRs, so an analysis of all strands assessed was conducted. The nine strands are listed below.

Content Strands

- _ Number sense
- _ Measurement
- _ Geometric sense
- _ Probability and statistics
- _ Algebraic sense

Process Strands

- _ Problem solving
- _ Reasons logically
- _ Communicating understanding
- _ Makes connections

The test specifications call for a range in the number of items that assess each strand. These items are to have a range in the number of points that can be awarded. For example, there are to be 4–8 items on the assessment that measure “number sense,” with a total of 5–9 points awarded for this strand. Each item assesses only one strand, even though information related to other strands may be included.

understanding of computation skills were not included), the percentages would be 36.7 percent of all items and 18.3 percent of the items allowing the use of calculators.

Our analysis found that the assessments conformed to the test specifications, that is, the number of items assessed were within the required range. Table 3-2 shows the number and percentage of items assessed for each strand for the assessments given in 1998, 1999, and 2000.

Table 3-2: Analysis of Items Assessing Mathematics Strands

Essential Learning Strands	Number of items in test specifications	Number of items assessing strand			Percent of items assessing strand		
		1998	1999	2000	1998	1999	2000
Concepts and Procedures		26	26	25	65%	65%	62.5%
Number sense	4–8	6	6	5	15.0%	15.0%	12.5%
Measurement	4–7	5	5	5	12.5%	12.5%	12.5%
Geometric sense	3–6	5	5	5	12.5%	12.5%	12.5%
Probability/statistics	4–7	5	5	5	12.5%	12.5%	12.5%
Algebraic sense	3–6	5	5	5	12.5%	12.5%	12.5%
Processes		14	14	15	35%	35%	37.5%
Solving problems	2–6	4	3	3	10.0%	7.5%	7.5%
Reasoning logically	2–6	4	4	3	10.0%	10.0%	7.5%
Communicating	2–6	3	3	4	7.5%	7.5%	10.0%
Making connections	2–6	3	4	5	7.5%	10.0%	12.5%

Analysis of Strands Assessed Indirectly

Table 3-3 shows the extent to which concepts from other strands were included on each item. These strands were not assessed directly, but students needed to have an understanding of concepts in these other strands in order to complete the item successfully. Analyses of the items found that nearly every item contained concepts related to another strand. In some cases, items included concepts related to four or more other strands. “Number sense” and “communications” were the other strands found most frequently in the items, while some strands were not present very often (i.e., algebraic sense, geometric sense, and making connections). OSPI should keep these proportions in mind when developing and designing future assessments.

Appendix B contains a more detailed analysis of the test items from the three years, including the type of item, the strand assessed, other strands related to the item, and the relative difficulty of each item. Data about how students scored on each item on the 1998 and 1999 tests are also provided in the appendix.

Table 3-3: Analysis of Items Related to Mathematics Strands Not Being Assessed

Essential Learning Strands	Percent of items assessing strand			3-year average
	1998	1999	2000	
Concepts and Procedures				
Number Sense	57.5%	60.0%	75.0%	64.2%
Measurement	15.0%	20.0%	20.0%	18.3%
Geometric Sense	7.5%	7.5%	2.5%	5.8%
Probability & Statistics	25.0%	25.0%	17.5%	22.5%
Algebraic Sense	5.0%	5.0%	2.5%	4.2%
Processes				
Solving Problems	20.0%	10.0%	20.0%	16.7%
Reasoning Logically	60.0%	32.5%	30.0%	40.8%
Communicating	42.5%	65.0%	70.0%	59.2%
Making Connections	5.0%	5.0%	7.5%	5.8%

Chapter 4

ASSESSING THE TEST'S DIFFICULTY

The Legislature required OSPI to study the developmental level required for students to achieve the 4th grade WASL standard. Chapter 2 discussed how the standard setting committee established a standard based on what a well taught, hard working student should know and be able to do in the spring of grade 4. The individual items developed for the WASL assess various aspects of the mathematics curriculum associated with the EALRs. By summing student performance from each item of the test, a total score is calculated. Each WASL has a specific total score that students are expected to achieve in order to meet the standard. These scores are equated and placed on a common scale so that comparisons can be made across school years.

Various analyses are required to determine whether this standard is an appropriate level on the actual tests and whether there is any aspect of the assessment that would raise questions about its validity (i.e., whether the WASL results accurately reflect student progress toward meeting the EALRs). To identify possible problems with the assessment, OSPI commissioned the Northwest Regional Educational Laboratory (NWREL) to analyze the items and results of the 4th grade mathematics WASL to determine if the individual items and the test as a whole were inappropriate in any way. (More information about the analyses is provided in Appendix H.) In addition, OSPI had independent experts analyze each item of the assessments to determine if they were age-appropriate and aligned with the EALRs. This chapter provides a summary of these various analyses.

ANALYZING TEST DIFFICULTY

The purpose of the work conducted by NWREL was to identify the sources of item complexity and determine whether items found on the 4th grade mathematics WASL and the test as a whole are age-appropriate. Test items are not appropriate if students in the test grade level do not have the cognitive ability to understand the required mathematics knowledge and skills. Three types of complexity can make a test difficult and potentially inappropriate—cognitive complexity, mathematics complexity, and format complexity.

- *Cognitive or developmental complexity* includes sources of difficulty introduced when students are asked to classify and relate more and more information to solve a problem. As an item requires a student to do more, the student reaches a point at which they may be unable to mentally manage all the information and relationships necessary to solve the problem correctly. This type of complexity is appropriate when schools provide students with opportunities to develop their cognitive skills through appropriate curricular and instructional decisions. Students' ability to control and coordinate cognitive activity is partly determined by their cognitive development and partly determined by their learning within a particular

context. A school's mathematics "environment" should provide sufficient opportunity to learn by offering quality curriculum and instruction. Additional factors such as school leadership and parental involvement also influence and support this growth.

- *Mathematics complexity* includes difficulty associated with an item due to the mathematics knowledge and skills required to answer the item correctly. A learning environment influences the development of such knowledge and skills as students are exposed to and practice mathematical concepts. Mathematics complexity is a desirable source of age-appropriate item difficulty.
- *Format complexity* includes sources of difficulty resulting from the way an item is presented, including the directions, question and response formats, response spaces and options, and the way information is presented in tables, graphs, and illustrations. Unusual, unclear, or difficult formats are inappropriate because they challenge the student's capability to understand the intent of the item writer. Format complexity is a less desirable source of item difficulty because it can interfere with students' demonstration of mathematical and cognitive ability.

Definitions and Assumptions

For this study, NWREL defined *developmentally-appropriate* (age-appropriate) items as those that test skills that are directly sensitive to instruction and that 4th grade students have the opportunity to master. In other words, students who have reached an appropriate developmental level can be taught the skills necessary for solving age-appropriate problems. *Developmentally-inappropriate* (i.e., not age-appropriate) items test skills that are not sensitive to instruction. When items are not age-appropriate, students who have not reached an appropriate developmental level cannot learn the required skills necessary for solving the items, regardless of the extent or form of instruction.

NWREL made certain assumptions when conducting their analyses. First, they assumed that the 4th grade mathematics WASL *could be* developmentally beyond the population it was designed to assess. This assumption was based on qualitative evidence taken from newspapers, letters, and emails that expressed concern about the test's difficulty. The study looked critically for evidence that the test was not age-appropriate.

Second, NWREL assumed that the test population was students enrolled in Washington's elementary schools who attended the 4th grade in the school year during which the test was administered. This definition includes special education students and students who use English as a second language. However, students who were only in the country for one year are exempted from testing if they did not score well on a language test, and certain special education students may receive accommodations or an exemption if stipulated by their Individualized Education Plan (IEP).

Third, performance was assumed to be influenced by factors observed in schools. Since the WASL is a criterion-referenced test, its purpose is to make inferences about whether students have reached some specified standard relative to the EALRs. The standard is a preconceived criterion or measure of performance that students are expected to meet. Instruction and curriculum aligned with the EALRs should result in better test performance over time.

Finally, the quality of the curricular and instructional experiences students receive varies among schools. A school's ability to develop a proper learning environment will improve student performance. The required skills necessary for solving any problem are jointly determined by the constructive actions of students and the context that supports these actions. If the students are to meet standards, schools must actively support the learner in order for students to attain the specified standard.

RESULTS OF THE NWREL ANALYSIS

NWREL's analysis focused on describing the cognitive and mathematical demands placed on students in solving any item found on the 1998 and 1999 mathematics WASL and determining whether the items were age-appropriate for 4th grade students. The analysis has three parts and comes to some general conclusions regarding the age-appropriateness of WASL mathematics items. In summary:

1. Scale scores were calculated for items from the 1998 and 1999 assessments and were analyzed. The distributions of scale scores suggest that some items on the assessments are very easy and some are very difficult. If the very difficult items were beyond the developmental level of 4th grade students, the students still had ample opportunity to meet the standard by doing well on other items.
2. The three sources of complexity were analyzed to determine the extent to which they explain the tests' difficulty. All three types of complexity contributed to the tests' difficulty, with cognitive complexity explaining the most variation in both years.
3. Analyses were conducted to determine the level of improvement that had occurred on mathematics anchor items in 1998 and 1999. The analyses found that students statewide were performing better on the test, with some schools showing improvement far beyond that of the state averages. Improvement was made among all ethnic/racial groups statewide, and dramatic improvement was made by all ethnic/racial groups in some schools. These results suggest that the test is sensitive to instruction and curriculum changes.

The sections below discuss each part of the analysis. Together, they suggest that in general the test is appropriate for Washington students, but it can be improved. (See Appendix H for more information about NWREL's analysis.)

Item Scale Scores

Test items can be ranked according to their relative difficulty. Scaling techniques can be used to put items of different difficulty on a numerical scale that includes the state performance standard. Figure 4-1 depicts the distribution of scale scores for items on the 1998 and 1999 tests. This numerical scale not only ranks the items from easy to difficult, but it also weights the item difficulties so that the actual distance between these scale values may be accurately calculated. In the case of the WASL, by anchoring the scale at the predetermined standard of 400, the relative item difficulty on both tests are transferred to a numerical scale ranging from 295 to 440.

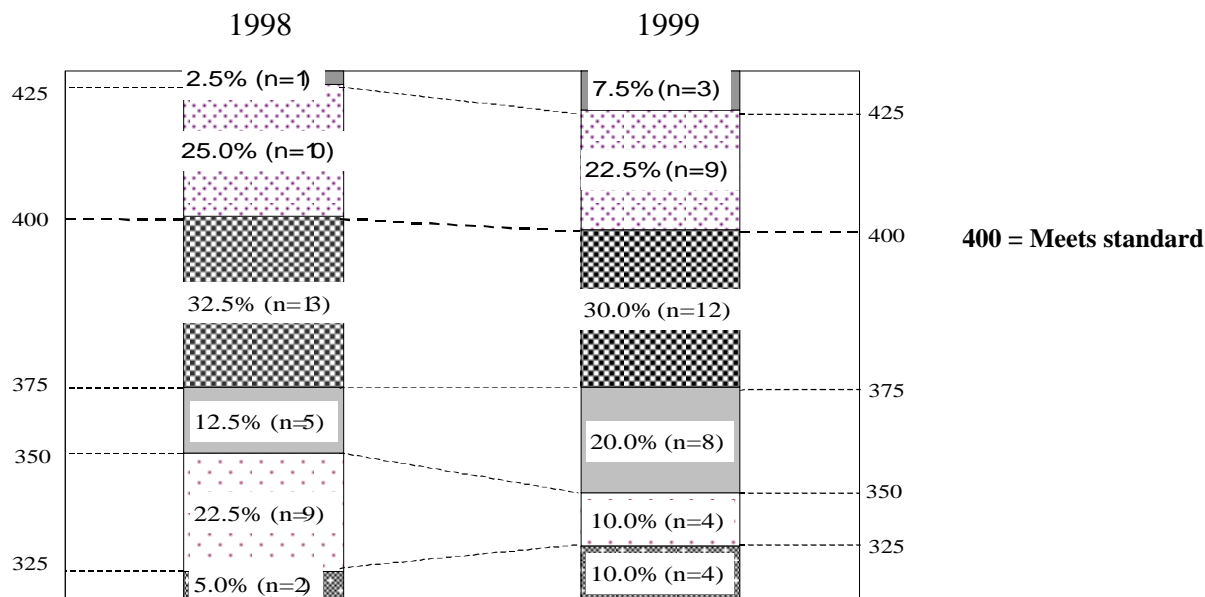
The distributions in Figure 4-1 describe the relative percentage of the item scale scores around the standard of 400. Items whose scaled difficulties are distributed relatively closer to the standard are

more informative than items scaled well below or well above the standard. That is, the farther away the item scale scores are from 400, the less of a contribution the test information makes for judging whether a student meets or fails to meet the standard.

When examining the relative distributions of the item scale scores, some items were found to have scale scores far above and far below the 400 standard. This suggests that a few items were very easy (two in each year) while a few were very difficult (one in 1998 and three in 1999). Since the test is intended to provide information about student ability across a range of items and not just around the 400 standard, it is appropriate to have some easy and some difficult items on the test. If one assumed that the very difficult items were not age-appropriate, then 5 percent (4 of 80) of the items on the 1998 and 1999 tests would not have been appropriate. However, student performance on these items indicate that some students received full or partial credit for their answers.

As shown in the figure, a student would need to get 29 of the 40 items correct on the 1998 assessment to meet the standard. In the 1999 assessment, a student would need to get 28 of the 40 items correct to meet the standard. If one assumed the very hard items were beyond the developmental level of a well-taught, hard-working student and were eliminated from the possible point total, there would still be ample opportunity for a student to meet the standard by doing well on the other items.¹⁶

Figure 4-1: Distribution of Scale Scores of All Items



Sources of Complexity

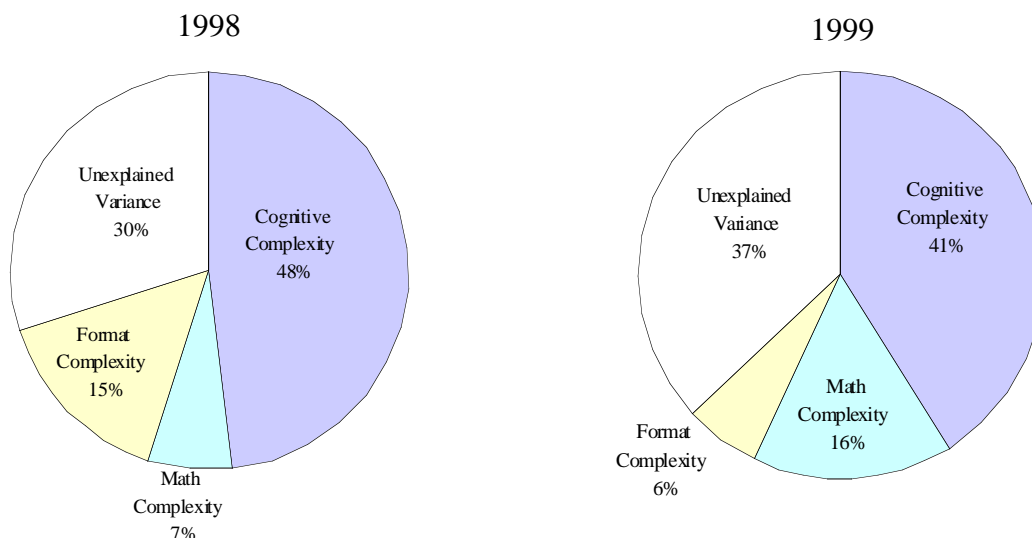
Items can be difficult for different reasons. NWREL created a complex model to analyze each item in terms of the three types of complexity—cognitive, mathematics, and format (see Appendix H). NWREL then performed regression analyses using item difficulty as the dependent variable

¹⁶ Recall from Chapter 2 that a student needed to score 38 out of a possible 62 points on the 1998 assessment to meet the standard and score 35 out of a possible 62 points on the 1999 assessment to meet the standard.

and cognitive complexity, mathematics complexity, and format complexity as the independent variables. This type of analysis shows what contributes to the difficulty of the test.

In both the 1998 and 1999 tests, cognitive complexity contributed the most to the test's difficulty. Mathematics and format complexity both accounted for a smaller portion of test difficulty. Together, the three types of complexity explained about two-thirds of the variation in the tests' difficulty (see Figure 4-2).¹⁷ Cognitive and mathematics complexity are appropriate and desired to test student performance. However, format complexity is less desirable because it may mask a student's cognitive and mathematical abilities. In 1998, format complexity contributed more to the test's difficulty than mathematical complexity.

Figure 4-2: Relative Influence of Different Types of Complexity



Two well-known sources of influence on test scores were not included in the analysis and could account for much of the remaining unexplained variance.

- Analyses of test results have found a strong link between student performance (i.e., how students score on standardized tests) and a student's socioeconomic status. Students from low-income families do not perform as well as students who are not from low-income families. Thus, the family background of students is likely to explain some of the remaining variance.
- Research has also shown that quality instruction and other school-related factors can help students perform better on tests and overcome a low socioeconomic background. Thus, the quality of instructional practices, mathematics curriculum, and the opportunity to learn specific mathematics knowledge and skills are also likely to be responsible for part of the remaining variance. Schools provide students with differing opportunities to practice and master specific mathematics content and cognitive processes. If some students are taught ways to solve certain types of problems, they will perform better than students who have not been taught these methods.

¹⁷ All three variables were statistically significant in both years. The R-square was .695 for 1998 and .631 for 1999.

Comparative Analyses

Another way to analyze the age-appropriateness of the WASL items is to investigate the progress of schools from year to year. Test items that are not sensitive to instruction (i.e., students fail to show growth following instruction and mastery of skills and knowledge) are likely to be inappropriate. Conversely, if more 4th grade students are able to correctly respond to “hard” items over time, there is evidence that the items are sensitive to changes in curriculum, instruction, and student motivation and are therefore appropriate for well taught, hard working students.

Table 4-1 and Figures 4-2 and 4-3 present data that show improvement on the 1998 and 1999 assessments, both for students statewide and students in 35 “achieving schools” identified by OSPI.¹⁸ (See Chapters 5 and 6 for more information about the progress being made on the mathematics WASL.)

Table 4-1 shows 15 items that were common in both years of the test (i.e., anchor items). Column 1 presents the item number, column 2 describes whether the item is multiple-choice or short-answer, and column 3 lists the mathematics strand assessed by the item. The data in the last four columns represent the percentage of correct answers on the test for each of the multiple choice and short answer items.¹⁹ Short-answer items have three percentages because these items have three possible scores (0, 1, and 2). Each percentage provides the distribution for these three scores on short-answer items (i.e., the first percent is for a score of 0, the second for a score of 1, and the third for a score of 2). Gains in item achievement of schools in the entire state are shown along with those of the 35 achieving schools.

The table indicates that students in the 35 schools out-performed the state average on nearly all the common items. Often the difference in improvement between the state average and the average of the 35 schools was relatively large. On item 16, for example, students statewide had an improvement of 9 percentage points (42 to 51) from 1998 to 1999, while the achieving schools improved 16 percentage points (45 to 61) during the same period.

¹⁸ These schools were those identified as having made dramatic improvement on the 4th grade mathematics WASL between 1997 and 1999. Individual student scores were available for 35 of the 38 identified schools at the time of the analysis. See Chapter 6 for more information about these schools.

¹⁹ The numbers in the table differ slightly from those that appear in Appendix B because different methods were used to determine the percentage of students with each possible score.

Table 4-1: Percent Correct for Common Items, State vs. Achieving Schools

Anchor Item	Item Type ¹	Strand Assessed	State Results (percent correct)		Achieving Schools (percent correct)	
			1998	1999	1998	1999
1	MC	Solves Problem	70.4	70.7	73.5	75.6
2	MC	Makes Connections	88.3	89.5	87.2	91.4
9	MC	Measurement	39.3	40.6	38.9	45.8
11	MC	Algebraic Sense	68.4	69.9	70.3	74.9
12	MC	Geometric Sense	72.1	76.0	75.3	81.8
16	MC	Logical Reasoning	42.2	51.1	44.6	60.6
17	MC	Prob./Statistics	41.3	51.1	44.8	57.1
20	SA	Algebraic Sense	47, 15, 35	44, 17, 37	39, 16, 41	31, 15, 50
21	MC	Algebraic Sense	66.8	65.7	69.1	67.5
23	SA	Logical Reasoning	44, 20, 35	42, 20, 37	38, 16, 41	32, 19, 46
24	MC	Number Sense	26.2	27.0	31.1	35.8
27	MC	Geometric Sense	40.4	42.0	41.1	51.9
28	MC	Prob./Statistics	67.4	67.0	69.4	73.5
32	SA	Problem Solving	67, 8, 19	69, 8, 19	64, 8, 22	56, 10, 30
40	MC	Measurement	53.0	56.4	51.4	59.2

¹ MC = Multiple-choice SA = Short-answer

Analyses of ethnic/racial groups were conducted to determine if different groups of students were more likely to show improvement. Figure 4-3 shows the extent to which students of each ethnic/racial group had met the standard in 1998 and 1999. The figure compares students statewide and those in the 35 achieving schools. Figure 4-4 shows the average scale scores for students statewide and in the 35 achieving schools according to ethnic/racial group for the same two years. The analyses found that more students of each ethnic/racial group were meeting the standard in 1999 than in 1998. From a statewide perspective, some of the gains were small. However, dramatic improvement occurred across all ethnic/racial groups in the 35 achieving schools, even after proportionately weighting the results to reflect the state's racial composition.²⁰ Analyses of the average scale scores for each group found the same pattern (Figure 4-4). The overall average and the average scores of three ethnic/racial groups in the achieving schools exceeded the state standard.

These results show that some schools are more effective than others at increasing students' cognitive and mathematics growth relative to the WASL items. In other words, schools can affect achievement on the WASL test items. These school effects were present regardless of the ethnic/racial or socioeconomic composition of the students. Based on these findings, it seems reasonable to conclude that the range of difficulty associated with most WASL items is within a developmental range appropriate for 4th grade students.

²⁰ The 1998 level for the achieving schools is generally higher than the 1998 level statewide because the achieving schools had made improvements greater than the state average between 1997 and 1998.

Figure 4-3: Percent Students Meeting Standard by Ethnic/Racial Group, 1998 and 1999

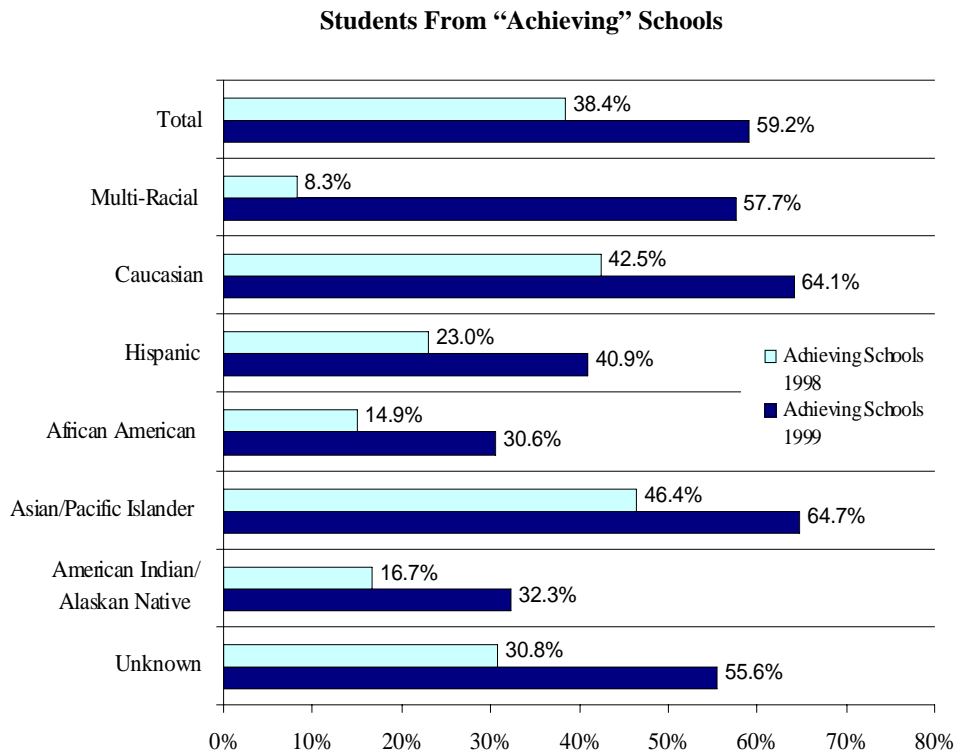
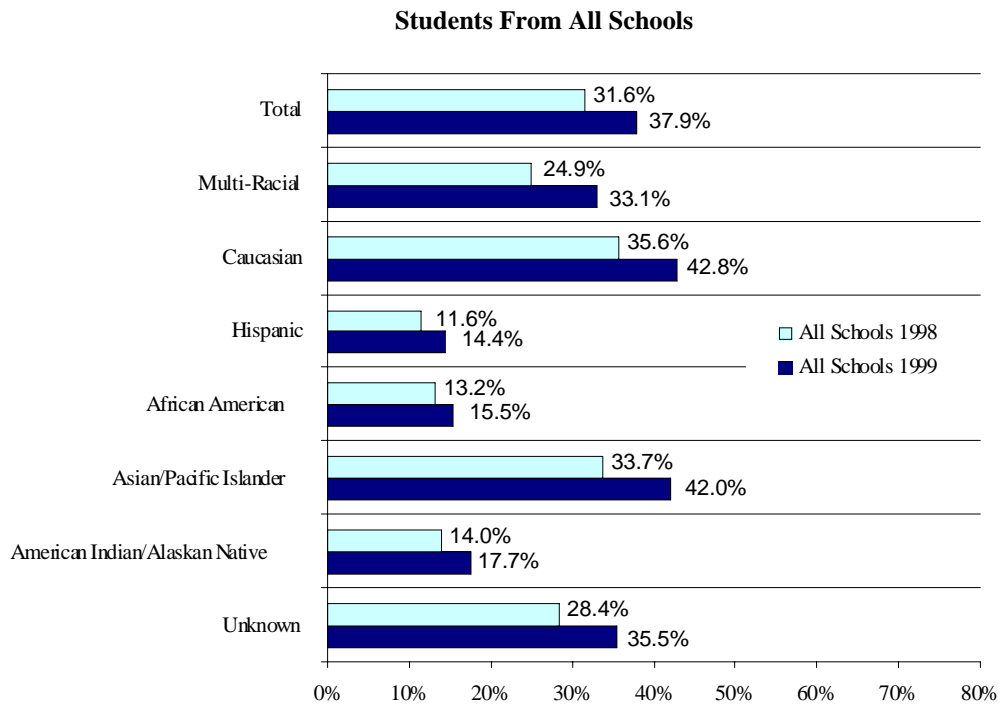
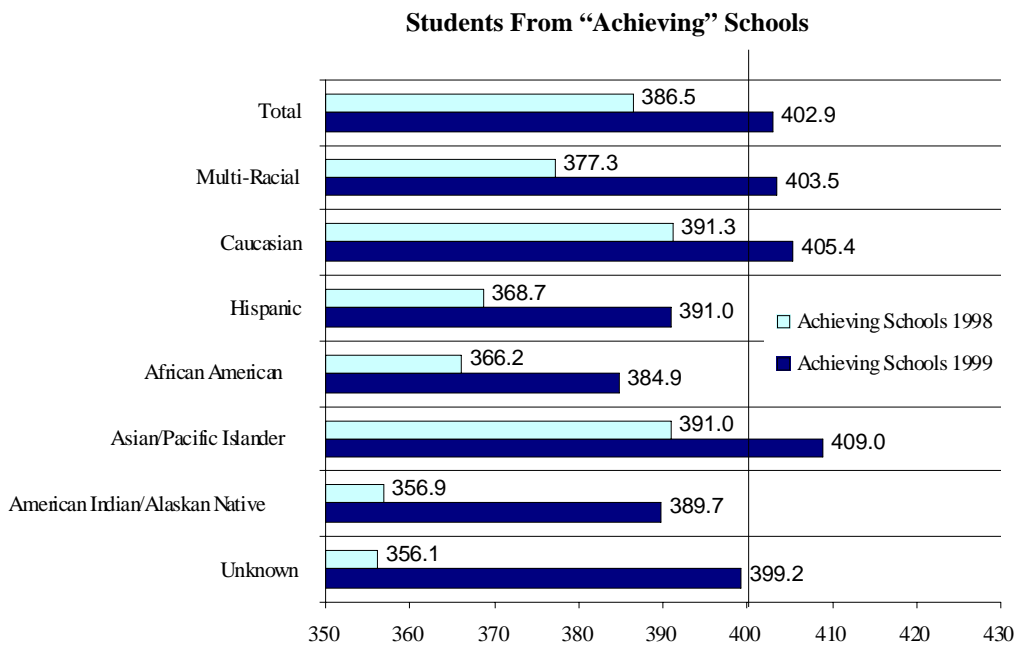
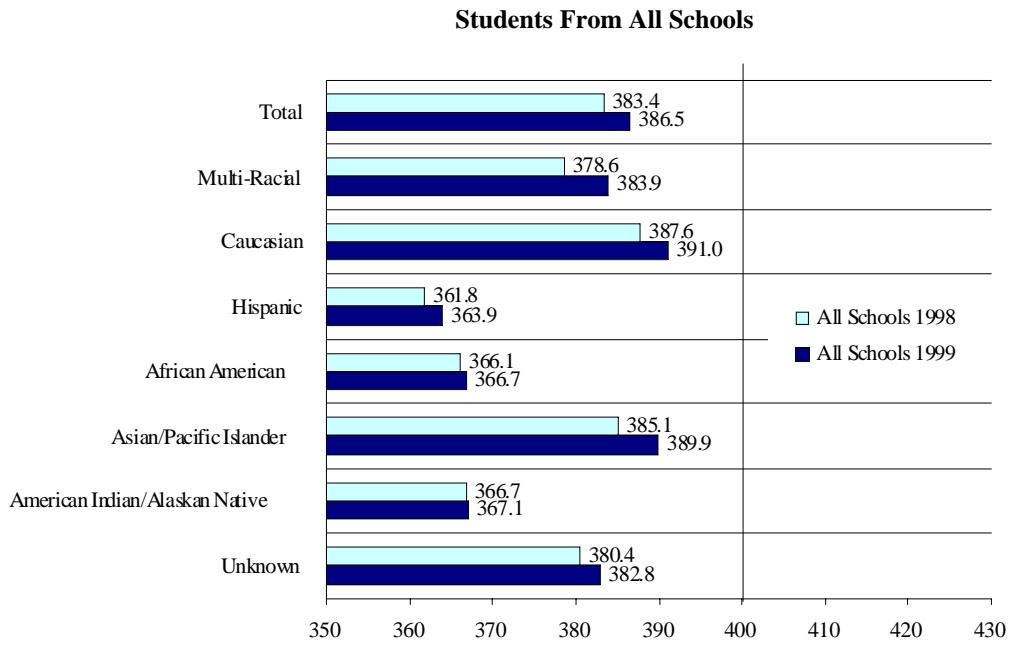


Figure 4-4: Average Scale Score by Ethnic/Racial Group, 1998 and 1999
(400 = Meets Standard)



ALIGNMENT WITH THE ESSENTIAL LEARNING REQUIREMENTS

To supplement the work performed by NWREL, two independent mathematics experts familiar with the EALRs and the WASL conducted analyses to determine if each item on the 1998, 1999, and 2000 tests and on the Example Test was aligned with the 4th grade mathematics EALRs and benchmarks.²¹ An important issue in test development is making sure the test actually measures the concepts and skills that it is supposed to measure. If the content of the test is not the same as the content of the intended curriculum, the test may lack validity.

Several problems make this type of analysis difficult.

- The EALRs and benchmarks are fairly general in some areas and subject to interpretation, making it difficult to say if certain items are aligned with the 4th grade benchmarks. Some items may be clearly aligned or not aligned, while others may be “on the edge” of alignment.
- Some types of items, such as those involving proportional reasoning, assess competence in areas that develop over time and require skills that begin to develop before 4th grade and continue to develop in later years. An item may appear in relatively simple form in the early grades and be more complex in tests in later grades.
- Adults view test items differently than do 4th grade students. Adults may view an item with greater sophistication than is required to solve the problem. Seen from the adult perspective, an item may not be aligned. However, when less sophisticated solution paths are identified, solutions are usually accessible by building on expected 4th grade knowledge. Thus, an item that appears difficult to an adult may be relatively simple to a 4th grade student.

Despite these challenges, the experts agreed that 10 of the 120 items on the tests from 1998, 1999, and 2000 were not aligned with the 4th grade EALR benchmarks. Seven of the 10 items were not aligned because they required students to “create a plan,” which is a task currently associated with the 7th grade EALRs. (Benchmark 4.1.1 for “communicating understanding” says that students should be able to “follow a plan,” but the test specifications *CUOI* says students should “create a plan.” Test developers follow the test specifications when constructing the assessment.) These “create a plan” items were among the more difficult items on the test, although over half the 4th grade students still received partial or full credit on several of these items. For example, 61 percent received full credit on one such item in 1998.²² This suggests that “create a plan” is within the developmental capability of 4th grade students.²³ Nevertheless, if teachers do not prepare their students to create a plan because such a task is not one of the benchmarks, their students are not likely to do well on such items.

The remaining three items that were not aligned with the EALRs were found to be above the 4th grade level. So, together with “create a plan” items, these 10 items represent 8 percent (10 of 120)

²¹ Verna Adams (Washington State University) and John Woodward (University of Puget Sound) conducted these analyses.

²² Three of the seven items not aligned with the EALRs were on the 2000 test. Student results on these items were not available when this study was completed.

²³ A further indication that “creating a plan” is within the ability of 4th grade students is the inclusion of “Creating a plan” in the state’s *Directions for Communicating through Mathematics: A Washington Model for Classroom-Based Evidence* for the primary grades (see page 26). This document is part of the Early Years Classroom-Based Assessment Tool Kit.

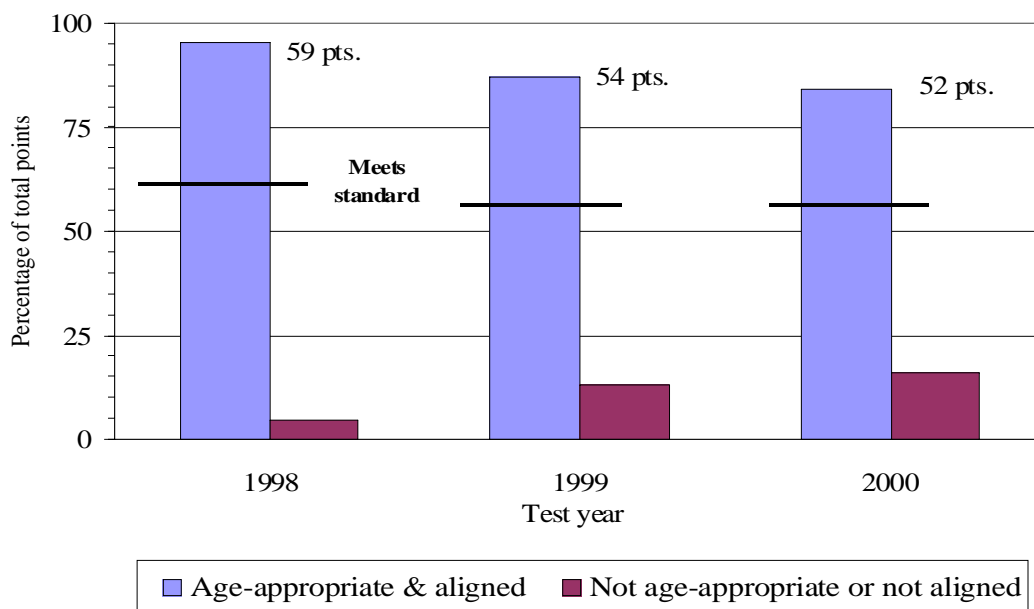
of all the items from the three tests. Eight of the 10 were short-answer items, one was an extended-response item, and one was a multiple-choice item. The experts also found that many of the concerns expressed about items in the Example Test were unfounded, although they found that some items on the Example Test were not aligned with the 4th grade benchmarks.

Table 4-2 provides more information about the 10 items. Although many students scored points on these 10 items, if the points from these 10 items were excluded from the totals, students still had ample opportunity to meet the standard in each year (see Figure 4-5).

Table 4-2: Analysis of Non-Aligned Test Items

	Year of Assessment			3-year Total
	1998	1999	2000	
Total items on the test	40	40	40	120
Number of non-aligned items	2	4	4	10
Percent not aligned	5%	10%	10%	8.3%
Total points on the test	62	62	62	186
Points of non-aligned items	3	8	10	21
Percent non aligned	5%	13%	16%	11.3%

Figure 4-5: Students Could Meet Standard By Doing Well on Other Items



Example of Item Not Aligned with EALRs

A look at an item from the Example Test provides an illustration of a question that is not aligned with the EALR benchmarks.²⁴ Item 19 in the Example Test, shown below, has responses that require students to understand measurement in the metric system. The 4th grade EALRs and benchmarks do not include knowing the metric system, although the 7th grade EALR benchmarks requires knowledge of metric measurements.

This item also includes format complexity. The word “middle” is confusing because it could be interpreted in several ways (it could mean either the circumference of the pencil or the pencil lead). In addition, 4th grade students are exposed to pencils of different lengths and widths, which could add further confusion.

19. Which of the following is closest to the distance around the middle of an unsharpened pencil?
- A. 25 millimeters
 - B. 25 centimeters
 - C. 25 meters

IMPLICATIONS

These analyses suggest that overall the 4th grade mathematics WASL is appropriate for 4th grade students, although some changes need to be made to improve the test.

- Overall student performance on the WASL gradually improved over the first three years, and students of all ethnic/racial groups showed dramatic improvement when given the opportunity to learn. This implies that the level needed to meet the standard is appropriate from a developmental point of view. Continued research is needed to investigate the educational practices in schools with students who exhibit mathematics and cognitive skills that are significantly higher than the state average.
- A few items on the assessments were found to be very difficult and possibly outside the developmental level of most 4th grade students. In addition, some items were found to assess concepts outside the 4th grade EALR benchmarks. OSPI has begun work with its contractors and technical advisors to ensure that items on operational and example tests are age-appropriate for 4th grade students and are aligned with appropriate EALR benchmarks.
- Some items have format problems that may mask students' cognitive and mathematical abilities. OSPI will take steps to rewrite or revise items with format problems so that the test only measures relevant aspects of cognitive and mathematical complexity.

Any changes to the test could have implications regarding the performance level needed to meet the standard. OSPI will consult with its technical advisors to determine if the changes identified in this study would require a change in the level needed to meet the state standard.

²⁴ Details about items on the actual test are confidential and cannot be disclosed.

Chapter 5

PERFORMANCE OF GRADE 4 STUDENTS ON THE MATHEMATICS WASL

The performance of various groups of 4th grade students on the mathematics WASL has improved over time. This provides evidence that the developmental level needed to meet the standard is appropriate to 4th graders. The assessment results are also useful for tracking the state's progress in helping students meet the EALRs.

This chapter provides information about how 4th grade students have performed on the mathematics WASL over time. The chapter also provides information on schools that have shown dramatic improvement in the percentage of students meeting the standard. Appendix G provides more detail on how students of different genders, ethnicities/races, and programs have performed on the test.

STATEWIDE TRENDS

Each year 4th grade students have performed better on the mathematics WASL. Since the test was administered the first time in 1997, the percentage of students meeting the standard has nearly doubled, with about 42 percent meeting the standard in 2000. The average scale score has gradually risen, and if the current rate of improvement continues, the average 4th grade score will be 400 (meets standard) in 2002. Table 5-1 and Figures 5-1 and 5-2 provide data on student performance statewide from 1997–2000.

Table 5-1: Performance by 4th Grade Students on the Mathematics WASL

	1997	1998	1999	2000
Level 1	47.2	37.8	33.6	31.4
Level 2	28.9	29.8	27.4	24.9
Level 3	14.7	20.2	23.3	22.4
Level 4	6.6	11.0	13.9	19.3
% meeting standard*	21.4	31.2	37.3	41.8
Average scale score	374.0	383.5	386.5	391.2

* Total may be different from the sum of Levels 3 and 4 due to rounding.

Figure 5-1: Change in Levels on 4th Grade Mathematics WASL, 1997–2000

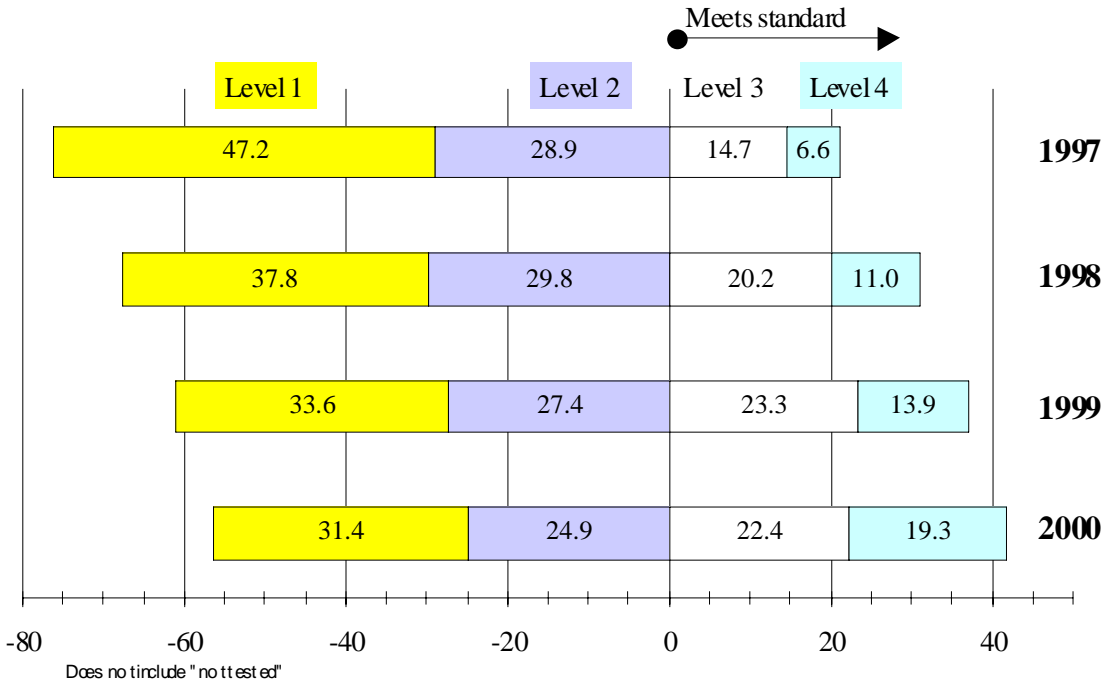
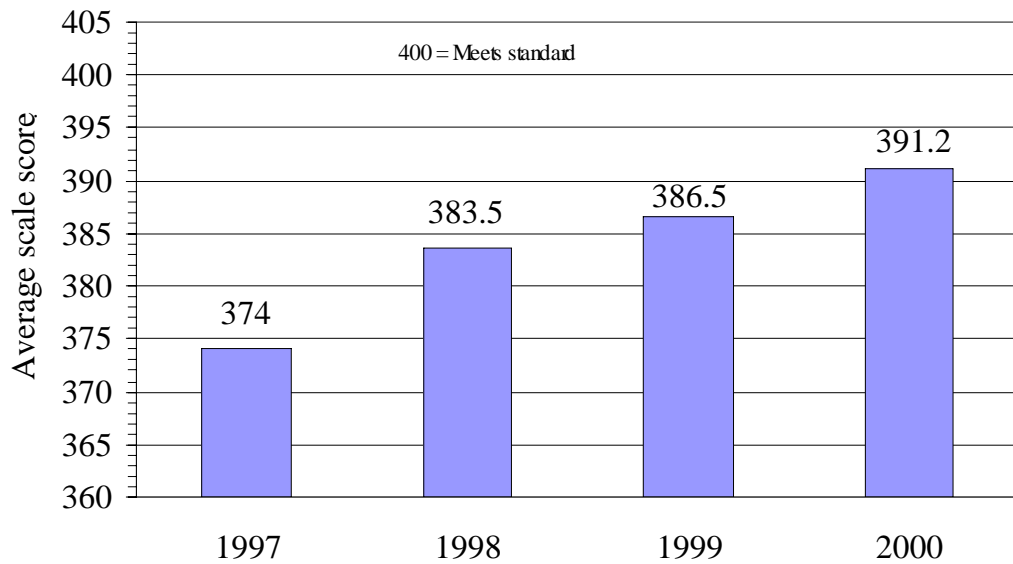


Figure 5-2: Average Score Is Approaching 400 (Meets Standard) on the 4th Grade Mathematics WASL



ANALYSIS OF SCHOOLS SHOWING THE MOST IMPROVEMENT

Some schools have shown much greater improvement on the 4th grade mathematics WASL than the state average. OSPI identified and contacted the 38 schools that showed the most dramatic improvement in the percentage of students meeting the standard between 1997 and 1999 to determine the factors that contributed to their improvement. This section discusses the methods used to identify these schools, their demographic context, the results of their mathematics WASL assessments, and factors that teachers and principals in those schools considered to be responsible for the improvement.

Criteria for Selecting Schools with Exemplary Improvement

OSPI used several criteria to identify truly exemplary schools. To be exemplary, a school's overall gains should be much larger than the average statewide gains. From 1997 to 1998, the percentage of students meeting the mathematics standard statewide increased by 9.8 percentage points, and from 1998 to 1999, the statewide increase was 6.1 percentage points. Thus, the total increase from 1997 to 1999 was about 16 percentage points statewide. To be considered exemplary for this study, a school had to meet or exceed the state average in the first year, sustain the gains in the second year beyond the state average, and have an overall gain that was roughly twice the state average. Specifically, the schools had to meet the following conditions.

1. The school had to have at least a 15-point gain from 1998 to 1999 and at least a 30-point gain from 1997 to 1999 (2-year gain) **or** at least a 10-point gain from 1998 to 1999 and at least a 35-point gain from 1997 to 1999 (2-year gain).
2. The school also had to have at least a 10-point gain from 1997 to 1998 and at least a 10-point gain from 1998 to 1999. (To be inclusive, gains were rounded up, so two schools with improvements of 9.5 from 1998 to 1999 were included.)
3. Only schools that had at least 40 students tested in each of the 3 years were considered in order to ensure valid inference of our results (i.e., the increase in test scores could be attributed to instructional interventions and not to statistical fluctuations normally attributed to few students tested). Requiring at least 40 students also ensured that there were at least two teachers in the school for 4th grade students. Of the 1,125 schools that administered the 4th grade WASL in 1999, 730 schools (65 percent) tested at least 40 students.

Using these criteria, OSPI identified 38 schools that made the most significant “gains” or progress on the 4th grade mathematics WASL over a 3-year testing period (Spring 1997 to Spring 1999).

Profile of Schools Meeting the Criteria

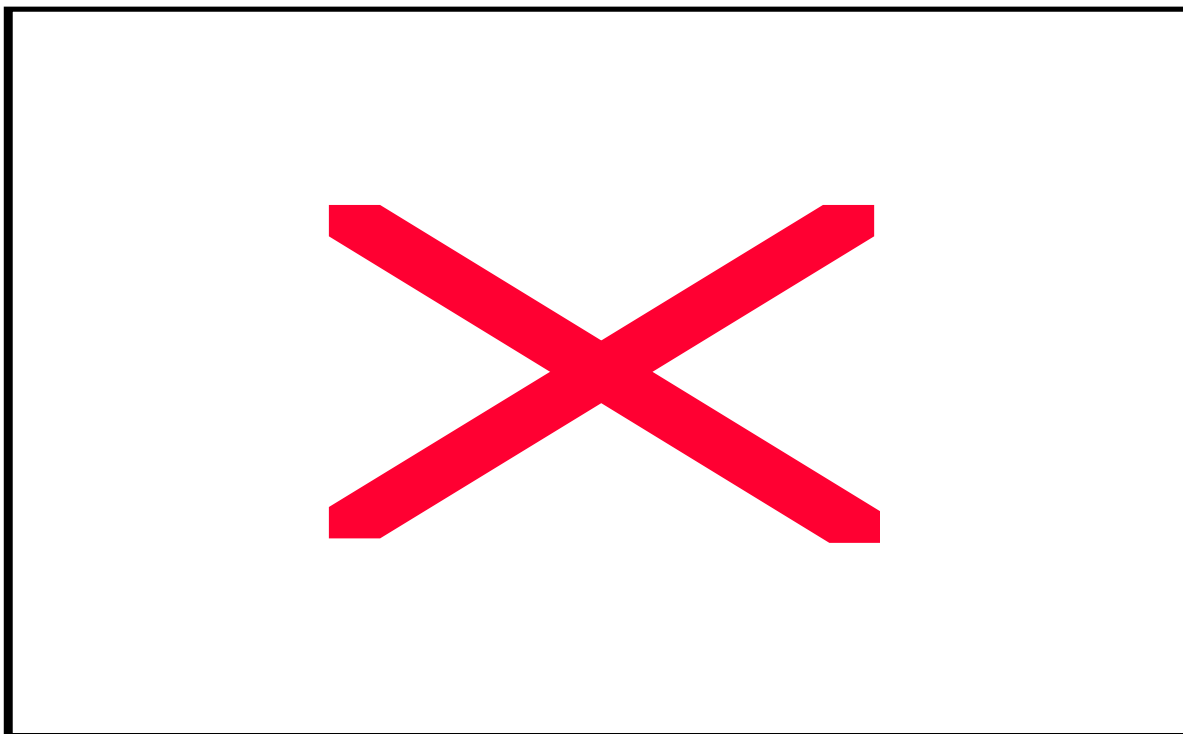
The 38 schools that met the selection criteria were fairly representative of other schools in the state from a demographic point of view. As a group, the schools were fairly typical of schools statewide—they had about the same percentage of students from low-income families and had the same percentage of students meeting the standard on the 4th grade mathematics WASL in 1997 (see Table 5-2). Their levels of teacher experience and education were also similar to the state average. The schools were located in all parts of the state (see Figure 5-3).

Table 5-2: Data for 38 Schools Showing the Most Improvement

DISTRICT	SCHOOL	Percent Low- Income*	Percent Meeting Mathematics WASL Standard – 4th Grade			Level of Improvement (Percentage Point Gain)		
			1997	1998	1999	1997 to 1998	1998 to 1999	1997 to 1999
Bellevue	Cherry Crest	4.4	32.9	56.7	78.8	23.8	22.1	45.9
Bellingham	Parkview	36.3	23.6	37.8	53.8	14.2	16.0	30.2
Central Kitsap	Emerald Heights	23.8	31.7	42.1	62.6	10.4	20.5	30.9
Edmonds	Maplewood	8.6	29.6	49.1	63.6	19.5	14.5	34.0
Ferndale	North Bellingham	31.6	24.6	43.8	60.3	19.2	16.5	35.7
Issaquah	Clark	8.3	30.0	42.9	61.0	12.9	18.1	31.0
Issaquah	Maple Hills	6.0	33.8	48.8	69.4	15.0	20.6	35.6
Kent	Carriage Crest	14.2	21.1	35.4	50.6	14.3	15.2	29.5
Kent	Pine Tree	32.0	15.1	29.8	49.5	14.7	19.7	34.4
Lake Washington	Laura Ingalls Wilder	0.2	39.8	61.7	82.9	21.9	21.2	43.1
Marysville	Allen Creek	8.2	24.2	34.9	58.9	10.7	24.0	34.7
Marysville	Shoultes	35.1	6.5	36.0	56.2	29.5	20.2	49.7
Mead	Farwell	46.0	13.5	38.9	48.4	25.4	9.5	34.9
Montesano	Beacon Avenue	36.9	21.3	42.6	57.1	21.3	14.5	35.8
Mukilteo	Challenger	46.3	7.0	18.0	64.6	11.0	46.6	57.6
Mukilteo	Columbia	13.6	29.5	50.5	72.1	21.0	21.6	42.6
Mukilteo	Olivia Park	32.6	11.5	38.7	53.5	27.2	14.8	42.0
North Kitsap	Hilder Pearson	21.7	29.2	41.7	60.4	12.5	18.7	31.2
North Kitsap	Richard Gordon	19.7	17.2	40.2	57.5	23.0	17.3	40.3
Northshore	Moorlands	7.9	19.5	56.1	69.2	36.6	13.1	49.7
Oak Harbor	Broad View	6.1	30.1	48.4	64.0	18.3	15.6	33.9
Olympia	Centennial	7.8	30.6	41.2	68.4	10.6	27.2	37.8
Olympia	Pioneer	6.1	25.7	54.5	68.2	28.8	13.7	42.5
Renton	Sierra Heights	27.5	11.0	60.2	81.0	49.2	20.8	70.0
Ridgefield	South Ridge	17.0	20.3	33.3	50.8	13.0	17.5	30.5
Shoreline	Lake Forest Park	5.8	44.2	54.5	75.0	10.3	20.5	30.8
Shoreline	Meridian Park	16.9	30.3	62.5	76.0	32.2	13.5	45.7
Shoreline	Parkwood	23.7	26.9	38.3	59.3	11.4	21.0	32.4
Snoqualmie Valley	Snoqualmie	17.2	20.3	33.3	50.0	13.0	16.7	29.7
Spokane	Bemiss	89.0	11.6	41.2	52.3	29.6	11.1	40.7
Spokane	Jefferson	23.8	30.6	57.6	67.1	27.0	9.5	36.5
Spokane	Lidgerwood	69.2	14.8	27.3	56.6	12.5	29.3	41.8
Tukwila	Thorndyke	63.9	6.3	25.9	54.2	19.6	28.3	47.9
Tumwater	Littlerock	25.4	9.6	31.4	49.3	21.8	17.9	39.7
Vancouver	M. L. King, Jr.	61.9	10.5	26.3	56.9	15.8	30.6	46.4
Wenatchee	Washington	32.3	16.3	40.0	51.5	23.7	11.5	35.2
White River	Mountain Meadow	26.6	10.6	28.3	43.1	17.7	14.8	32.5
Zillah	Zillah	54.1	3.3	15.2	55.0	11.9	39.8	51.7
	Group Average	26.5	21.4	41.2	60.8	19.8	19.6	39.3
	State Average	31.4	21.4	31.2	37.2	9.8	6.1	15.9

* Percent low-income was measured in terms of the percentage of students eligible to receive a free or reduced-priced lunch in the 1998-1999 school year.

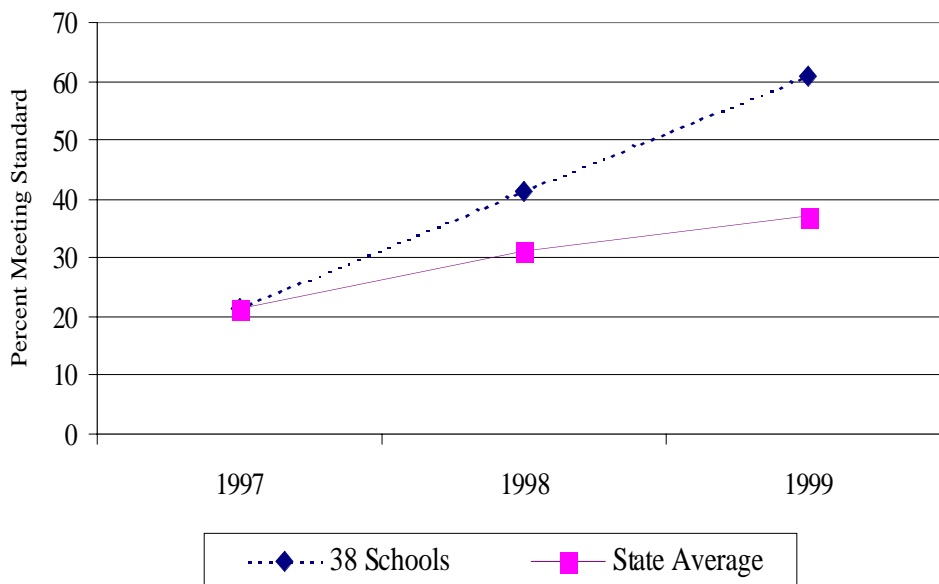
Figure 5-3: Districts Where 38 Schools Are Located



We analyzed the schools' demographic profiles over time to determine if a change in student population over the 3-year period may have accounted for the improvement. We found that the average percentage of low-income and minority students in the 38 schools was about the same in all three years, and the individual schools did not experience much change in the percentage of low-income or minority students.

Results of 4th Grade Mathematics WASL

In 1997, the first year the WASL was administered, the group of 38 schools had an average of 21 percent of their 4th grade students meeting the mathematics standard. This percentage was the same as the state average. In 1998, the group's average nearly doubled to 42 percent meeting standard, twice the increase of the state average. In 1999, the group sustained the same level of improvement, with their average jumping again by nearly the same amount (20 percentage points). This gain was more than triple the state gain. Figure 5-4 shows the group and state averages for the three years.

Figure 5-4: Results of Mathematics WASL, State and 38 Schools

Survey Results

To determine the factors that contributed to the dramatic improvement in the 38 schools, we sent a survey to all 38 schools. Principals and teachers involved in providing instruction in the 38 schools were asked to indicate the extent to which various factors may have contributed to the improvement in the percentage of students meeting the mathematics standard over the 1997–1999 period. Additional information was also solicited from teachers about their professional development and familiarity with the EALRs. Staff from all 38 schools responded.

The analysis of the 205 surveys found that a number of factors contributed to the improvement on the mathematics WASL. Improved mathematics curriculum and instruction, more time given to mathematics instruction, greater teamwork and staff collaboration, and greater alignment of the curriculum with the EALRs were the most important factors noted by staff in the 38 schools. The respondents were familiar with the EALRs and nearly half said that mathematics instruction was provided for more than an hour each day. More preparation for the WASL and focused professional development also contributed a great deal to the improvement, according to the staff. Changes in a school’s demographic profile or in the types of students taking the test did not contribute to the improvement.

The survey results suggest that dramatic improvement can be made across the range of students taking the mathematics WASL. The results are also consistent with research conducted by OSPI and others.²⁵ However, the factors ranked on the survey are not an exhaustive list of possible reasons for improvement. For instance, the survey did not ask about changes in funding or class

²⁵ For example, see *Organizing for Success: Improving Mathematics Performance in Washington State*, OSPI, July 2000; recent RAND research on the effects of education reform in Washington state (see OSPI’s website at www.k12.wa.us/publications/ for information on both studies); and studies completed by the University of Washington for the Partnership for Learning (see www.partnership-wa.org).

size which can contribute to improved test scores. But research would suggest that without other changes within the school, providing additional funding and reducing class size may have little impact on student achievement.

Survey results from the 38 schools are shown in the tables and figures below.

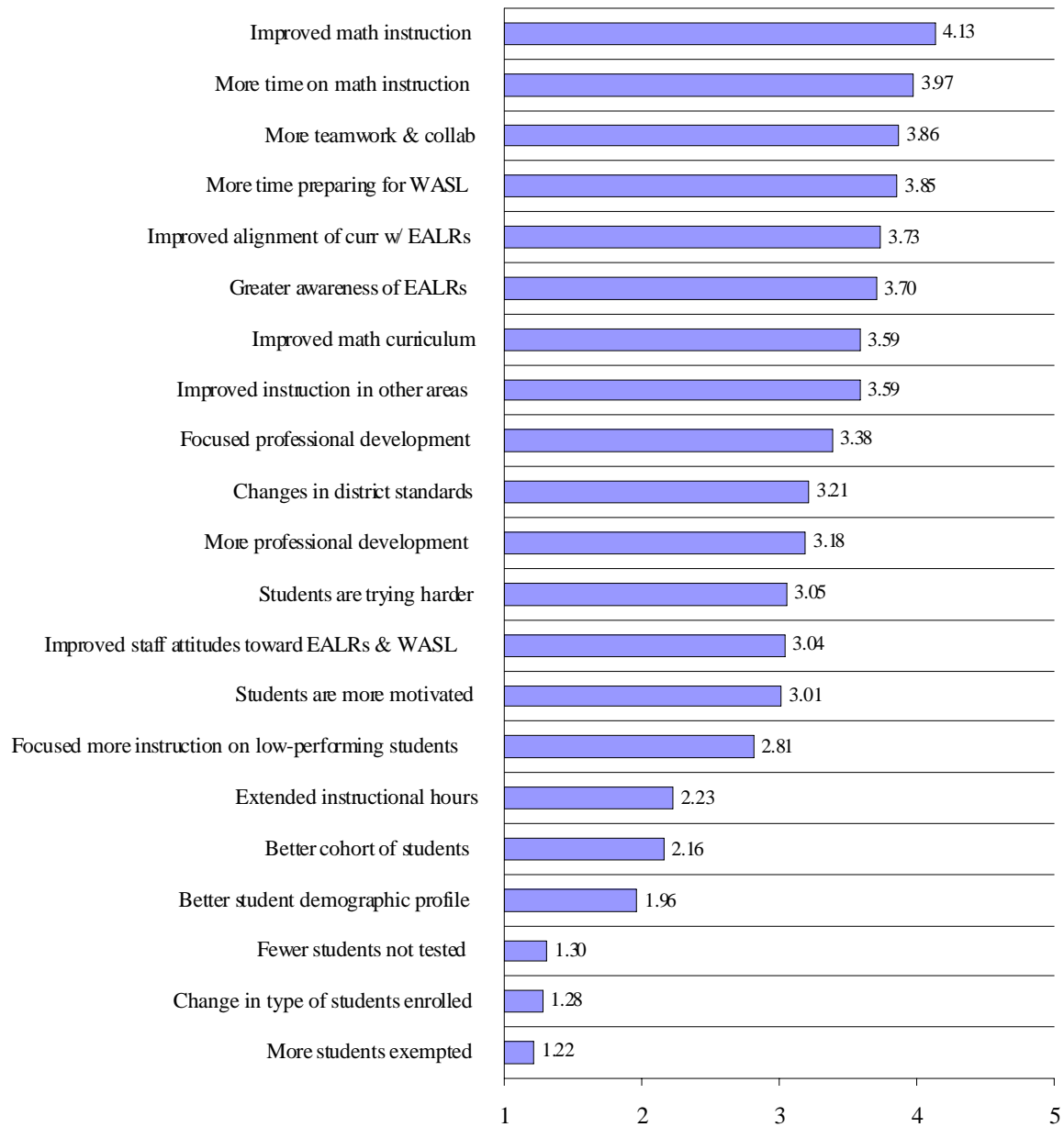
Table 5-3: Factors Contributing to Improvement on Mathematics WASL

Reason for Improvement	Mean*	% Answering Great or Very Great Extent
Improved mathematics instruction	4.13	77.7
More time on mathematics instruction	3.97	75.9
More teamwork and collaboration among staff	3.86	69.9
More time preparing for WASL	3.85	69.8
Improved alignment of curriculum with EALRs	3.73	66.2
Greater awareness of EALRs	3.70	63.1
Improved mathematics curriculum	3.59	59.4
Improved instruction in other areas	3.59	56.8
Focused professional development	3.38	50.2
Changes in district standards	3.21	42.5
More professional development	3.18	41.9
Students are trying harder	3.05	38.8
Improved staff attitudes toward EALRs and WASL	3.04	39.6
Students are more motivated	3.01	34.9
Focused more instruction on low-performing students	2.81	25.0
Extended instructional hours	2.23	19.3
Better cohort of students	2.16	17.3
Better student demographic profile	1.96	15.4
Fewer students not tested	1.30	2.0
Change in type of students enrolled	1.28	2.5
More students exempted	1.22	0.0

N=205

* Scale of 1 to 5
 1 = Little or no extent
 2 = Some extent
 3 = Moderate extent
 4 = Great extent
 5 = Very great extent

Figure 5-5: Average Score for Factors Contributing to Improvement



Scale of 1 to 5

- 1 = Little or no extent
- 2 = Some extent
- 3 = Moderate extent
- 4 = Great extent
- 5 = Very great extent

Figure 5-6: Relative Importance of Factors Contributing to Improvement

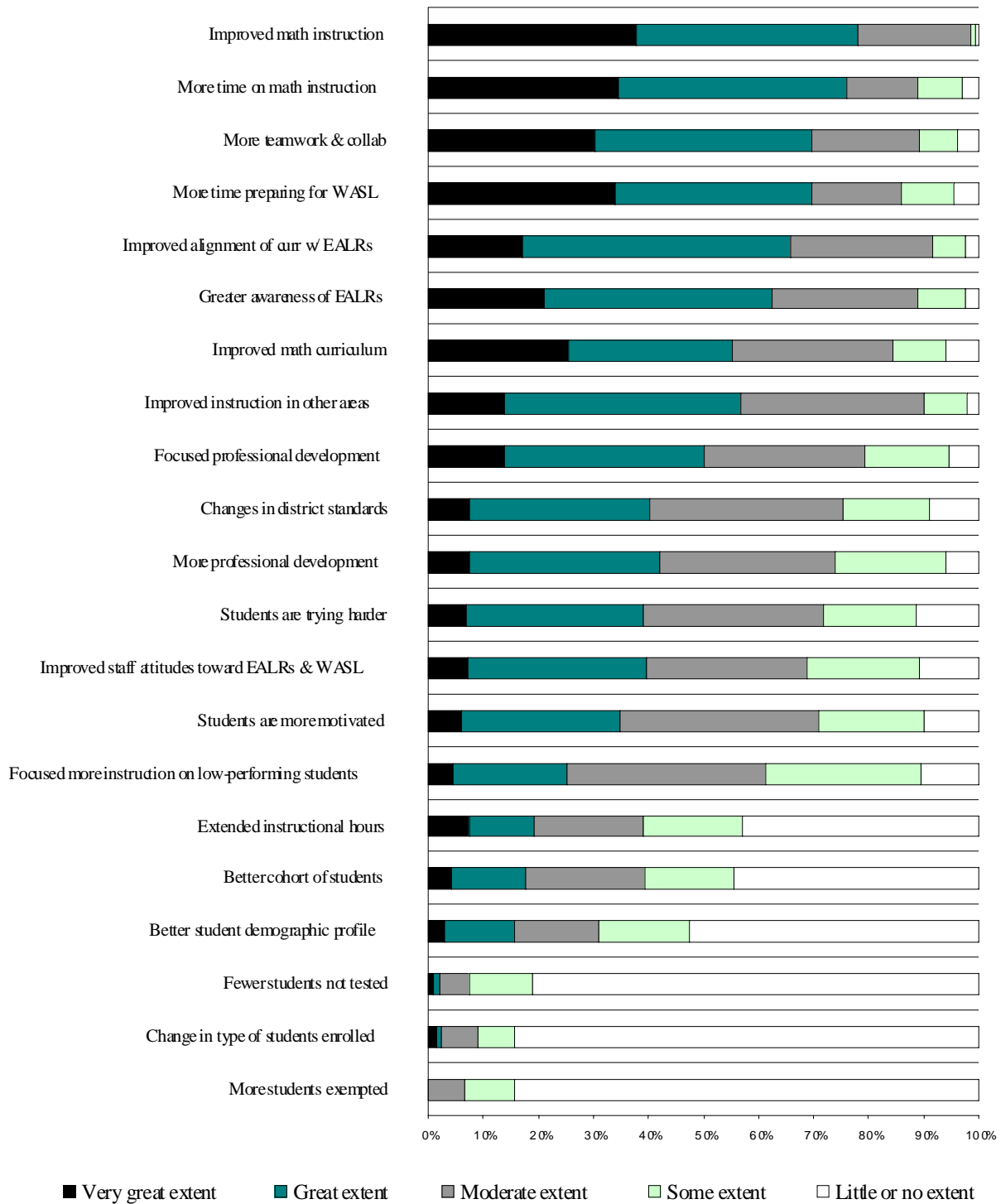


Table 5-4: Teacher Familiarity with EALRs

1. To what extent are you familiar with the mathematics EALRs for the 4 th grade?		
Not familiar at all		0.0%
Slightly familiar		6.2%
Moderately familiar		38.9%
Very familiar		54.9%
2. To what extent have you received training on math instruction related to the math EALRs during the past 2 years?		
Received little or no training		6.1%
Received some training		23.2%
Received a moderate amount of training		43.3%
Received a great amount of training		20.7%
Received a very great amount of training		6.7%
3. To what extent do you feel you were prepared to teach mathematics as it relates to the EALRs at the elementary level both last year and this year?		
	<u>Last year</u>	<u>This year</u>
Not prepared at all	0.6%	0.0%
Slightly prepared	9.6%	3.1%
Moderately prepared	33.1%	22.4%
Well prepared	45.9%	54.7%
Very well prepared	10.8%	19.9%

Table 5-5: Amount of Mathematics Instruction Time Provided

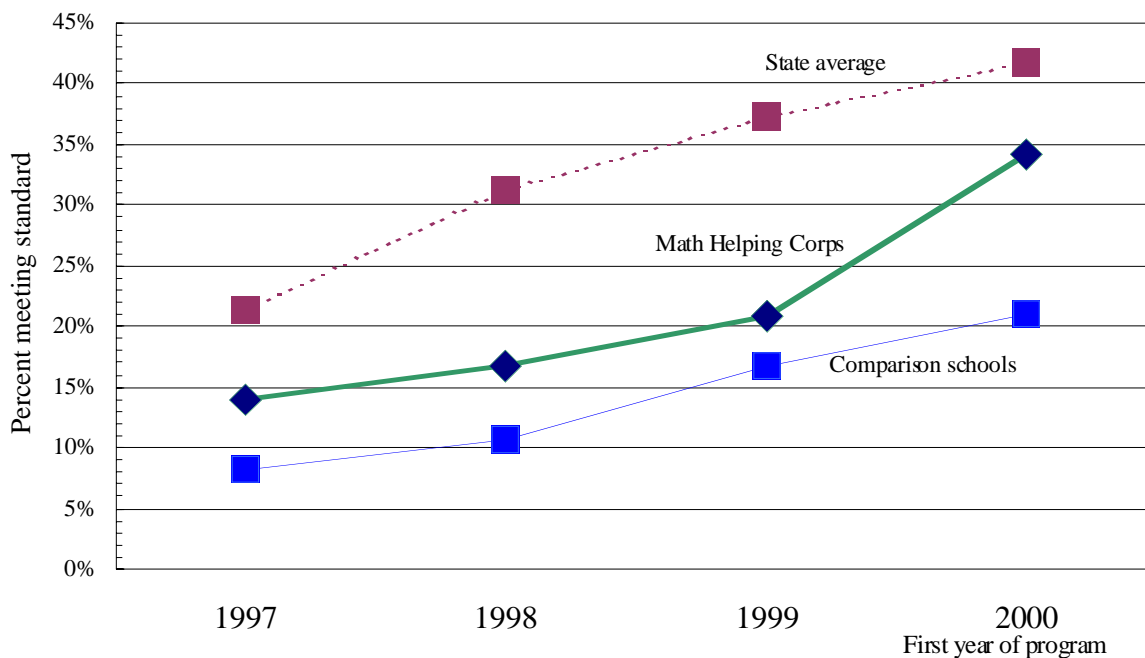
1. During the current school year, about how much time is spent on teaching and learning mathematics in your classroom?	
0 – 15 minutes a day	0.6%
16 – 30 minutes a day	0.6%
31 – 45 minutes a day	7.3%
46 – 60 minutes a day	43.9%
61 – 75 minutes a day	34.1%
76 minutes or more a day	13.4%
2. How does the amount of time spent teaching and learning mathematics in your classroom this year compare with the amount last year?	
Much less time this year	0.0%
Somewhat less time this year	1.9%
About the same amount of time both years	66.0%
Somewhat more time this year	27.0%
Much more time this year	5.0%

ANALYSIS OF SCHOOLS RECEIVING MATH HELPING CORPS ASSISTANCE

The 1999 Legislature established the Washington Helping Corps to provide direct assistance to schools with low performance in mathematics. The goal of the Helping Corps is to increase student proficiency in mathematics primarily by providing training to teachers and administrators. During the 1999-2000 school year, 16 schools from different parts of the state received assistance from 8 teachers who have mathematics expertise and were temporarily released by their districts to staff the Corps. Funding for the Helping Corps covers assistance for the 2000-2001 school year as well. Because funding was limited, however, many schools applied for but did not receive this technical assistance.

Students in 13 of the Math Helping Corps schools took the 4th grade mathematics WASL in 2000. Since the program began at the beginning of the school year, the schools had less than one year of assistance from the Corps. Nevertheless, 4th grade students in these schools performed much better than previous groups of 4th grade students at those schools, and the rate of increase was twice that of the schools that had applied for but did not receive the assistance (see Figure 5-7). While the reasons for the increase are not fully known at this time, these scores provide preliminary evidence that the technical assistance to staff can improve mathematics curriculum and instruction, which can lead to improved mathematics scores on the WASL.

Figure 5-7: Large Increase in Student Performance in Math Helping Corps Schools



Chapter 6

OTHER ANALYSES OF THE TEST

OSPI examined several more issues that have been raised about the assessment: The amount of time students need to take the mathematics WASL, the way the test is administered, and additional issues that still need to be studied. This chapter describes the results of this work.

ESTIMATED TESTING TIME PER SESSION

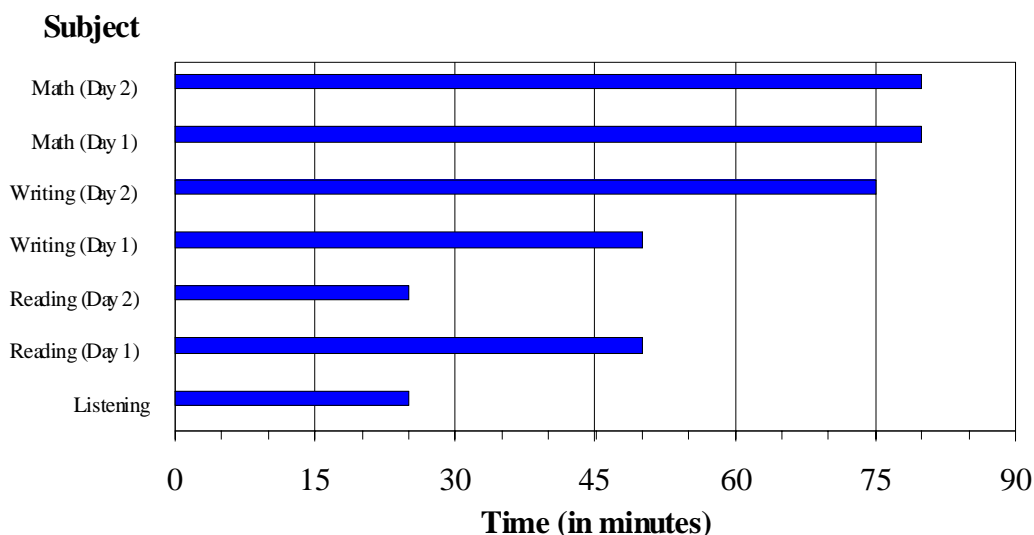
Subject area WASL tests are not timed. Students are to have as much time as they need to work on the tests, although some schools impose a time limit on the students. Professional judgment should determine when a student is no longer “productively engaged.” When the majority of students have finished, testing guidelines suggest that those still working may be moved to new location to finish. Teachers’ knowledge of students’ work habits or special needs may suggest that some students who work very slowly should be tested separately or grouped with similar students for the entire assessment.

For planning purposes, the estimated testing times required *for most students* are given in the test administration instructions. The times are estimates for actual testing time. Additional time is required to distribute and collect materials and cover the directions for test-taking. Testing sessions need not follow on consecutive days. Individual sessions should not be split but may be spaced with one or more days between testing periods.

Table 6-1 and Figure 6-1 show that the estimated time to complete the mathematics sessions of the WASL is the longest of all the tested subjects. Observations and anecdotal reports suggest that the actual time provided to students is often longer than these estimated times. Nearly all the experts OSPI consulted about the assessment agreed that the mathematics test, as currently constructed, is too long for 4th grade students.

Table 6-1: Estimated Testing Times for Grade 4 WASL

Session	Subject	Estimated Time
1	Listening	25 minutes
	Reading (Day One)	50 minutes
2	Reading (Day Two)	25 minutes
	Writing (Day One)	50 minutes
3	Writing (Day Two)	75 minutes
4	Mathematics (Day One) with tools	80 minutes
5	Mathematics (Day Two) without tools	80 minutes

Figure 6-1: Estimated Testing Time Per Session

To address this issue, OSPI and its contractor will consider ways to make the testing period more appropriate for 4th graders. Options that will be considered include:

- _ Having fewer items on the test, possibly by having more than one strand assessed on each item.
- _ Breaking the assessment into smaller periods of time (e.g., having three or four shorter sessions rather than two long ones).
- _ Limiting or standardizing the amount of time students are allowed to take the test.

TEST ADMINISTRATION

In addition to timing, other issues have been raised that suggest the test may not be administered under “standardized” conditions. Some educators say that they lack clear guidance about testing conditions, such as the types of materials that can be kept on the walls of the classroom and the kind of guidance or encouragement they can offer students during the test. Some teachers provide breaks and snacks, while others do not. Some teachers allow students to leave when they are finished, while others require all students to stay in the room until all students are finished. Moreover, schools have the discretion to administer the tests in whatever order they wish (e.g., mathematics can come before or after the reading WASL) and may administer more than one test in a single day. Thus, OSPI will take steps to improve the guidelines for administering the test and will consider using a more standardized approach.

ADDITIONAL RESEARCH NEEDED

As discussed above, anecdotal information indicates that the amount of time most students require to complete the assessment is more than the estimated time. Research is needed to determine the actual length of time that most students need to complete the mathematics and other assessments. When this information is available, adjustments can be made in the test administration booklets to help educators plan their assessment schedule. The information will also be helpful when designing future tests.

OSPI did not study how individual students approach and respond to test items. Having students “think aloud” while taking the test provides insights into the cognitive processes students use when approaching a test item. Analyses of how students approach each item are being conducted by independent researchers but have not completed. More research needs to be done in this area to strengthen the validity of the WASL.

Finally, reliable data on the readability of the mathematics test is lacking. Teachers on the test development committees reviewed the vocabulary and overall reading level of test items to ensure they were below grade level. Nevertheless, some people have expressed concerns that the reading level of the test is too difficult, particularly for students whose primary language is not English. If the reading level is too high, the test may confuse reading ability with mathematics ability.

Preliminary analyses suggest that the reading level of the mathematics WASL is at grade 4. However, the current models used to assess readability are not designed to evaluate mathematics items and result in an inflated reading level. These models count the number of syllables in a word to determine readability, even though some mathematical terms have many syllables (e.g., multiplication) but are content-specific vocabulary that should be taught by the 4th grade. The current models also convert numbers to words (150 becomes “one-hundred fifty”) which inflates the reading level by inflating the number of syllables in a word.²⁶ In addition, individual test items do not provide enough “text” to analyze their readability. Finally, the readability of graphs, tables, and figures is not easily quantified. When better models are available that take into consideration mathematics terminology, numbers, and formats, more accurate readability analyses can be conducted.

²⁶ OSPI’s reading specialist investigated the readability of the entire test. One of the analyses made adjustments for numbers in the text and found the reading level to be at grade 4.4.

Chapter 7

SUMMARY AND NEXT STEPS

The Legislature required that performance standards on the WASL be set at internationally competitive levels. After the development of the assessments, concerns were raised regarding the level of difficulty of the 4th grade mathematics test. OSPI staff and independent experts conducted an objective study to determine the student developmental level required to achieve the 4th grade standard successfully and to investigate other issues raised about the test.

The various analyses conducted for this study yielded a number of conclusions.

- _ The processes used to develop the test and set the standard were sound and consistent with national test development standards.
- _ The level needed to meet the standard was within the developmental capability of well taught, hard working students in the 4th grade.
- _ A few items were beyond the developmental level of most 4th grade students. However, there was ample opportunity for students to meet the standard by doing well on the other items.
- _ The performance of 4th grade students on the mathematics WASL has improved over time. Improvement has occurred statewide and regardless of gender, ethnic/racial group, and socioeconomic status. Performance in some schools has improved dramatically. Such progress is evidence that the developmental level needed to meet the standard is appropriate and that students are capable of meeting high standards when given exposure to appropriate curriculum and instruction.
- _ A few items on the test were not aligned with the 4th grade EALR benchmarks. Most of these were not aligned because they required students to “create a plan,” which is a task associated with the 7th grade EALRs. However, students have performed well on such items.
- _ On average, 55 percent of the test items required computation skills, and calculators were allowed on half (27.5%) of these. The percentage of items assessing the various mathematical strands was balanced and conformed to test specifications. Some items included concepts related to four or more strands.
- _ Some items had formats that influenced students’ performance. Items with difficult or unusual formats may mask students’ true cognitive and mathematics abilities.

- _ The mathematics portion of the WASL, as currently constructed, is too long for 4th grade students. The estimated time to take each part of the test is the longest of all subjects (80 minutes each session), and students may need even more time to complete the test.
- _ The test is not administered under the same conditions. Some schools impose a time limit on the students, even though the test is not to be timed. Some educators lack clear guidance about testing conditions, such as the types of materials that can be kept on classroom walls and the kind of guidance or encouragement they may offer students during the test. Other differences in test administration were identified.

This study is part of OSPI's ongoing review of the state's assessment system and has identified some areas that need to be changed or reviewed in the 4th grade mathematics WASL. OSPI will work with its contractors and technical advisors to:

- _ Make sure each test item is grade-appropriate and aligned with the EALR benchmarks.
- _ Revise some items to reduce the level of format complexity and avoid confusion.
- _ Consider ways to shorten the time of the test sessions and/or shorten the test while maintaining appropriate test reliability.
- _ Provide more guidance to improve the administration of the test.

The study identified other issues that should be studied to improve the quality of the test. More information is needed to determine (1) the amount of time most students need to complete the mathematics and other assessments, (2) what and how individual students think while taking test items, and (3) the readability of the mathematics test.

Appendix A

4TH GRADE MATHEMATICS EALRS, STRANDS, AND LEARNING TARGETS

As part of the test development process, test and item specifications were developed for each content area so they are aligned with the EALRs. This appendix lists the components and benchmarks for the 4th grade mathematics EALRs and the strands and learning targets related to the EALRs.²⁷

EALR COMPONENTS AND BENCHMARKS

1. Student Understands and Applies the Concepts and Procedures of Mathematics

1.1 Understand and Apply Concepts and Procedures from Number Sense

- 1.1.1 use objects, pictures, or symbols to demonstrate understanding of whole and fractional numbers, place value in whole numbers, and properties of the whole number system
- 1.1.2 identify, compare, and order whole numbers and simple fractions
- 1.1.3 show understanding of whole number operations (addition, subtraction, multiplication and division) using blocks, sticks, beans, etc.
- 1.1.4 add, subtract, multiply and divide whole numbers
- 1.1.5 use mental arithmetic, pencil and paper, or calculator as appropriate to the task involving whole numbers
- 1.1.6 identify situations involving whole numbers in which estimation is useful
- 1.1.7 use estimation to predict computation results and to determine the reasonableness of answers, for example, estimating a grocery bill

1.2 Understand and Apply Concepts and Procedures from Measurement

- 1.2.1 understand concepts of perimeter, area, and volume
- 1.2.2 use directly measurable attributes such as length, perimeter, area, volume/capacity, angle, weight/mass, money, and temperature to describe and compare objects
- 1.2.3 understand that measurement is approximate
- 1.2.4 know how to estimate to predict and to determine when measurements are reasonable, for example, estimating the length of the playground by pacing it off
- 1.2.5 understand the benefits of using standard units of measurement for measuring length, area, and volume

²⁷ More information about the EALRs is found on OSPI's website: www.k12.wa.us/reform/EALR/default.asp. Information about the test and item specifications are on OSPI's website: www.k12.wa.us/assessment/assessproginfo.

- 1.2.6 know appropriate units of measure for time, money, length, area, volume, mass, and temperature
- 1.2.7 use appropriate tools for measuring time, money, length, area, volume, mass, and temperature

1.3 Understand and Apply Concepts and Procedures from Geometric Sense

- 1.3.1 use shape and size to identify, name, and sort geometric shapes
- 1.3.2 recognize geometric shapes in the surrounding environment, for example, identify rectangles within windows
- 1.3.3 describe the relative location of objects relative to each other on grids or maps
- 1.3.4 understand concepts of parallel and perpendicular
- 1.3.5 understand concepts of symmetry, congruence, and similarity
- 1.3.6 understand and construct simple geometric transformations using slides, flips, and turns
- 1.3.7 construct simple shapes using appropriate tools such as a straightedge or a ruler

1.4 Understand and Apply Concepts and Procedures from Probability and Statistics

- 1.4.1 understand the difference between certain and uncertain events
- 1.4.2 know how to list all possible outcomes of simple experiments
- 1.4.3 understand and use experiments to investigate uncertain events
- 1.4.4 know that data can be represented in different forms such as tabulations of events, objects, or occurrences
- 1.4.5 can collect data in an organized way
- 1.4.6 organize and display data in numerical and graphical forms such as tables, charts, pictographs, and bar graphs
- 1.4.7 use different measures of central tendency such as “most often” and “middle” describing a set of data
- 1.4.8 predict outcomes of simple activities and compares the predictions to experimental results
- 1.4.9 understand and make inferences based on experimental results using coins, number cubes, spinners, etc.

1.5 Understand and Apply Concepts and Procedures from Algebraic Sense

- 1.5.1 recognize, create and extend patterns of objects and numbers using a variety of materials such as beans, toothpicks, pattern blocks, calculator, cubes, or colored tiles
- 1.5.2 understand the use of guess and check in the search for patterns
- 1.5.3 represent number patterns symbolically, for example, using tiles, boxes, or numbers
- 1.5.4 use standard notation in reading and writing open sentences, for example, $3 \cdot _ = 18$
- 1.5.5 evaluate simple expressions using blocks, sticks, beans, pictures, etc.
- 1.5.6 solve simple equations using blocks, sticks, beans, pictures, etc.

2. Student Uses Mathematics to Define and Solve Problems

2.1 Investigate Situations

- 2.1.1 search for patterns in simple situations
- 2.1.2 use a variety of strategies and approaches
- 2.1.3 recognize when information is missing or extraneous
- 2.1.4 recognize when an approach is unproductive and tries a new approach

2.2 Formulate Questions and Define the Problem

- 2.2.1 identify questions to be answered in familiar situations
- 2.2.2 define problems in familiar situations
- 2.2.3 identify the unknowns in familiar situations

2.3 Construct Solutions

- 2.3.1 organize relevant information
- 2.3.2 select and use appropriate mathematical tools
- 2.3.3 apply appropriate methods, operations, and processes to construct a solution

3. Student Uses Mathematical Reasoning

3.1 Analyze Information

- 3.1.1 interpret and compare information in familiar situations
- 3.1.2 validate thinking using models, known facts, patterns, and relationships

3.2 Predict Results and Make Inferences

- 3.2.1 make conjectures and inferences based on analysis of familiar problem situations

3.3 Draw Conclusions and Verify Results

- 3.3.1 test conjectures by finding examples to support or contradict them
- 3.3.2 support arguments and justify results based on own experiences
- 3.3.3 check for reasonableness of results
- 3.3.4 reflect on and evaluates procedures and results in familiar situations

4. Student Communicates Knowledge and Understanding in Both Everyday and Mathematical Language

4.1 Gather Information

- 4.1.1 follow a plan for collecting information
- 4.1.2 use reading, listening, and observation skills to access and extract mathematical information from a variety of sources such as pictures, diagrams, physical models, classmates, oral narratives, and symbolic representations
- 4.1.3 use available technology to browse and retrieve mathematical information from a variety of sources

4.2 Organize and Interpret Information

- 4.2.1 organize and clarify mathematical information in at least one way—reflecting, verbalizing, discussing, or writing

4.3 Represent and Share Information

- 4.3.1 express ideas using mathematical language and notation such as physical or pictorial models, tables, charts, graphs, or symbols
- 4.3.2 express mathematical ideas to familiar people in everyday language

5. Student Understands How Mathematical Ideas Connect Within Mathematics, to Other Subject Areas, and to Real-life Situations

5.1 Relate Concepts and Procedures Within Mathematics

- 5.1.1 connect conceptual and procedural understandings among familiar mathematical content areas
- 5.1.2 recognize equivalent mathematical models and representations in familiar situations

5.2 Relate Mathematical Concepts and Procedures to Other Disciplines

- 5.2.1 recognize mathematical patterns and ideas in familiar situations in other disciplines
- 5.2.2 use mathematical thinking and modeling in familiar situations in other disciplines
- 5.2.3 describe examples of contributions to the development of mathematics such as the contributions of women, men, and different cultures

5.3 Relate Mathematical Concepts and Procedures to Real-life Situations

- 5.3.1 give examples of how mathematics is used in everyday life
- 5.3.2 identify how mathematics is used in career settings

STRANDS AND LEARNING TARGETS

The following *learning targets* are intended to summarize the knowledge or EALRs (benchmarks, content, or process examples). These benchmarks are identified by numbers in parentheses of the related EALR.

Number Sense (NS)

- **NS01** Identify and illustrate whole numbers and fractions in a variety of forms and representations, using pictures, models, and symbols. (1.1.1)
- **NS02** Demonstrate an understanding of place value and magnitude in identifying, ordering, and comparing whole numbers and common or simple fractions. (1.1.1, 1.1.2)
- **NS03** Add, subtract, multiply, and divide whole numbers; demonstrate an understanding of whole number operations and fraction operations at the concrete level. (1.1.3, 1.1.4)
- **NS04** Determine appropriateness of estimation and use estimation to predict computation results and determine reasonableness of answers. (1.1.6, 1.1.7)
- **NS05** Identify and illustrate properties of whole numbers and break down (decompose), combine, compare, pattern/sequence, and order numbers. (1.1.1, 1.1.2)

Measurement (ME)

- **ME01** Describe and compare objects and measurable attributes of objects (such as length, perimeter, area, volume or capacity, angle, weight, money, and temperature) in standard units. (1.2.2)
- **ME02** Select, use, and evaluate appropriate instruments, units (standard or nonstandard), and procedures for measuring time, money, length, area, volume, weight, and temperature. (1.2.6, 1.2.7)
- **ME03** Use estimation to predict or determine the reasonableness of measurements and to obtain reasonable approximations. (1.2.4)
- **ME04** Demonstrate an understanding of the appropriate uses of standard and nonstandard units of measure and the approximate nature of measurement. (1.2.5, 1.2.3)

Geometric Sense (GS)

- **GS01** Identify, describe, sort, and compare geometric figures using their attributes; describe how geometric shapes and objects in the surrounding environment are related; and construct geometric figures. (1.3.7, 1.3.1, 1.3.2)
- **GS02** Identify and describe the relative location of objects to one another; identify and describe the location of objects on a location grid (map, grid, number line); identify and construct simple geometric transformations using slides, flips, and turns. (1.3.3, 1.3.6)

- **GS03** Identify, describe, and compare parallel, perpendicular, and intersecting lines, as well as congruent, symmetrical, and similar figures, in two-dimensional and real-world constructions. (1.3.4, 1.3.5)

Probability and Statistics (PS)

- **PS01** Predict, show, and evaluate the possible outcomes and probabilities of simple experiments and activities; distinguish between certain and uncertain events; and compare predictions to experimental results. (1.4.1, 1.4.2, 1.4.3, 1.4.8)
- **PS02** Identify, describe, and evaluate methods for the effective collection of data. (1.4.5)
- **PS03** Collect, organize, analyze, and display data in graphs, tables, charts, and other pictorial representations (e.g., icons); make and evaluate inferences from data and experimental results. (1.4.6, 1.4.9)
- **PS04** Identify, find, and use defined measures of central tendency (mean, median, mode) and other characteristics to describe a set or sets of data and sample populations. (1.4.7)

Algebraic Sense (AS)

- **AS01** Recognize, create, and extend patterns of objects and numbers. (1.5.1)
- **AS02** Identify and use appropriate symbols/notation to represent number patterns and operations, and to translate problem situations into mathematical symbols. (1.5.3)
- **AS03** Set up and solve simple equations at the concrete or pictorial level. (1.5.6)

Solving Problems (SP)

- **SP01** Use, modify, create, and evaluate strategies and approaches to conduct explorations and perform operations. (2.3.3)
- **SP02** Formulate questions; define problems; and identify patterns, questions to be answered, missing or unnecessary data, and unknowns. (2.2.1, 2.2.3, 2.1.3)
- **SP03** Collect needed information, select and use tools, use a variety of strategies, and apply concepts and procedures in constructing solutions. (2.1.2, 2.3.2)

Reasoning Logically (RL)

- **RL01** Compare and contrast information, and interpret information from a variety of sources. (3.1.1)
- **RL02** Identify and use models, known facts, patterns, relationships, counterexamples, and deductive and inductive reasoning to validate thinking, support arguments, and evaluate procedures and results. (3.3.1, 3.1.2, 3.3.4)

- **RL03** Make inferences, predictions, and conclusions based on analysis of problem situations. (3.2.1)

Communicating Understanding (CU)

- **CU01** Create a plan for collecting information. (4.1.1)
- **CU02** Use reading, listening, and observation skills to gather, extract, and interpret mathematical information from a variety of sources—pictures, diagrams, models, text, symbolic representations, and technology. (4.1.2, 4.1.3)
- **CU03** Represent, organize, and express mathematical information, understandings, and ideas using models, tables, charts, graphs, written reflections, and algebraic notation, and explain these ideas in ways appropriate to a given audience. (4.2.1, 4.3.1, 4.3.2)

Making Connections (MC)²⁸

- **MC01** Link conceptual and procedural understandings among the areas of number sense, measurement, geometric sense, probability and statistics, and algebraic sense. (5.1.1)
- **MC02** Use, create, and evaluate equivalent graphical, numerical, physical, algebraic, geometric, and verbal mathematical models and representations. (5.1.2)
- **MC03** Identify and apply mathematical thinking, modeling, patterns, and ideas in other disciplines, real-life situations, and job-related applications. (5.2.2, 5.3.1, 5.3.2)

The following learning targets have been identified or recommended as bases for classroom-based assessment activities:

Demonstrate ability to use mental arithmetic, pencil and paper, and calculator as appropriate; choose the appropriate strategy. (1.1.5)

Use the guess and check strategy in searching for and evaluating patterns. (1.5.2)

²⁸ Mathematics relations and applications within real-world situations, other disciplines, or within mathematics permeate the test. Whenever possible, these relations and applications will be made. Specific items, however, will also be constructed to assess students' ability to use, identify, or construct such applications or relations.

Appendix B

DETAILED ANALYSIS OF TEST ITEMS

The tables on the following pages provide a detailed analysis of each item on the operational tests used in 1998, 1999, and 2000. Each test has 40 items, with 20 given in each session. The table gives the following information.

- _ The first column lists the items in the order they appear on the test. An asterisk (*) indicates which items allow the use of calculators. Only the first session allows calculator use, so half (20 of 40) items allow calculators.
- _ In the next set of columns, an “x” indicates the type of item (i.e., multiple-choice, short-answer, or extended-response), and the maximum number of points that can be scored on the item is given.
- _ The “difficulty rating” is the Rasch score for that item (see Appendix F for more information on Rasch scores). Higher scores indicate a more difficult item, while lower scores (e.g., negative scores) indicate a less difficult item.
- _ An C indicates whether the item may utilize computation skills in some part of the item.
- _ The nine mathematics strands—five content strands and four process strands—are listed in the middle set of columns. An X indicates which strand the item assesses; an item assesses only one strand. However, concepts from other strands are usually included in the item as well. Concepts from these other (“secondary”) strands are indicated with a small dot.
- _ The final set of columns on the right shows the percentage of students that scored each number of points on the item. These data were only available for the 1998 and 1999 tests.
- _ Cumulative totals or averages for each column are shown at the bottom of each table.

Table B-1: Analysis of Items on 4th Grade Mathematics WASL, Spring 1998

Item	Item Type/Points				Difficulty Rating ¹	Computation May Apply	Strand Assessed								Points Scored (percent of all students)									
	Multiple-choice	Short-answer	Extended-response	Total Points Given			Content					Process			Omitted or other	0	1	2	3	4				
							Number Sense	Measurement	Geometric Sense	Probability/Statistics	Algebraic Sense	Solves Problems	Reasons Logically	Communications							Makes Connections			
1*	x			1	-0.76			•				X	•				3.7	28.2	68.1					
2*	x			1	-2.05	C	•								X		3.7	10.9	85.4					
3*		x		2	-0.70		•	X				•	•	•			3.9	9.5	38.2	48.4				
4*	x			1	-1.93	C	X						•				3.6	10.9	85.5					
5*		x		2	1.00	C	X						•	•			5.4	69.5	8.5	16.6				
6*	x			1	-1.01	C	•			X			•				3.8	24.5	71.7					
7*	x			1	0.05			X	•				•				3.7	40.5	55.8					
8*		x		2	-0.87				X				•	•			4.1	10.4	22.2	63.4				
9*	x			1	0.72	C	X						•				4.0	58.1	38.0					
10*			x	4	0.92		•					•	•	X			9.2	37.7	21.4	10.7	7.6	13.4		
11*	x			1	-0.54	C	•			X			•				3.9	29.9	66.2					
12*	x			1	-0.88				X			•					3.8	26.3	69.8					
13*	x			1	0.33	C	•		•				•	•	X		4.1	49.7	46.2					
14*		x		2	0.74	C	•					X		•			8.5	38.7	29.7	23.1				
15*	x			1	0.66		X										5.3	53.9	40.8					
16*	x			1	0.31								X				4.9	46.3	48.8					
17*	x			1	0.65	C	•	•		X			•				4.4	55.6	40.0					
18*			x	4	0.33		•						X	•			8.0	20.3	6.7	29.6	19.2	16.2		
19*	x			1	1.02	C	•			X							4.5	64.0	31.5					
20*		x		2	0.50	C	•				X		•				6.4	45.2	14.4	34.0				
Total	13	5	2	31	-0.08	11	4	2	2	3	2	2	2	1	2									
End of First Session																								
21	x			1	-0.53	C	•			X		•					4.1	31.3	64.6					
22	x			1	0.26	C	•	•				•	X				4.3	47.7	48.0					
23		x		2	0.39		•						X	•			4.5	42.4	18.9	34.2				
24	x			1	1.49	C	X	•		•			•	•			4.3	70.3	25.4					
25		x		2	-1.29			X					•	•			4.9	7.0	6.6	81.5				
26		x		2	-0.52									X			7.4	15.9	16.2	60.5				
27	x			1	0.70				X			•	•				4.3	56.6	39.1					
28	x			1	-0.60	C	•			X			•				4.2	30.6	65.2					
29		x		2	-0.08		•			X			•	•			5.6	22.6	34.4	37.3				
30	x			1	0.00				X				•				4.2	40.5	55.3					
31	x			1	0.68	C	•				X		•				4.6	58.6	36.8					
32		x		2	1.20						•	X	•	•			9.1	64.8	7.8	18.3				
33	x			1	-0.83		•	X					•	•			4.2	27.8	68.0					
34		x		2	1.05	C	•	•					•		X		6.9	48.3	29.8	15.0				
35		x		2	0.78				X				•	•			7.0	47.4	24.7	20.9				
36	x			1	0.20	C	•			X			•				4.5	45.4	50.1					
37			x	4	0.56	C	•				•	X	•	•	•		9.8	32.3	19.7	9.7	7.7	20.8		
38	x			1	1.24	C	•	X	•						•		4.9	65.6	29.5					
39		x		2	-1.28		X						•				5.6	6.5	14.2	73.7				
40	x			1	0.21		•							X			5.4	43.3	51.3					
Total	11	8	1	31	0.18	9	2	3	3	2	3	2	2	2	1									
End of Second Session																								
2-day total	24	13	3	62	0.05	20																		
	Total assessed (X)						6	5	5	5	5	4	4	3	3									
	Secondary strand (•)						23	6	3	10	2	8	24	17	2									
	Total points						8	7	7	6	6	9	8	7	4									
Average difficulty ¹						-0.01	-0.14	-0.05	.00	.06	.44	.32	.20	-.22										

¹ Rasch scores: Higher scores indicate a more difficult item; lower scores (e.g., negative scores) indicate a less difficult item.

* = Use of technology (calculators) allowed.

X = Primary stand assessed on the item.

• = Secondary strand (item also includes an understanding of concepts in this strand).

Table B-2: Analysis of Items on 4th Grade Mathematics WASL, Spring 1999

Item	Item Type/Points				Difficulty Rating ¹	Computation May Apply	Strand Assessed								Points Scored (percent of all students)										
	Multiple-choice	Short-answer	Extended-response	Total Points Given			Content					Process			Omitted or other	0	1	2	3	4					
							Number Sense	Measurement	Geometric Sense	Probability/Statistics	Algebraic Sense	Solves Problems	Reasons Logically	Communications							Makes Connections				
1*	x			1	-0.76			•				X	•			4.3	27.8	67.9							
2*	x			1	-2.05	C	•							X		4.3	9.8	85.9							
3*		x		2	0.47			X	•				•			5.4	33.6	29.3	31.7						
4*	x			1	-0.39				X				•			4.3	33.5	62.2							
5*		x		2	1.45					•			X			7.3	56.6	25.6	10.5						
6*			x	4	1.02	C	•			•			X	•		7.3	33.2	14.7	19.7	18.0	7.1				
7*	x			1	0.19	C	•			•			•	X		4.7	44.2	51.1							
8*		x		2	1.20	C	•			•			•	•	X	6.4	50.0	29.6	14.0						
9*	x			1	0.72	C	X			•			•			4.8	56.2	39.0							
10*		x		2	0.85				X				•			7.6	39.2	36.0	17.2						
11*	x			1	-0.54	C	•			X			•			4.8	28.1	67.1							
12*	x			1	-0.88			X			•					4.7	22.3	73.0							
13*	x			1	-1.16	C	•			•	X		•			6.6	17.6	75.8							
14*	x			1	1.14		X						•			5.4	62.1	32.5							
15*			x	4	0.42		•	•		•	•	X	•			6.6	9.2	37.1	16.8	18.3	12.0				
16*	x			1	0.31					•		X				5.3	43.5	51.2							
17*	x			1	0.31	C	•			X			•	•		5.2	45.7	49.1							
18*	x			1	-1.35		X							•		6.0	16.7	77.3							
19*	x			1	1.14	C	•	•					•	X		5.0	62.7	32.3							
20*		x		2	0.50	C	•			X			•			6.5	41.8	16.0	35.7						
Total	13	5	2	31	0.13	10	3	1	2	2	3	1	2	2	4										
End of First Session																									
21	x			1	-0.53	C	•			X	•					5.1	31.8	63.1							
22	x			1	-1.33		•			X			•			4.4	18.0	77.6							
23		x		2	0.39		•	•				X	•			5.1	40.7	19.1	35.1						
24	x			1	1.49	C	X	•		•			•	•		4.7	69.4	25.9							
25		x		2	-1.38		X						•			5.1	6.4	6.5	82.0						
26		x		2	0.39			•	•	X			•	•		5.4	28.9	43.6	22.1						
27	x			1	0.70				X	•			•			4.8	54.9	40.3							
28	x			1	-0.60	C	•			X			•			4.6	31.1	64.3							
29	x			1	0.37		X	•					•	•		4.9	47.8	47.4							
30			x	4	1.13	C	•		•			X		•		5.4	14.4	41.0	19.7	17.5	2.0				
31	x			1	-0.32	C	•			X			•			4.6	34.5	60.9							
32		x		2	1.20					•	X		•	•		8.4	66.3	7.2	18.1						
33	x			1	0.18	C	•	X					•			5.2	43.8	51.0							
34		x		2	-0.06		•						X	•		5.6	12.5	57.6	24.3						
35	x			1	-0.46	C	•	X					•	•		5.1	32.1	62.8							
36		x		2	0.36		•	•	X				•	•		8.3	32.9	21.0	37.8						
37		x		2	1.78	C	•		X				•	•		9.4	74.9	5.4	10.3						
38	x			1	-0.20	C	•	X								7.3	32.6	60.1							
39		x		2	-0.76	C	•	X						•		5.5	11.3	20.9	62.3						
40	x			1	0.21		•			•				X		5.6	40.2	54.2							
Total	11	8	1	31	0.13	10	3	4	3	3	2	2	2	1	0										
End of Second Session																									
2-day total	24	13	3	62	0.13	20																			
	Total assessed (X)						6	5	5	5	5	3	4	3	4										
	Secondary strand (•)						24	8	3	10	2	4	13	26	2										
	Total points						7	7	7	7	6	7	9	7	5										
Average difficulty ¹						.17	-.15	.31	-.08	-.41	.52	.27	.89	.12											

¹ Rasch scores: Higher scores indicate a more difficult item; lower scores (e.g., negative scores) indicate a less difficult item.
 * = Use of technology (calculators) allowed.
 X = Primary stand assessed on the item.
 • = Secondary strand (item also includes an understanding of concepts in this strand).

Table B-3: Analysis of Items on 4th Grade Mathematics WASL, Spring 2000

(Note: Student scores were not available for this assessment.)

Item	Item Type/Points				Computation May Apply	Strand Assessed									
	Multiple-choice	Short-answer	Extended-response	Total Points Given		Content					Process				
						Number Sense	Measurement	Geometric Sense	Probability/Statistics	Algebraic Sense	Solves Problems	Reasons Logically	Communications	Makes Connections	
1*	x			1	C	•				X	•				
2*	x			1	C	•	X				•	•	•		
3*	x			1	C	•			X				•	•	
4*	x			1				X					•		
5*		x		2					•				X		
6*		x		2	C	•	X						•		
7*	x			1	C	•			•				•	X	
8*			x	4	C	•					X	•	•		
9*	x			1	C	•		X					•		
10*		x		2					X				•		
11*	x			1	C	•		X			•	•	•		
12*		x		2	C	•				X	•		•		
13*	x			1	C	•			•	X			•		
14*	x			1		X							•		
15*		x		2	C	•							X		
16*			x	4		•	•				X		•		
17*	x			1		•			X		•	•			
18*	x			1		X								•	
19*	x			1	C	•	•						•	X	
20*		x		2		•			•				•	•	
Total	12	6	2	32	12	2	2	3	3	3	1	1	3	2	
End of First Session															
21	x			1				X	•				•		
22	x			1		•			X				•		
23		x		2	C	X							•		
24	x			1	C	•			•				X		
25		x		2	C	•				X	•	•	•		
26		x		2		•	•				•	X	•		
27	x			1	C	•						X			
28			x	4	C	•					X		•		
29	x			1		X	•						•		
30	x			1	C	•					X	•			
31	x			1	C	•				X			•		
32		x		2	C	•	•	•			•	•		X	
33	x			1	C	•	X					•			
34	x			1	C	•	•		X				•		
35	x			1	C	•	X					•	•		
36		x		2		•	•	X					•		
37		x		2		•						•	•	X	
38	x			1	C	•	X								
39		x		2	C	X			•			•	•		
40	x			1	C	•			•				•	X	
Total	12	7	1	30	14	3	3	2	2	2	2	2	1	3	
End of Second Session															
Grand Totals	24	13	3	62	26										
	Total assessed (X)					5	5	5	5	5	3	3	4	5	
	Secondary strand (•)					30	8	1	7	1	8	12	28	3	
Total points					7	6	6	6	7	9	7	7	7		

* = Use of technology (calculators) allowed.

X = Primary stand assessed on the item.

• = Secondary strand (item also includes an understanding of concepts in this strand).

Appendix C

TYPES OF ASSESSMENTS

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 subject. In order to decide "how well" a student has done, some external frame of reference is needed. When we compare a student's performance to a desired performance, this is considered a criterion-referenced interpretation. When we compare a student's performance to the performance of other students, this is considered a norm-referenced interpretation.

Criterion-referenced tests are intended to provide a measure of the degree to which students have achieved a desired set of learning targets that have been identified as appropriate for a given grade or developmental level. Careful attention is given to making certain that the items on the test represent only the desired learning targets and that there are sufficient items for each learning target to make dependable statements about students' degree of achievement related to that target. When a standard is set for a criterion-referenced test, examinee scores are compared to the standard in order to draw inferences about whether students have attained the desired level of achievement. Scores on the test are used to make statements like, "the student meets the minimum mathematics requirements for this class," or "the student knows how to apply computational skills to solve a complex word problem." The WASL is a criterion-referenced test.

Norm-referenced tests are intended to provide a general measure of achievement in a particular subject. The primary purpose of norm-referenced tests is to make comparisons between students, schools and districts. Careful attention is given to creating items that vary in difficulty so that even the most gifted students may find that some of the items are challenging and even the student who has difficulty in school may respond correctly to some items. Items are included on the test that measure below-grade-level, on-grade-level, and above-grade-level concepts and skills. Items are spread broadly across the subject matter. While some norm-referenced tests provide objective-level information, items for each objective may represent concepts skills that are not easily learned by most students until later years in school. Examinee scores on a norm-referenced test are compared to the performances of a norm-group (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, "the student is the best student in the class," or "the student knows mathematical concepts better than 75% of the students in the norm group." The ITBS and ITED are norm-referenced tests.

To test all of the desired concepts and skills in a particular subject, testing time would be inordinately long. Well designed state or national achievement tests, whether norm-or criterion-referenced, always include samples from the full range of subject matter 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 subject as a whole. To have a broader measure of student achievement in some subject, 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, sometimes even desirable, to have both norm-referenced and criterion-referenced information about students' performance. Determining the type of test to administer should be based on the intended use of the test. If tests are being used to make decisions about the success of instruction, 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; therefore, student performance should be interpreted in terms of how well students have achieved the Washington State EALRs.

Appendix D

GUIDELINES FOR DESIGNING 4TH GRADE TEST ITEMS

This appendix provides general information that guides the construction of individual test items for the 4th grade mathematics WASL. This information is part of the test and item specifications used to create the test.

DESIGNING MULTIPLE-CHOICE ITEMS

- All items must clearly indicate what is expected in a response and must help students focus their response.
- Each multiple-choice item will have a stem (question, statement, or incomplete statement) and three answer or completion options, only one of which is correct. Correct answers will be distributed as evenly as possible among A, B, and C options.
- Multiple-choice item stems will present a complete problem so that students will know what to do before looking at the answer choices. Students should not need to read all answer choices before knowing what is expected.
- The three answer choices will be approximately the same length, have the same format, and have parallel syntax and semantic structures. Students should not be able to rule out a wrong answer or identify a correct response simply because it looks or sounds different.
- Distracters will reflect common errors or misunderstandings, naïve pre-conceptions, or other conceptual problems so that students do not simply eliminate incorrect responses by virtue of a distracter’s obviously inappropriate nature.
- Distracters will not be partially correct responses nor will they “trick” students.

DESIGNING SHORT-ANSWER ITEMS

- Short-answer items will require responses that range in length from two words or numbers, to simple number sentences, to simple figures or diagrams, to no more than two simple sentences.
- Short-answer items will give clear indications of what is required of students (e.g., if a number sentence is required, the stem will indicate this).

- Anything required by the scoring rule will be asked for in the item stem.
- To the extent possible, short-answer items should be answerable via numbers, words, lists, pictures, or simple phrases.
- Short-answer items will involve no more than two steps to arrive at a viable solution.

DESIGNING EXTENDED-RESPONSE ITEMS

- Item stems will contain only one question or prompt and no more than two sentences, one to set up the item and a second to prompt or question.
- Anything required by the scoring rule will be asked for in the item stem. The item will give clear indications of what is required of students (e.g., if a picture and two sentences are required, the stem will indicate this).
- Extended-response items will require responses that range in length from a list of up to five numbers to a response that requires three or four scaffolded steps.
- Extended-response items may require a figure/diagram/table with labels or with one or two words, sentences, or number sentences to support the figure/diagram/table. To the extent possible, extended-response items should be answerable via words, lists, pictures/figures/tables/graphs, or simple phrases.
- Any extended-response item that requires the student to use information from a stimulus will specifically ask for the information from the stimulus that was needed to respond (e.g., “Which numbers in the table could you use to . . .?”).

LANGUAGE AND READABILITY

- All items are reviewed to eliminate language or content that is biased, offensive, or disadvantageous to a particular group of students. No items that display or imply unfair representations of gender, race, persons with disabilities, or cultural or religious groups are included.
- Character names on each form are representative of the ethnic diversity of Washington students. Names are generally short and simple to read.
- Items in each form are balanced by gender frequency and are gender-neutral for active/passive roles.
- The readability level is targeted for the end of 3rd grade but no higher than 4th grade.
- To the extent possible, items testing application and problem solving involve understandable, realistic situations to which most 4th grade students should be able to relate.
- Item stems will be short and succinct with simple syntax and familiar words.

NOTATIONS FOR GRADE 4

- In the item stems, numbers (other than years) having more than three digits to the left of the decimal point include commas to group digits in the usual manner (e.g., 135,000).
- Units are given when appropriate. Standard abbreviations may be used (e.g., cm or ft). However, the unit is spelled out if any confusion is reasonably possible.
- The symbols \cdot and $*$ are not used as multiplication signs in items to test students' multiplication abilities. Only the symbol \times is used as the multiplication sign. The variable x is not used to avoid confusion with the multiplication sign.
- Fractions have horizontal lines separating numerator and denominator.
- Only common and simple fractions are used to test learning targets at the 4th grade level. For operations such as addition and subtraction, as well as for comparisons and ordering, only common and simple fractions are used, and then generally with pictorial representations.
- Grids are used in test items that involve finding the area of a geometric figure. Illustrations are used in test items that involve finding the volume.
- Decimals are used only when expressing monetary units; expressions for monetary units use the *cents* sign (not decimals) for items less than a dollar—e.g., 25¢. Dollar signs and decimals are used in mixed cases (e.g., \$1.25).

CHARACTERISTICS OF ITEMS AND ITEM STEMS OR FOILS

- Each item begins with a stem that asks a question. A stem usually asks a direct question. It seldom uses an incomplete sentence, is worded negatively, or asks for a “best” answer.
- A stimulus that gives information may precede a question or a set of questions. A stimulus consists of brief written material and sometimes a graphic, such as a simple diagram, graph, chart, table, or drawing.
- The stimulus for an item is always factually correct. Stimuli are adapted specifically for the test. A test item always focuses on what is essential and consequential in the stimulus to minimize the impact of, or need for, outside (prior) knowledge.
- To the extent possible, no stimulus, stem, or response for an item will serve as a clue to the correct response for another item.
- Graphs, tables, or figures are clearly associated with their intended items and are not separated from their intended items. When there is any reasonable chance of confusion, page references direct students to the correct graphic.
- Test items are independent and not “linked” (i.e., the answer for any test item does not depend on knowing the correct answer to another item). This prohibition applies to different

items, but not necessarily to parts within a single item. For instance, an enhanced multiple-choice item may ask students to both select a response and then explain their reason for selecting that particular response.

- When appropriate, several items may center around a particular stimulus, graph, chart, or scenario. When this happens, the items will appear on the same page or facing page to the stimulus.
- All items should clearly indicate what is expected in a response and help students focus their response. That is, items clearly state (or imply) the criteria by which the response will be evaluated so that students understand what they are expected to do (e.g., create a table, provide a written explanation, calculate an answer, etc.). General directions that allow the student more freedom in response format may read as follows: “In the space below, use words, numbers, pictures, or any combination of these to explain your answer.” In such cases, any one of these response modes is acceptable as long as it is complete and responsive to the item stem.
- Pictorial representations are realistic and authentic for 4th grader students.
- On items for which manipulatives and/or tools are encouraged or required, students may be given the opportunity to use any punch-out or overlay manipulatives provided (e.g., a ruler or metric ruler), or may use classroom manipulatives (e.g., rulers) or tools with which they are most familiar/comfortable, as long as nothing about the tools would introduce bias into the test results.
- Care is taken to avoid items for which incorrect or inappropriate methods yield the correct response. For example, “Simplify the fraction $64/16$ ” is a poor item, because the correct response can be obtained by canceling the two sixes.
- If a question is stated in terms of one measurement system, all response options will be given in terms of the same measurement system. Units may not always be included in the stem, but they will appear in every distracter or response when appropriate.

Appendix E

SCORING OPEN-ENDED ITEMS

This appendix provides a brief overview of the scoring of open-ended items on the 4th grade mathematics WASL. It gives a summary of the procedures used to score these items and the general scoring guidelines used to score both short-answer and extended-response items (more specific content-related criteria are established for individual test items). The appendix concludes with a summary of analyses conducted to check the reliability of the 1999 scoring.

SCORING PROCEDURES

Each 4th grade mathematics WASL test has 16 open-ended items that require students to construct a written response. Two types of open-ended items are found on the test, short-answer and extended-response. Short-answer items are scored on a scale of 0 to 2 points, and extended-response items are scored on a scale of 0 to 4 points. During item development, individual scoring criteria are developed for each open-ended item. During item reviews, these scoring criteria are reviewed.

The following procedures are used to develop scoring procedures for the open-ended WASL mathematics test items (as well as for open-ended items on other tested subjects). These procedures are used for the full pool of pilot-tested items as well as for the operational tests.

Qualifications of Scorers Qualified and experienced scorers are essential to achieving and maintaining consistency and reliability when scoring open-ended responses. Scorers selected for the mathematics tests are required to have the following qualifications:

- A minimum of a bachelor's degree in an appropriate academic discipline, such as mathematics or mathematics education.
- Demonstrable ability in performance assessment scoring.
- Teaching experience, especially at the elementary or secondary level, is preferred.

Team and table leaders are responsible for supervising small groups of scorers. These leaders 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. A range-finding committee identifies anchor papers—exemplars that clearly and unambiguously represent the solid center of a score point as described in the scoring criteria. The anchor papers form the basis not only of scorer training, but of subsequent range-finding discussions as well.

As part of the range-finding process, assessment and curriculum specialists work with team and table leaders and teachers from Washington to become thoroughly familiar with and reach consensus on the scoring criteria (rubrics) approved by the content committees for each open-ended item. 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 place 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 total range finding group works together to discuss both the common and divergent scores. From this work, they assemble tentative sets of example responses for each prompt.

Discussion is ongoing with the goal of identifying a sufficient pool of additional student responses for which consensus scores can be achieved and which illustrate the full range of student performance. This pool of responses include borderline responses—ones which appear to be between rather than clearly within a score level and which therefore represent a decision-making problem that scorers (with training) need to resolve.

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 13 papers, three examples of 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 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.

Training Following the range-finding sessions, the performance assessment specialists and team leaders finalize the training materials. Qualifying sets of responses are selected for use 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 introduce the scorers 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. Training continues throughout the scoring of all responses to maintain high inter- and intra-reader reliability. Therefore, training is a continuous process and readers are consistently given feedback as they score student responses.

Monitoring After training has occurred, different methods are used to monitor the consistency of each scorer's performance over time. The primary method is through a process called "back-reading." In this process, each table leader checks scores on an average of 5–10 percent of each scorer's work each day, with a higher percentage early in the scoring. If a reader is consistently assigning scores other than those the table leader would assign, the team leader and performance assessment specialist work together to retrain the scorer, using the original anchor papers and training materials. This continuous checking provides an effective guard against scorer "drift" (i.e., 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.

SCORING GUIDELINES

Individual scoring guidelines are developed for each open-ended item. Short-answer items were scored on a scale of 0 to 2 points, and extended-response items were scored on a scale of 0 to 4 points. Scoring criteria for all open-ended items focus on the clear communication of mathematical ideas, information, and solutions. The conventions of writing (sentence structure, word choice, usage, grammar, spelling, and mechanics) are disregarded, as long as they do not interfere with communicating the response.

The following scoring criteria are used to guide item writers in their development of item-specific scoring criteria for short-answer and extended-response items. The criteria help to ensure that the item scoring criteria were clearly focused on the appropriate dimension of mathematics performance.

Rules for Developing Scoring Guides

- An specific scoring guide is developed for each short-answer and extended-response item. The scoring guide for each item will follow the general scoring criteria (see next section). Information from the pilot will be used to refine these scoring guides for use with the final tests.
- Scoring guides for concept and procedures will focus on conceptual understanding and accuracy. Scoring guides for processes will focus on effectiveness, reasonableness, selection of useful procedures, and degree to which solutions are viable.
- Scoring guides will follow a “focused holistic” model in which the score for the response is not only based on overall quality but also results from focusing on several important features of the student's performance.
- Short-answer items will be scored with a 3-level scoring guide (0–2) in which students may receive full credit, partial credit, or no credit. Extended-response items will scored with a 5-level scoring guide (0–4); the levels may be summarized as Extensive, Essential, Partial, Minimal, and Unsatisfactory.

General Scoring Criteria: Short-Answer Items (0–2 Points)

Mathematical Concepts and Procedures

- 2** A 2-point response shows complete understanding of the concept or task, as well as consistent and correct use of applicable information and/or procedures. Set-up and computations are accurate.
- 1** A 1-point response shows partial understanding of the concept or task. There may be minor errors in the use of applicable information and/or procedures. Set-up or computations may have minor errors.
- 0** A 0 point response shows little or no understanding of the concept or task.

Communicating Mathematical Understanding

- 2** A 2-point response shows understanding of how to effectively and appropriately interpret, organize, and/or represent mathematical information relevant to the concept.
- 1** A 1-point response shows some understanding of how to interpret, organize, and/or represent mathematical information relevant to the concept; however, the response is not complete or effectively presented.
- 0** A 0 point response shows little or no understanding of how to interpret, organize and/or represent mathematical information relevant to the concept.

Solving Mathematical Problems

- 2** A 2-point response shows thorough investigation, clear understanding of the problem, and/or effective and viable solution.
- 1** A 1-point response shows partial investigation and/or understanding of the problem, and/or a partially complete or partially accurate solution.
- 0** A 0-point response shows very little or no investigation and/or understanding of the problem, and/or no visible solution; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

Mathematical Reasoning

- 2** A 2-point response shows effective reasoning through a complete analysis or thorough interpretation, supported predictions, and/or verification.
- 1** A 1-point response shows somewhat flawed reasoning either through incomplete analysis or interpretation, prediction that lacks support, or inadequate verification.
- 0** A 0-point response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

Making Mathematical Connections

- 2** A 2-point response makes clear and effective connections within and/or between conceptual or procedural areas.
- 1** A 1-point response makes vague or partially accurate connections within and/or between conceptual or procedural areas.
- 0** A 0-point response makes little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

General Scoring Criteria: Extended-Response Items (0–4 Points)

Solving Mathematical Problems

4 points – Meets all relevant criteria

- Thoroughly investigates the situation
- Uses all applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs elegant, efficient, valid solution using applicable tools and workable strategies

3 points – Meets all or most relevant criteria

- Investigates the situation
- Uses most applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs viable/acceptable solution using applicable tools and workable strategies

2 points – Meets some relevant criteria

- Investigates the situation, but may omit issues or information
- Uses some applicable information related to the problem
- Uses some applicable mathematical concepts and procedures
- Constructs solution using applicable tools and workable strategies, solution may not completely address all issues or strategies may have flaws

1 point – Meets few relevant criteria

- Attempts to investigate the situation
- Uses some applicable information related to the problem
- Uses few applicable mathematical concepts and procedures
- Attempts solution, however, mostly incomplete or not effective

0 points – Student's response provides no evidence of problem-solving skills or shows very little or no understanding of the task; or the prompt may simply be recopied, or the response may indicate “I don't know” or a question mark (?).

Communicating Mathematical Understanding

4 points – Meets all relevant criteria

- Gathers all applicable information from appropriate sources
- Demonstrates interpretations and understandings in a clear, systematic, and organized manner
- Represents mathematical information and ideas in an effective format for the task, situation, and audience

3 points – Meets most relevant criteria

- Gather applicable information from appropriate sources
- Demonstrates interpretations and understandings in a clear and organized manner
- Represents mathematical information and ideas in an expected format for the task, situation, and audience

2 points – Meets some relevant criteria

- Gathers information from appropriate sources
- Demonstrates interpretation and understandings in an understandable manner
- Represents mathematical information in an acceptable format for the task, situation, and audiences

1 point – Meets few relevant criteria

- Gathers little information from appropriate sources
- Demonstrates interpretations and understandings in a manner that may be disorganized or difficult to understand
- Represents mathematical information and ideas in a format that may be inappropriate for the task, situation, and audience.

0-points – Student's response shows little or no understanding of how to interpret, organize or represent mathematical information relevant to the concept; or the prompt may simply be recopied, or the response may indicate “I don't know” or a question mark (?).

Mathematical Reasoning

4 points – Meets all relevant criteria

- Makes insightful interpretations, comparisons, or contrasts of information from sources
- Effectively uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes insightful conjectures and inferences, if asked
- Systematically and successfully evaluates effectiveness of procedures and results, if asked
- Gives comprehensive support for arguments and results

3 points – Meets most relevant criteria

- Makes thoughtful interpretations, comparisons, or contrasts of information from sources
- Uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes expected conjectures and inferences, if asked
- Successfully evaluates effectiveness of procedures and results, if asked
- Gives substantial support for arguments and results

2 points – Meets some relevant criteria

- Makes routine interpretations, comparisons, or contrasts of information from sources
- Includes examples, models, facts, patterns, or relationships to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Partially evaluates effectiveness of procedures and results, if asked
- Gives partial support for arguments and results

1 point – Meets few relevant criteria

- Makes superficial interpretations, comparisons, or contrasts of information from sources
- Examples, models, facts, patterns, or relationships may not be included to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Attends to wrong information and/or persists with faulty strategy when evaluating effectiveness of procedures and results
- Support for arguments and results may not be included

0-points – Student's response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or the response may indicate “I don't know” or a question mark (?).

Making Mathematical Connections:

4 points – Meets all relevant criteria

- Shows a thorough understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in a clear and insightful manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in a clear and insightful manner

3 points – Meets most relevant criteria

- Shows a general understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in an obvious/expected manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in an obvious/expected manner

2 points – Meets some relevant criteria

- Shows a partial understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines AND/OR

- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations

1 point – Meets few relevant criteria

- Shows a little understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, mathematical patterns and concepts in other disciplines AND/OR
- Identifies applies mathematical patterns and concepts in real-life situations

0-points – Student's response makes very little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or the response may indicate “I don't know” or a question mark (?).

RELIABILITY OF TEST SCORES

The reliability of test scores is a measure of the degree to which the scores on the test are a “true” measure of the students’ knowledge and skill relevant to the tested knowledge and skills. There are several ways to obtain estimates of score reliability including: (1) test-retest and alternate forms, (2) internal consistency, (3) standard error of measurement, and (4) inter-judge agreement.

1. Test-Retest and Alternate Forms Test-retest estimates of reliability require the administration of the same test at two different times to the same individuals. Typically the testing times for achievement tests are close together so that new learning does not impact scores. Items are judged to be reliable if the test-takers respond in the same way each time the test is administered. Alternate forms reliability estimates require the administration of two parallel tests. These tests must be created in a way that ensures that they measure the same domain of knowledge and skills using different items.

Both test-retest and alternate forms analyses related to score reliability require significant testing time for students and are generally avoided when there is a concern that fatigue or loss of motivation might impact the results. The WASL is a rigorous assessment that requires significant concentration on the part of students for a sustained period of time. Given the length of the 4th grade mathematics WASL, these two reliability measures were not used because they were unlikely to yield accurate estimates of score reliability.

2. Internal Consistency Internal consistency reliability is an indication of how similarly students perform across items measuring the same knowledge and skills—in other words, how consistent each student performs across all of the items within a test. Internal consistency can be estimated using Cronbach's alpha coefficient when multiple-point items are included on a test (as is the case in the WASL mathematics tests). Two of the demands of applying this method when estimating score reliability are that (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 type and measuring very similar knowledge and skills).

The mathematics WASL has sufficient items to address the issue of test length in internal consistency reliability. The test also combines multiple-choice, short-answer, and extended-

response items across multiple strands. Hence, student performance may differ markedly from one item to another due to prior knowledge, educational experiences, exposure to similar content, etc. Because of this heterogeneity of items, use of Cronbach's alpha for estimating score reliability could result in an *under-estimate* of the reliability of scores. Generally it is believed that the true score reliability is higher than the estimate obtained through the use of Cronbach's alpha when items are heterogeneous as they are in the WASL.

In 1999, the alpha coefficient for the mathematics WASL was .88, which indicates that the test scores can be trusted to represent students' performance on the concepts and skills measured by the test.

3. Standard Error of Measurement Another way to interpret the reliability of test scores is through the use of the Standard Error of Measurement (SEM). The SEM is the standardized distribution of error around a given observed score.
 - When one SEM is added and subtracted from an observed score, we can be about 68 percent certain that the student's true score lies within the band. For example, if the SEM for the WASL mathematics scores is 11.0 and a student's scale score was 400, we could be about 68 percent certain that the student's true score was between 389 and 411 (that is, $400 - 11$ and $400 + 11$).
 - When two SEMs are added and subtracted, we can be about 95 percent certain that the student's true score lies within the band.
 - When three SEMs are added and subtracted, we can be about 99 percent certain that the student's true score lies within the band.

In 1999, the standard error of measurement of the mathematics WASL was 11.24. This level of error is large enough that caution should be used when making decisions based on individual students' scores.

4. Inter-Judge Agreement Inter-judge agreement is another source of evidence for the reliability of mathematics WASL scores. When two trained judges agree with the score given to a student's work, this gives support for the score on the short-answer or extended-response item.

Two methods were used to determine the degree to which judges gave equivalent scores to the same student work: correlations between totals when scores for open-ended items are summed, and percent agreement. The methods were calculated using random selections of student work in the test booklets of 10 percent of the students.

- For total score agreement on the open-ended items, the correlations for the mathematics WASL was quite high (.98) for the 1999 test, with virtually no difference between the means of the total scores summed across open-ended items. Similar results were found for scoring of open-ended items in other subjects.

- Exact agreement between two judges on scores for the mathematics open-ended items ranged from 81 to 98 percent for the 1999 test. Exact and adjacent agreement ranged from 99 to 100 percent.

These findings, which are similar to those of 1998, indicate that the judges can consistently score performances using the scoring criteria developed for each item. The tables below summarize the two types of inter-judge agreement, the score agreement across the total score for the open-ended item set for each content area, and the score agreement for individual items.

Table E-1: Correlations Between and Means of Total Scores of First and Second Readings for Open-Ended Items, 1999 Grade 4 Subjects

Test	Correlation	First Reading Mean	Second Reading Mean
Mathematics	.98	15.84	15.81
Listening & reading	.97	14.63	14.57
Writing	.97	6.93	6.96

Table E-2: Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for 1999 Grade 4 Mathematics WASL

Item	Points Possible	Exact Score Match	Adjacent Scores	Discrepant by Two Points	Discrepant by Three Points	Discrepant by Four Points
3	2	7834	348	20		
5	2	7000	1144	58		
6	4	6696	1288	209	8	1
8	2	7234	933	35		
10	2	6670	1454	78		
15	4	6893	1225	74	9	1
20	2	7712	408	82		
23	2	7540	619	43		
25	2	8009	157	36		
26	2	7078	1092	32		
30	4	6762	1340	94	6	
32	2	7742	402	58		
34	2	7319	868	15		
36	2	7701	464	37		
37	2	7819	309	74		
39	2	7730	457	15		

SUMMARY

The 16 open-ended items on each mathematics WASL are analyzed by qualified scorers who assign points based on criteria established by Washington curriculum experts and teachers. Training is provided to the scorers, and systematic monitoring helps to ensure that the scorers understand how to score items consistently and in accordance with the criteria. Reliability analyses confirm that the test scores can be trusted to represent students' performance on the concepts and skills measured by the test and that the scorers can consistently judge student performances using the scoring criteria developed for each item. However, the standard error of measurement is large enough that caution should be used when making decisions based on individual students' scores.

Appendix F

STATISTICAL ANALYSES OF TEST ITEMS

After student responses are scored, various statistical analyses are conducted to determine the effectiveness of the items and to check for item bias that may have been missed by the earlier reviews. Three types of statistical analyses have been conducted for the 4th grade mathematics WASL administered in 1997, 1998, and 1999.

- Rasch analysis examines the item location and item fit.
- Classical item analysis examines the item means and item-test correlations for each item.
- Bias analysis investigates whether there is different performance on items for examinees of the same abilities who differ by virtue of gender or ethnicity.

This appendix provides an overview of these types of analyses and how they are used when designing the test.

RASCH ANALYSIS

Rasch analysis is an Item Response Theory (IRT) analysis that places all items and student responses on a unique continuous scale. The Rasch analysis process separates item difficulty from the abilities of the students in the sample that is tested. Item difficulties and student abilities can then be estimated for a given test. The item difficulty threshold is the point on the ability scale where students have a 50/50 chance of getting an item correct.

Because the Rasch model can obtain an equal interval scale independent of item difficulty and student performance, the meaning of test scores can be interpreted in terms of scaled scores rather than the number of correct scores. For example, an examinee gets the first eight items correct on Mathematics Test 1 and the first six items right on Mathematics Test 2. The examinee is the same and has the same mathematics knowledge and skill. However, the ease or difficulty of the items result in a different number of correct scores. The Rasch model indicates the true distance of items from one another across the scale so that student test scores reflect the relative distance along the scale rather than the number of items answered correctly. The Rasch model separates item difficulty from student ability so that scores can be interpreted in terms of an underlying ability scale.

For items that have multiple points, a partial credit Rasch model is used to estimate the difficulty threshold of each score for an item. For example, items with 2 possible points can have two item thresholds: one for the point on the scale (location) at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 0 or 1, and one for the point on the scale at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 1 or 2.

Once items and item scores are placed on a scale, items are assessed for “fit” to the Rasch model. The Rasch model assumes there was no guessing on multiple-choice items and that, even though the items differ in terms of difficulty (or location on the scale), the items all function equally in discriminating between students below and above a given location on the scale. In order to be retained in the item pool, items must measure relevant knowledge and skill, represent desired locations on the ability scale, and fit the Rasch model.

Rasch analyses were conducted independently for each part of the mathematics WASL and are noted in Appendix B.²⁹ The fit of items depends upon whether the items in a scale were all measuring a similar body of knowledge and skill—in other words, whether the scale was unidimensional. Just as height, weight, and body temperature are different dimensions of the human body, so are reading, writing, and mathematics different dimensions of learning. Therefore, the items and scales for each test are examined independently.

In order to place all items across test forms on the same Rasch scale, a subset of items was repeated in adjacent forms. In other words, five items in Form 1 were repeated in Form 2; a different five items in Form 2 were repeated in Form 3; a different five items in Form 3 were repeated in Form 4; a different five items in Form 4 were repeated in Form 5; a different five items in Form 5 were repeated in Form 6; a different five items in Form 6 were repeated in Form 7; a different five items in Form 8 were repeated in Form 1. In this way, Form 1 could be the anchor form and all items could be calibrated back to the item locations for the items in Form 1.

CLASSICAL ITEM ANALYSIS

For multiple-choice items, item means and item-test correlations constitute p-values and point-biserials, respectively. These are the classical test theory equivalent of item difficulties and item discriminations. The p-value tells the percent of examinees who respond correctly to an item. Its value can range from 0 to 1.0. The point-biserial gives a measure of the relationship between performance on an item and performance on the test as a whole and can range from -1.0 to 1.0. Item means indicate, for multiple-point items, the average earned score for examinees in the tryout sample. For 2-point items, item means can range from 0 to 2. For 4-point items, item means can range from 0 to 4. Item-test correlations, for multiple-point items, indicate the relationship between item performance and test performance. Item-test correlations can range from -1.0 to 1.0. Item-test correlations are computed using the test scores relevant to the item.

Unlike the Rasch item data, item means and item-test correlations are dependent on the sample of students who took the various tests. If the students are exceptionally well prepared in the concepts and skills tested, item means will be fairly high and the items will appear to be easy. If students are not well prepared in the concepts and skills tested, item means will be fairly low and items will appear to be difficult. If performance on an item does not relate well to performance on the test as a whole, item test correlations will be low or even negative. Hence, both Rasch data and traditional item analysis data are used in item selection.

²⁹ The Difficulty Rating in Tables B-1 to B-3 are the Rasch scores for the results for 1998 and 1999. For open-ended items, the Rasch scores shown in Appendix B were computed using a partial credit model.

BIAS ANALYSIS

The Mantel Haenszel statistic is a chi-square (χ^2) statistic. Examinees are separated into relevant groups based on ethnicity or gender. Examinees in each group are ranked in terms of their total score on the relevant test. Examinees in the focal group (e.g., females) are compared with examinees in the reference group (e.g., males) in terms of their performance on individual items. Multiple 2x2 tables are created for each item (one for each total test score) indicating, for that score, the number of examinees in each group who got the item right and the number of examinees in each group who got the item wrong. Table F-1 shows an example 2x2 table for performance on a hypothetical item for males and females with a total test score of 10 on a 40 point test. It appears that the item is more difficult for females than it is for males who had a total test score of 10.

Table F-1: Responses to a Hypothetical Item for Males and Females with a Total Test Score of 10

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

Note: All examinees had a total test score of 10.

To complete the Mantel-Haenszel statistic, similar 2x2 tables are created for every test score. A chi-square statistic is computed for each 2x2 table and the sum of all of the statistics across all test scores gives the total bias statistic for a single item. When items have multiple points, a generalized Mantel-Haenszel statistic is computed using all points. Items that have a high sum of χ^2 are flagged for potential bias. Generally, a certain percent of the 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 data is likely a result of statistical error. For the WASL analyses, the alpha (error) level was set at .01, that is, about 1 percent of the items are expected to be flagged for bias by chance alone.

HOW THESE ANALYSES ARE USED

Statistical review of items involves examining Rasch item difficulties (locations on the ability scale), item means, and item-test correlations to determine whether items are functioning well. In addition, statistical review requires examining the “fit” of items to the Rasch model. Items that have extremely poor fit to the Rasch model must be revised or removed from the item pool prior to building a final test form. Items that function very poorly (are too easy, too difficult, or have low or negative item-test correlations) must also be revised or removed from the item pool. Finally, items that are flagged for bias against a focal group are examined closely to decide whether they will be removed from the pool of items. Generally, when item tryouts are conducted, sufficient numbers of items are developed so that revision and new tryouts are not needed, and faulty items can be deleted from the item pool.

Appendix G

FURTHER EVIDENCE OF VALIDITY

The most important issue in test development is the degree to which the test actually measures the concepts and skills that it is supposed to measure. For instance, when one claims that a student must use logical reasoning skills to respond to an item, we need evidence that logical reasoning rather than memorization (or something else) was actually used in the student’s response. Validity is an evaluative judgment about the degree to which the test scores can be interpreted to mean what test developers claim that they mean.

Validity is also a complex concept that requires different types of analyses to provide evidence for the validity of test scores.³⁰ Different strategies were used to examine the validity of the 4th grade mathematics WASL. Several are discussed in Chapters 7—analyses of the content of the test in relation to the content of the subject matter (alignment of the WASL with the EALRs), and the need to understand the ways students respond to the items or tasks (Chapter 7).

This appendix summarizes additional evidence for the validity of the 4th grade mathematics WASL test scores as reported in technical reports.³¹ These reports were prepared in accordance with professional testing standards.³² Specifically, this appendix discusses (1) the relationships (correlations) among responses to the tasks, items, or parts of the test (internal validity), (2) the relationships of test scores with other tests (external validity), and (3) analyses of score differences over time and across groups.

INTERNAL VALIDITY

The first analysis examined the relationship between performance on the mathematics WASL and the other WASL subjects. In 1999, the correlation of the mathematics test with the 4th grade reading, writing and listening tests were moderately to strongly related, as seen in Table G-1.

Table G-1: 1999 Grade 4 Correlations Among WASL Test Scores

	<u>WASL Reading</u>	<u>WASL Writing</u>	<u>WASL Listening</u>
WASL Mathematics	.749	.604	.544

³⁰ See Messick, S., “Validity” in Educational Measurement, Robert Linn (Ed.), American Council on Education, U.S. Department of Education, Washington DC, 1989.

³¹ See the technical reports at www.k12.wa.us/assessment/assessproginfo/subdocuments/techreports.asp.

³² See *Standards for Educational and Psychological Testing*, a document prepared by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 1999.

The next type of analysis examined the correlation between performance on the various strand scores for mathematics, reading, and writing for the 1998 and 1999 tests (see Tables G-2 and G-3). The results of these analyses are as follows.

- Correlations among the mathematics *concepts* scores are moderate as would be expected given that these are diverse conceptual areas of mathematics. Prior research has shown that students perform differently on mathematical tasks that tap different areas of mathematics.³³
- Correlations among the mathematics *process* scores are also moderate. The highest correlation is between scores for “reasons logically” and scores for “solves problems.” It is likely that reasoning is an important aspect of problem-solving.
- All of the correlations in the mathematics domain reflect moderately positive relationships between different mathematics strands. Correlations between reading and mathematics strand scores are low to moderate.
- The strongest correlation in both 1998 and 1999 is between the “interpreting nonfiction” reading strand and the “reasons logically” mathematics strand. This suggests that students’ ability to analyze nonfiction text is an important aspect of successful application of mathematics concepts and skills. It is important to note that correlations between writing strand scores and mathematics strand scores are fairly weak, suggesting that writing ability is not very important in the mathematics test.

Table G-2: 1998 Grade 4 Correlations Among WASL Content Strands

Strands (Reading, Writing, Mathematics)	Mathematics Strands								
	NS	ME	GS	PS	AS	SP	RL	CU	MC
Ideas & Details Fiction	.403	.470	.436	.439	.399	.459	.513	.350	.425
Interpretation Fiction	.438	.498	.472	.472	.437	.505	.565	.387	.454
Ideas & Details Nonfiction	.436	.481	.473	.463	.445	.485	.553	.358	.451
Interpretation Nonfiction	.469	.512	.503	.490	.476	.524	.597	.397	.478
Content, Organization & Style	.371	.405	.378	.379	.366	.413	.468	.310	.372
Writing Mechanics	.347	.348	.340	.334	.328	.360	.402	.264	.322
Number Sense (NS)		.464	.452	.429	.460	.455	.530	.367	.452
Measurement (ME)			.487	.464	.456	.474	.550	.375	.477
Geometric Sense (GS)				.446	.456	.475	.557	.372	.474
Probability & Statistics (PS)					.436	.456	.531	.360	.442
Algebraic Sense (AS)						.467	.553	.354	.465
Solves Problems (SP)							.585	.397	.477
Reasons Logically (RL)								.456	.555
Communicates Understanding (CU)									.358

NS-Number Sense

ME-Measurement

GS-Geometric Sense

PS-Probability and Statistics

AS-Algebraic Sense

SP-Solves Problems

RL-Reasons Logically

CU-Communicates Understanding

MC-Makes Connections

³³ See Shavelson, R. J., Baxter, G. P., Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.

Table G-3: 1999 Grade 4 Correlations Among WASL Content Strands

Strands (Reading, Writing, Mathematics)	Mathematics Strands								
	NS	ME	GS	PS	AS	SP	RL	CU	MC
Ideas & Details Fiction	.402	.471	.399	.480	.429	.437	.496	.421	.445
Interpretation Fiction	.423	.490	.410	.499	.445	.456	.513	.436	.467
Ideas & Details Nonfiction	.458	.530	.457	.534	.492	.486	.555	.462	.506
Interpretation Nonfiction	.479	.553	.479	.560	.511	.526	.579	.499	.528
Content, Organization & Style	.345	.393	.330	.415	.374	.383	.392	.391	.384
Writing Mechanics	.362	.424	.363	.409	.398	.380	.409	.379	.395
Number Sense (NS)		.476	.411	.451	.450	.442	.470	.406	.452
Measurement (ME)			.525	.526	.514	.507	.556	.462	.508
Geometric Sense (GS)				.457	.456	.464	.494	.393	.445
Probability & Statistics (PS)					.479	.491	.548	.469	.500
Algebraic Sense (AS)						.467	.513	.432	.485
Solves Problems (SP)							.537	.463	.483
Reasons Logically (RL)								.469	.524
Communicates Understanding (CU)									.470

NS-Number Sense

SP-Solves Problems

ME-Measurement

RL-Reasons Logically

GS-Geometric Sense

CU-Communicates Understanding

PS-Probability and Statistics

MC-Makes Connections

AS-Algebraic Sense

EXTERNAL VALIDITY

We examined external validity by analyzing the relationship among scores for WASL tests and scores for subtests of the Comprehensive Test of Basic Skills (CTBS), a nationally standardized achievement test. We also conducted various factor analyses, which are described in detail in the technical reports.

Correlations Among WASL and CTBS Scores

To assess the external validity of WASL scores, we analyzed the correlations between the scores of the CTBS taken in the fall of 1997 with the scores of the WASL taken the following spring. Table G-4 gives the correlations among the test scores.

All of the total scores of CTBS are highly correlated, and the test scores of WASL are moderately to highly correlated. Correlations between WASL test scores and CTBS totals are moderately high, suggesting that the constructs measured are all interrelated.

The patterns between tests found that the highest correlation among WASL test scores is between reading scores and mathematics scores, higher than between reading scores and writing. WASL reading scores are moderately to highly correlated with nearly all CTBS scores and CTBS reading total scores are moderately to highly correlated with all WASL and CTBS scores. The high correlations between WASL mathematics scores and CTBS mathematics total scores provide evidence for the validity of WASL scores.

Table G-4: Correlations for 1997 Fall CTBS Total Scores and Spring 1998 WASL Scores

Tests	WASL Writing	WASL Mathematics	CTBS Reading	CTBS Language	CTBS Mathematics	CTBS Spelling
WASL Reading	.625	.762	.743	.711	.654	.575
WASL Writing		.579	.587	.619	.537	.529
WASL Mathematics			.661	.658	.698	.507
CTBS Reading				.830	.740	.716
CTBS Language					.780	.712
CTBS Mathematics						.609

Given the high correlations between the reading and mathematics scores, exploratory factor analyses were conducted to determine the extent to which the mathematics WASL is a test of reading. These analyses, which are described in the technical reports prepared for the 1998 and 1999 tests,³⁴ show that WASL reading and mathematics strand scores reflect different dimensions of achievement. So although reading may be prerequisite to all achievement, close examination shows that the WASL mathematics and writing tests are not reading tests.

STUDENT PERFORMANCE ON THE 4TH GRADE MATHEMATICS WASL

Another way to obtain evidence for the validity of test scores is to analyze differences over time, across groups, and in response to instructional interventions. Chapter 5 noted how improvement has occurred on the 4th grade mathematics WASL statewide. The tables on the following pages show that improvement has occurred over time across groups and in different programs.

- Tables G-5 and G-6 shows the number of points on the 1998 and 1999 tests for each of the nine mathematics strands and the results of student performance on each strand.
- Tables G-7 to G-9 provide results by gender.
- Tables G-10 to G-12 provide results by ethnic/racial group.
- Tables G-13 to G-15 provide results by type of program category.

Finally, Figures G-1 to G-4 show the distribution of the 4th mathematics WASL scale scores from 1997 to 2000. The figures show a gradual increase in the level of students meeting the standard, and a general movement from the lower levels to the higher levels. The number of students who took the test in 1997 was fewer than in the other years because the test was optional.

³⁴ See the technical reports at www.k12.wa.us/assessment/assessproinfo/subdocuments/techreports.asp.

STRAND ANALYSIS

Table G-5: Analysis of Stand Results, 1998

Strand	Points Possible	Mean	Standard Deviation	Percent with Strength in Strand
Number Sense	8	4.06	1.66	35.4
Measurement	7	4.74	1.55	33.8
Geometric Sense	7	3.92	1.69	36.2
Probability & Statistics	6	3.28	1.46	45.0
Algebraic Sense	6	3.10	1.61	39.2
Solves Problems	5	1.95	1.46	32.3
Reasons Logically	12	5.35	3.31	36.0
Communicates Understanding	7	3.18	1.85	36.3
Makes Connections	4	1.95	1.10	30.6

Table G-6: Analysis of Stand Results, 1999

Strand	Points Possible	Mean	Standard Deviation	Percent with Strength in Strand
Number Sense	7	4.09	1.43	37.3
Measurement	7	4.27	1.84	47.0
Geometric Sense	7	3.10	1.66	41.2
Probability & Statistics	7	3.64	1.67	31.1
Algebraic Sense	6	3.69	1.60	33.0
Solves Problems	7	2.62	1.65	42.9
Reasons Logically	9	4.22	2.13	31.1
Communicates Understanding	7	2.47	1.78	44.2
Makes Connections	5	2.36	1.28	42.2

GENDER ANALYSIS

Table G-7: Percent Meeting Standards, by Gender (1998)

Group	Meets Standard		Does Not Meet Standard		Percent Exempt
	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	
All Students	10.8	19.8	29.2	38.2	2.0
Males	11.6	19.6	28.0	38.6	2.3
Females	10.0	20.1	30.5	37.8	1.7

Table G-8: Percent Meeting Standards, by Gender (1999)

Group	Meets Standards		Does Not Meet Standards		Percent Exempt	Percent Not Tested
	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		
All Students	13.9	23.3	27.4	33.6	3.2	1.7
Females	13.4	22.9	27.3	32.8	2.9	0.7
Males	13.9	22.7	26.4	32.7	3.4	0.9

Table G-9: Scale Scores, by Gender

	1998			1999		
	Number Tested	Mean Scale Score	Standard Deviation	Number Tested	Mean Scale Score	Standard Deviation
Total	73,102	383.50	32.16	74,392	386.51	33.89
Males	37,448	383.67	33.04	36,242	386.70	33.21
Females	35,654	383.34	31.21	38,150	386.44	34.47

ETHNICITY ANALYSIS

Table G-10: Scale Scores, by Ethnicity

Ethnic Group	1998			1999		
	Number Tested	Mean	Standard Deviation	Number Tested	Mean	Standard Deviation
African American/Black	3,512	366.19	30.03	3,641	366.71	31.92
Alaska Native/Native American	1,972	366.77	29.81	2,040	367.14	34.08
Asian/Pacific Islander	5,104	385.09	33.63	4,832	389.94	34.05
Latino/Hispanic	6,240	361.90	30.60	6,399	363.93	32.88
White/Caucasian	54,877	387.60	30.87	54,944	391.03	32.36
Multi-Racial	832	378.74	29.51	1,805	383.93	31.89

Table G-11: Percent Meeting Standards, by Ethnicity (1998)

Group	No. of Students	Meets Standard		Does Not Meet Standard		Percent Exempt
		Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	
African American/Black	3,428	2.9	9.6	24.0	60.1	3.4
Alaska Native/Native American	1,905	3.7	9.6	24.1	59.6	3.1
Asian/Pacific Islander	5,053	12.7	20.0	27.8	37.4	2.0
Latino/Hispanic	6,048	2.7	8.2	20.6	64.6	3.9
White/Caucasian	54,113	12.4	22.3	30.8	32.8	1.7
Multi-Racial	816	6.6	18.0	30.6	44.1	0.7

Table G-12: Percent Meeting Standards, by Ethnicity (1999)

Group	No. of Students	Meets Standard		Does Not Meet Standard		Percent Exempt	Percent Not Tested
		Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		
African American/Black	3,871	3.5	11.1	24.2	55.2	4.7	1.2
Alaska Native/Native American	2,169	4.6	12.0	23.7	53.8	4.2	1.8
Asian/Pacific Islander	5,054	16.8	23.4	25.4	30.0	3.7	0.7
Latino/Hispanic	6,846	3.5	10.0	22.0	57.9	5.1	1.4
White/Caucasian	56,909	15.7	25.5	27.8	27.5	2.8	0.7
Multi-Racial	1,831	11.2	21.4	28.9	37.0	1.1	0.3

PROGRAM ANALYSIS

Table G-13: Scale Scores, by Program Type

Categorical Program	1998			1999		
	Number Tested	Mean	Standard Deviation	Number Tested	Mean	Standard Deviation
LAP Reading	2,494	360.88	27.29	3,647	364.87	29.24
LAP Mathematics	2,436	357.72	25.48	3,651	364.96	28.90
Title I Reading	3,992	359.80	25.91	8,249	366.27	30.90
Title I Mathematics	1,908	357.83	24.72	5,309	366.06	31.50
Section 504	388	371.36	33.02	497	374.25	33.65
Special Education	6,804	355.22	31.25	7,721	357.47	34.52
Title I Migrant Education	568	350.86	28.11	709	354.75	31.07
Bilingual/ESL	3,149	354.25	28.99	3,294	355.54	31.44
Gifted/Highly Capable	3,221	423.55	26.83	3,337	427.39	24.00

Table G-14: Percent Meeting Standards, by Program Type (1998)

Categorical Program	Number of Students	Meets Standard		Does Not Meet Standard		Percent Exempt
		Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	
LAP Reading	1,224	1.5	6.3	21.8	69.9	0.6
LAP Mathematics	1,906	0.4	4.3	20.2	74.5	0.6
Title I Reading	3,918	1.1	5.0	21.1	72.2	0.6
Title I Mathematics	1,866	0.5	3.6	20.2	74.9	0.8
Section 504	385	6.4	14.1	22.3	54.2	3.0
Special Education	6,412	1.5	5.8	16.4	69.4	6.8
Title I Migrant Education	548	1.0	2.8	14.6	77.1	4.4
Bilingual/ESL	3,014	1.5	4.8	14.5	73.5	5.6
Gifted/Highly Capable	3,207	51.7	32.9	12.2	3.1	0.1

Table G-15: Percent Meeting Standards, by Program Type (1999)

Categorical Program	Number of Students	Meets Standards		Does Not Meet Standard		Percent Exempt	Percent Not Tested
		Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1		
LAP Reading	3,735	2.6	9.2	25.1	60.8	1.6	0.8
LAP Mathematics	3,742	2.5	9.5	24.4	61.1	2.0	0.5
Title I Reading	8,508	3.5	11.2	23.8	58.5	2.4	0.7
Title I Mathematics	5,504	3.8	11.3	23.0	58.4	2.9	0.7
Section 504	523	6.7	15.3	26.6	46.5	3.8	1.1
Special Education	8,677	2.5	7.9	17.2	61.4	9.6	1.5
Title I Migrant Education	779	1.0	6.7	16.8	66.5	7.6	1.4
Bilingual/ESL	3,748	2.2	5.1	15.5	65.2	10.7	1.4
Gifted/Highly Capable	3,349	59.4	30.2	8.7	1.4	0.2	0.1

Figure G-1: Distribution of Scores, 1997

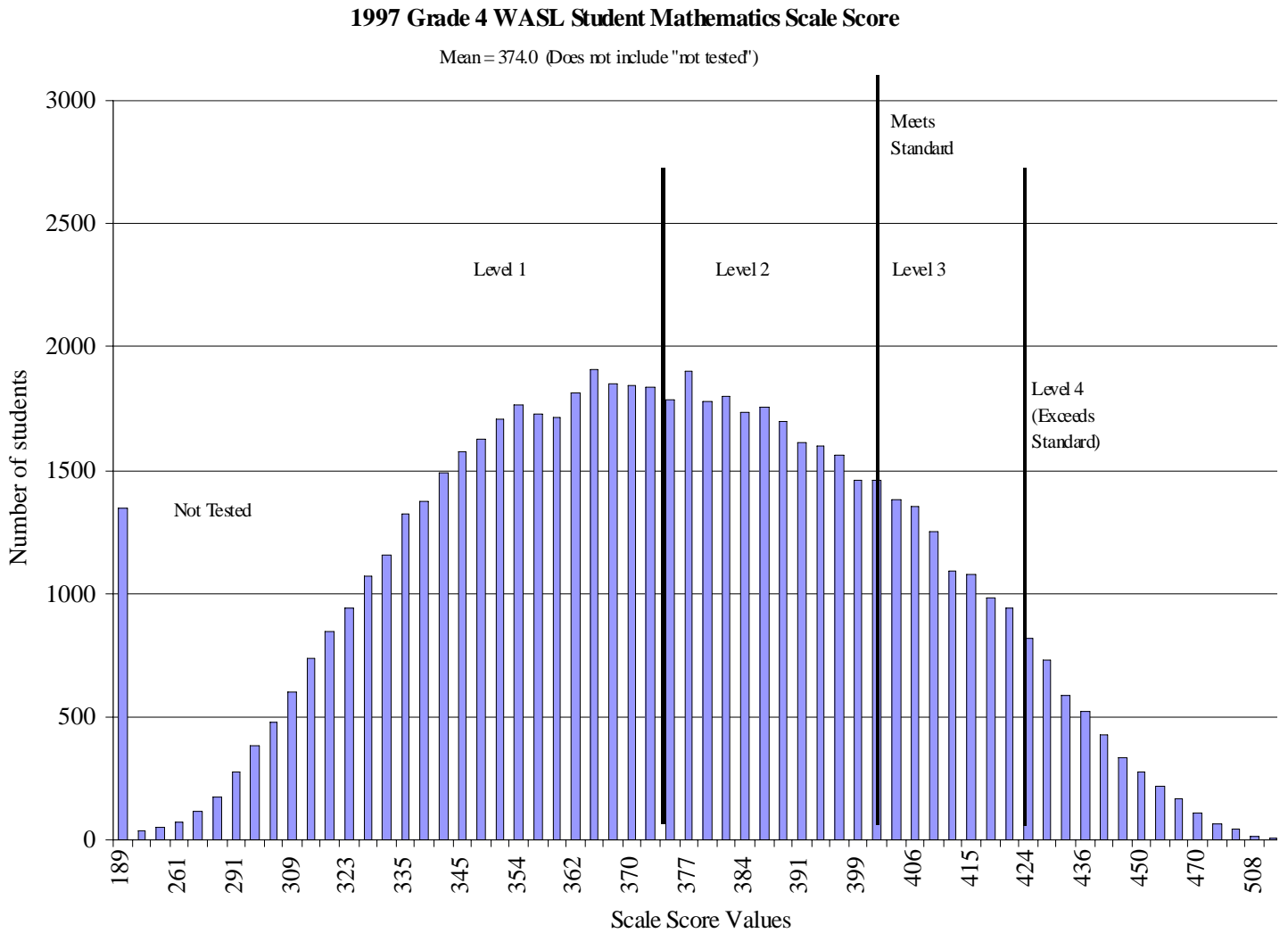


Figure G-2: Distribution of Scores, 1998

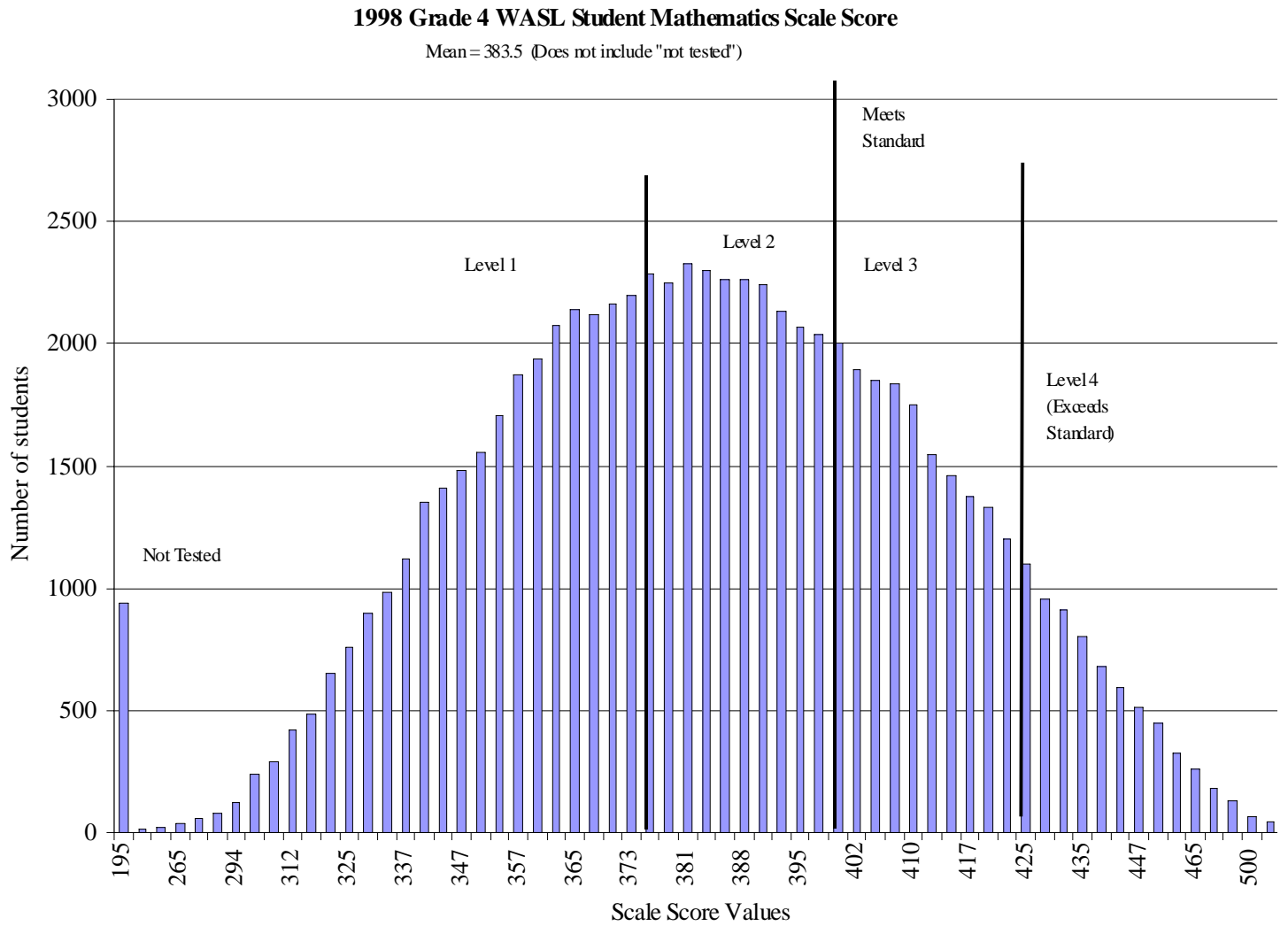


Figure G-3: Distribution of Scores, 1999

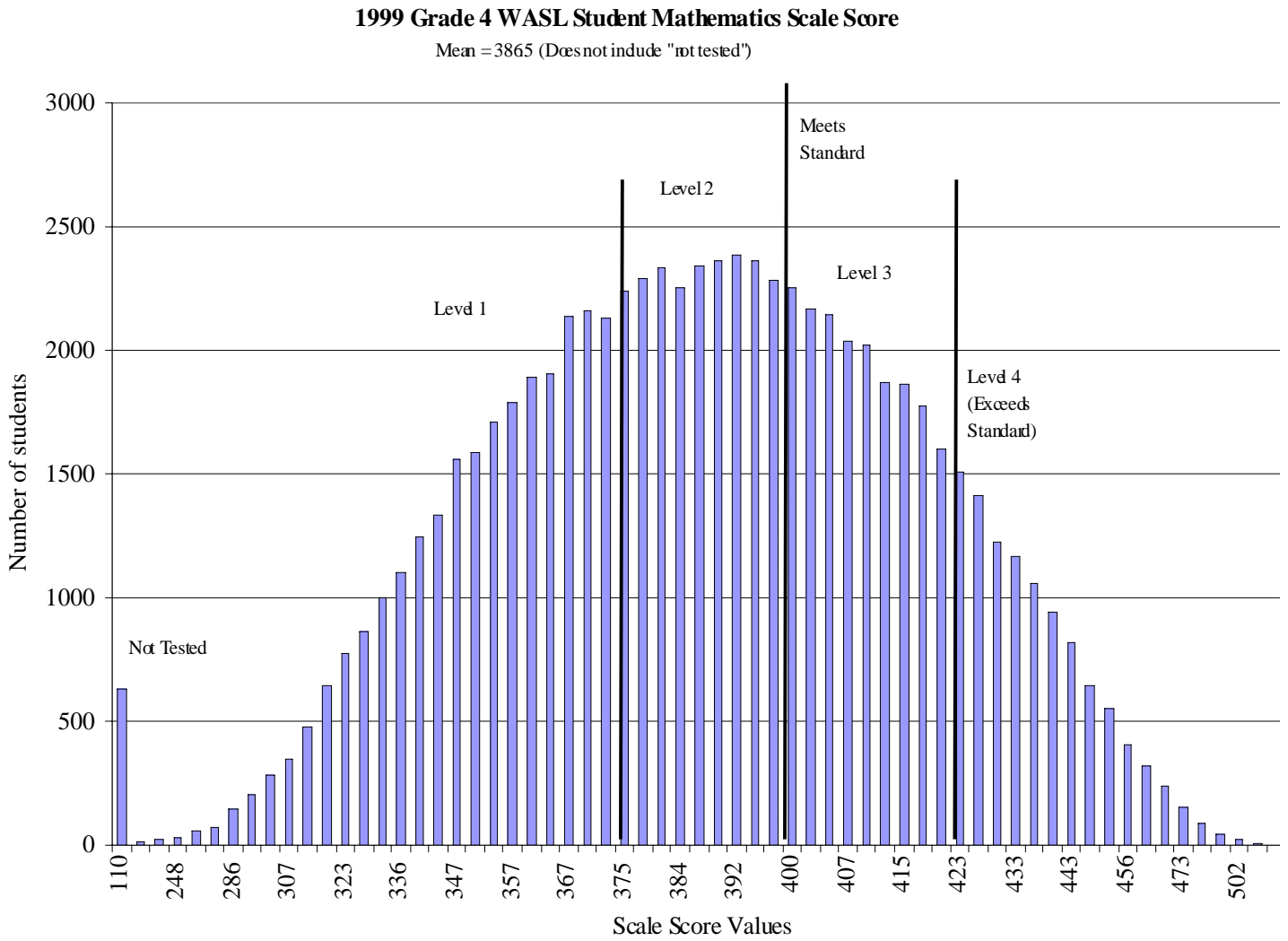
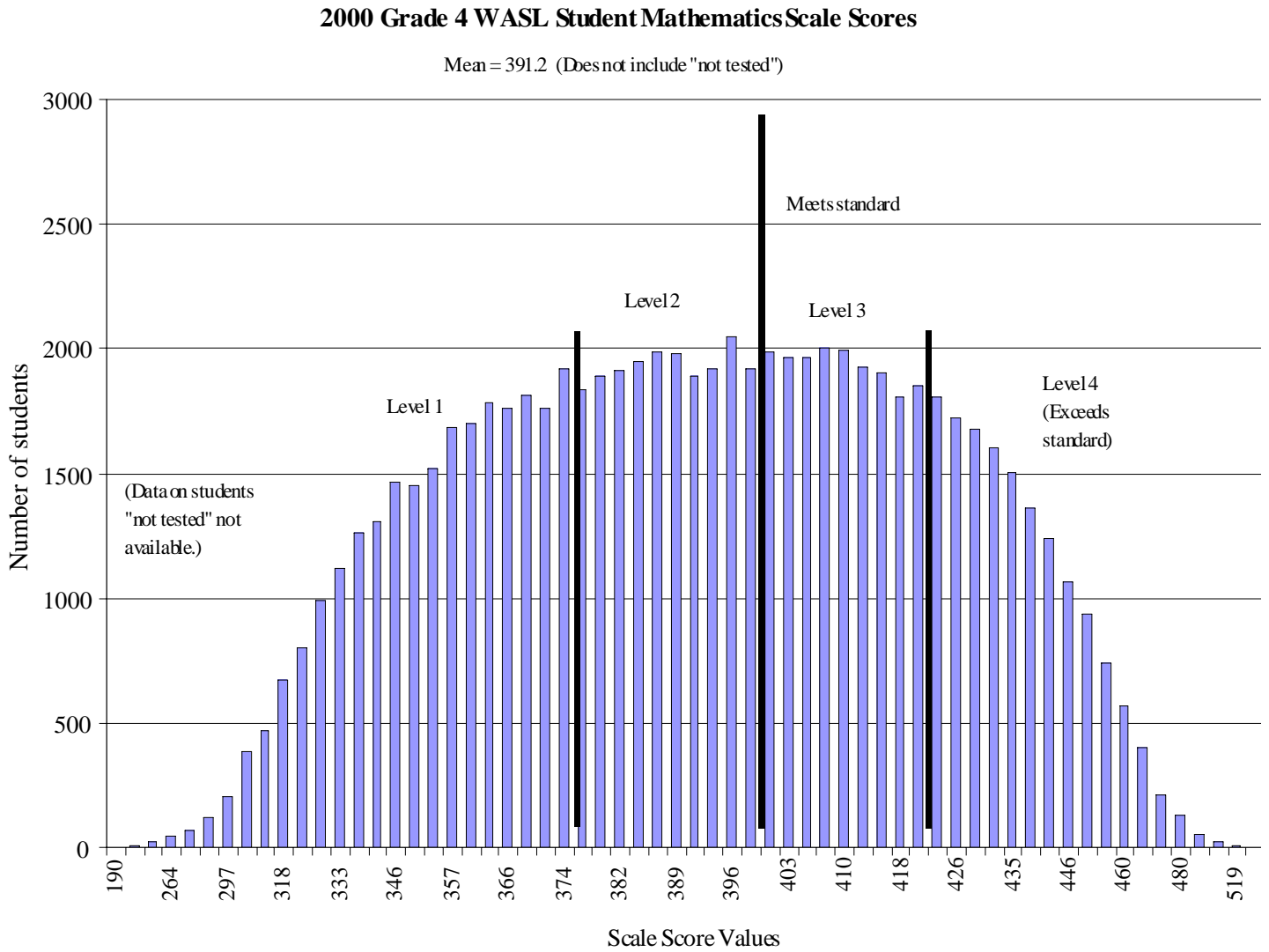


Figure G-4: Distribution of Scores, 2000



Appendix H

NWREL ANALYSIS METHODS

A test's difficulty is influenced by many factors. NWREL used a multiple regression model to predict test difficulty by considering three predictors—*format complexity*, *cognitive complexity* and *mathematics complexity*. In modeling the relationship between test difficulty and the three predictor variables, the analysis assumes that a linear model fits the data. In general, the analysis selects a linear combination of predictors that have a maximum correlation with the dependent variable.

In NWREL's analysis, the dependent variable was the Rasch item calibrations from the state table of specifications for both the 1998 and 1999 tests (see Appendixes B and F). These reflect how students actually performed on the test items. *Format complexity* was measured by classifying the items in three ways: (1) formats that provide clues to the student, (2) formats that are judged free from problems, and (3) formats that hinder student performance. *Cognitive complexity* was measured by identifying the number relevant dimensions in a problem, i.e., the independent chunks of information that students must manipulate to solve a problem. *Mathematics complexity* was measured in two ways: (1) the range of intended curricular exposure as defined by the EALRs framework; and (2) the range of intended curricular exposure as defined by National Council of Teachers of Mathematics (NCTM) standards. In both cases, the range of intended curricular exposure was coded "0" if not specified by the framework, "1" if specified for grades K–2, and "2" if specified for grades 3–5. Grade 5 was included in this range since the NCTM standards classify mathematics content using a grade 3–5 range. Mathematics complexity as measured by the EALRs was included in early phases of the analyses. However, the EALRs failed to explain significant portions of the variation in item difficulty, so the variable was dropped from later analyses.

The order in which the predictors enter the regression equation can make a great difference when determining how much variance in item difficulty is explained by each of the predictors. There are various ways of selecting and ordering predictor variables and cross-validating the resulting regression equation. NWREL used a stepwise regression to examine item difficulty of the test in each of the two years. Stepwise regression employs a test at each stage of the procedure where the least useful predictor is tested for removal. In other words, at each stage in the procedure, a test is made of the least useful independent variable. At any time, a variable that was previously entered could be dropped for the regression when combined with new entries because it was superfluous. Although some statisticians are critical of the misuses of stepwise procedures for noting the relative ordering of predictor variables, similar significant results are found for both years no matter what regression procedure is employed.

Linear regression assumes that the errors are independent and follow a normal distribution with constant variances. These assumptions were checked using various plots and statistical tests that are available in the Statistical Package for the Social Sciences (SPSS) for assessing the model. Although that analysis is not presented here, the assumptions for the regression model hold in each test year examined by the analysis.

The results of the analyses for the 1998 and 1999 tests are provided below. The unique contributions of each complexity variable in explaining the test difficulty in the 1998 and 1999 tests are displayed in Figure 4-2 of Chapter 4 of this report.

1998 RESULTS

For the 1998 test, the stepwise regression enters cognitive complexity first, format complexity second, and math complexity third. All variables are retained by the stepwise procedure since they each explain a significant portion of item difficulty. When using stepwise procedures, cognitive complexity explains nearly 48 percent of item difficulty on the 1998 test, while format complexity explains 14 percent, and mathematics complexity (as measured by NCTM standards) explains almost 7 percent. However, these three variables are consistently chosen no matter what variable selection procedure is employed. The statistical tests of each independent variable in the equation are also significant, regardless of the order in which the variables were entered into the model.

The model summary shown in Table H-1 displays the stepwise entry of the variables and its effects on the R-square, or coefficient of determination, for the 1998 test. *R-square* is a measure of “goodness of fit” to the linear model, and can be thought of as the square of the correlation between the predictor variables and task difficulty. Results indicated that only the EALRs predictor of mathematics complexity would be dropped from the analysis. The proportions of explained and unexplained variation that are found in Figure 4-2 of this report were calculated by finding the amount of variation added to R-square when an additional predictor was added to the model.

In Table H-2, the *B*'s are the estimates of the slope of the line associated with each of predictor variable. The standard error of the regression (*Std. Error*) is a measure of error associated with these slope estimates and can be used to build a confidence interval for each *B*. The standard error is smallest for cognitive complexity. The *Beta coefficient* describes one way of making regression coefficients (*B*'s) comparable by putting them on the same scale. The Beta coefficients describe the importance of each predictor relative to other predictors in explaining item difficulty. The Beta coefficient weighs cognitive complexity as having the most relative importance, followed by mathematics complexity, and then by format complexity. The *t statistic* and the two-tailed significance levels (*Sig.*) are also displayed. The small observed significance level (less than 0.006 for all predictors) associated with the slopes of the respective predictors support the hypotheses that cognitive complexity, format complexity, and mathematics complexity have a linear association with item difficulty.

Table H-1: 1998 Model Summary

Model	R	R Square	Sum of Squares	df	Mean Square	F	Sig.
1. Cognitive complexity	0.691	0.477	14.538	1	14.538	34.632	0.000
2. Cognitive complexity Format complexity	0.790	0.624	19.025	2	9.512	30.697	0.000
3. Cognitive complexity Format complexity Math complexity (NCTM)	0.834	0.695	21.205	3	7.068	27.402	0.000

Table H-2: 1998 Regression Coefficients

Model	B	Std. Error	Beta coefficient	t	Sig.
1. Cognitive complexity	0.480	0.082	0.691	5.885	0.000
2. Cognitive complexity Format complexity	0.415 0.552	0.072 0.145	0.597 0.395	5.759 3.805	0.000 0.001
3. Cognitive complexity Format complexity Math complexity (NCTM)	0.323 0.421 0.569	0.073 0.140 0.196	0.464 0.301 0.322	4.419 3.008 2.907	0.000 0.005 0.006

1999 RESULTS

For the 1999 test, the stepwise method enters cognitive complexity first, mathematics complexity second, and format complexity third in the regression analysis. Again, all variables are retained by the stepwise procedure since they each explained a significant portion of item difficulty. Cognitive complexity explains 41 percent of the difficulty in the test, with mathematics complexity explaining almost 16 percent of the difficulty, and format complexity explaining 6 percent. As before, all three variables are statistically significant no matter what the order of entry of the variables into the analysis.

The results of the analysis of the 1999 test are shown in Tables H-3 and H-4. Just as in the previous tables, these tables display the coefficient of determination for each model as an additional variable is added by the stepwise procedure. As before, only the mathematics complexity measured by the EALRs was dropped from the model by the stepwise procedure. The proportions of explained and unexplained variation that are displayed in Figure 4-2 of this report were calculated by finding the increases in R-square associated when an additional predictor was added to the model.

Just as in Tables H-1 and H-2, the estimated regression slopes for the 1999 test are displayed in the column labeled *B*, and, as with the 1998 test, cognitive complexity has the smallest standard error. Again, the *Beta coefficients* weigh the importance of each variable relative to the other predictor variables. As in 1998, Beta coefficients first weigh cognitive complexity as largest in magnitude, followed by mathematics complexity, and finally format complexity. The *t statistics* for each regression coefficients is again displayed for all three predictors. The smallness of the observed significance levels (*Sig.*)—less than 0.02—associated with the estimated slope of each predictor supports the hypotheses that the three variables are related in a linear manner. Since the

results of the 1999 analysis cross-validate the 1998 modeled results, the modeled predictors have utility in explaining variation in item difficulty. The 1999 results reaffirm the importance of the three predictors in explaining item difficulty.

Table H-3: 1999 Model Summary

Model	R	R Square	Sum of Squares	df	Mean Square	F	Sig.
1. Cognitive complexity	0.644	0.414	13.386	1	13.386	26.901	0.000
2. Cognitive complexity Math complexity (NCTM)	0.755	0.570	18.421	2	9.211	24.596	0.000
3. Cognitive complexity Math complexity (NCTM) Format complexity	0.794	0.631	20.378	3	6.793	20.522	0.000

Table H-4: 1999 Regression Coefficients

Model	B	Std. Error	Beta coefficient	t	Sig.
1 Cognitive complexity	0.329	0.063	0.644	5.187	0.000
2 Cognitive complexity	0.264	0.058	0.516	4.553	0.000
Math complexity (NCTM)	0.770	0.210	0.210	3.665	0.001
3 Cognitive complexity	0.218	0.058	0.426	3.778	0.001
Math complexity (NCTM)	0.615	0.208	0.332	2.965	0.005
Format complexity	0.430	0.177	0.283	2.432	0.020

SUMMARY OF ANALYSIS

Cognitive complexity explains most of the difficulty in both the 1998 and 1999 test. Cognitive complexity is appropriate, provided schools can induce enough cognitive growth through their curricular and instructional decisions. A student’s ability to control and coordinate cognitive activity while thinking is partly determined by development and partly by learning in an environment that offers quality curriculum and instruction. Additional factors such as school leadership and parental involvement are also presumed to influence and support this growth.

Math complexity is also found in the two test administrations—7 percent in 1998 and 16 percent in 1999. Math complexity, along with cognitive complexity, is the appropriate “content” to be measured by the WASL. It is assumed that this knowledge and skill is introduced, practiced, and mastered by students. A major effort should be made to ensure that the content and skills of the local curriculum align with the state EALRs. The mathematics complexity of items in an age-appropriate test should contribute to making the test a good predictor of student ability.

Finally, there is a moderate percentage of format complexity in the 1998 exam (14 percent) and a lesser amount (6 percent) in the exam administered in 1999. Regardless of the amount, format complexity is undesirable because it adds unnecessary difficulty to test items and challenges the student’s capability to understand the intent of the item writer. In doing so, format complexity may mask a student’s cognitive and mathematical abilities.

EXAMPLE OF ITEM TASK ANALYSIS

NWREL created a complex methodology to analyze the relative difficulty of each item on the 1998 and 1999 tests. They sought advice from the panel of experts regarding this methodology and regarding ways to improve individual test items. An example of this methodology is shown on the following pages. This methodology was applied to each item on the two tests. To preserve the security of the test items on the operational tests, the first item from the Example Test is used to illustrate the analysis method.

The analysis of the item provides a good example of how cognitive complexity, mathematics complexity, and format complexity interact to partially explain the difficult or ease of a test item. From a *cognitive complexity* perspective, the item is fairly easy for 4th grade students. There are only four variables involved in solving this problem. Cognitive research has shown that 4th grade students can usually handle four variables in a task without difficulty. The cognitive complexity is increased somewhat by requiring students to solve the two-step problem, again an appropriate level of complexity for 4th grade students.

Format complexity is contained in the language of the test item. There are no graphics, tables, or other representations of the problem. The only element of the problem that increases the format complexity is the need for the students to equate cans of paint with gallons of paint. For some, this may not be an experience they have had, or they may associate cans of paint with another measure of the quality of paint, such as a quart. This is a minor flaw that could be problematic for some students.

Mathematics complexity in the item is in the application of two mathematics skills—addition and estimation—in a simple problem. These involve skills and processes describe in the strands and learning targets contained in the Test Specifications. These skills are also found in early ages in the NCTM standards. The mathematics complexity should be easy for 4th grade students.

The suggested rewrite addresses the problem of equating cans and gallons by using gallons of paint throughout the test item.

Example Test Item 1		Kaitlin’s dad bought 6 gallons of white paint for the house. He also bought 3 gallons of yellow paint for the garage. Each can of paint costs \$14.95. About how much did Kaitlin’s dad pay for the paint?	
I. Cognitive Complexity			
Classification Complexity	Variables	Representation Complexity	
<ul style="list-style-type: none"> • Total 4 variables • Dependent variable (DV) is represented • All (100%) independent variables (IV) are represented 	DV: Approximate cost for the paint IV1: Cost of a can of paint IV2: Color of paint IV3: Number of cans of paint	None	
Relational Complexity	Intermediate IVs	Clarification of Intermediate Ivs	
<ul style="list-style-type: none"> • 1 intermediate IV • 0 interactions • 1 effect: 2x1 	IV4: Total number of cans of paint. Interaction of IVs None Effects of IVs on DV (IV4; IV1) ⇒ DV	<ul style="list-style-type: none"> • Students need to first find the total number of cans of white and yellow cans of paint. Clarification of Interactions <ul style="list-style-type: none"> • None Clarification of Effects <ul style="list-style-type: none"> • Students estimate the cost for the total number of cans of white and yellow cans of paint. Students can work the problem by first estimating the costs for the white and then the yellow cans of paint and add the cost together. This involves working with two intermediate independent variables to estimate the total cost. 	
Cognitive Process Complexity	<ul style="list-style-type: none"> • Two step solution • Application 	Clarification of Cognitive Process Complexity	
		<ul style="list-style-type: none"> • Students needs to use addition to determine the total number of cans of paint and then estimate the total cost of the paint. • Students apply mathematics skills of addition and estimation to solve the problem. 	

II. Format Complexity	
Location of Format Complexity <input type="checkbox"/> Instructions <input type="checkbox"/> Problem figure, table, graph, etc. <input checked="" type="checkbox"/> Language <input type="checkbox"/> Response options	Clarification of Format Complexity The students must assume that a gallon of paint is equal to a can of paint. This item can be clarified by using gallon of paint consistently throughout the problem.
III. Math Skills Complexity	
Content Strand <ul style="list-style-type: none"> • Number sense (NS03, NS04) 	Clarification of Content Strand <ul style="list-style-type: none"> • Students need to add two numbers together and estimate the total cost of a purchase.
Process Strand <ul style="list-style-type: none"> • Problem Solving (SP03)* 	Clarification of Process Strand <ul style="list-style-type: none"> • Students need to identify data that are necessary in solving a problem. • Students need to identify the relationships between the information presented and apply a strategy for solving the problem.
Skills/Strategies <ul style="list-style-type: none"> • Two-step problem 	Clarification of Skills/Strategies <ul style="list-style-type: none"> • Students need to apply a two-step solution process to this problem, either first finding the total number of gallons of paint and then estimating the total cost, or estimating the total cost for the white and yellow gallons of paint and estimating the sum of the two costs.
Concept Complexity <ul style="list-style-type: none"> • Concept of gallon, can and estimated cost 	Clarification of Concept Complexity <ul style="list-style-type: none"> • Students need to understand the relationship of number of gallons to the number of cans and the cost per can of paint.
IV. Suggested Rewrite	
1. Kaitlin’s dad bought 6 gallons of white paint for the house. He also bought 3 gallons of yellow paint for the garage. Each gallon of paint costs \$14.98. About how much did Kaitlin’s dad pay for the paint? a. \$80 b. \$100 c. \$140	

* Strand as identified and described in the EALR.

Appendix I

EXPERTS PROVIDING ASSISTANCE

OSPI relied primarily on outside experts to complete this study. The Northwest Regional Educational Laboratory (NWREL) in Oregon conducted various analyses related to the difficulty of the items on the 1998 and 1999 test and reviewed the processes used to develop the test and set the standards. Four experts conducted additional analyses for this study and provided technical assistance to NWREL and OSPI. This appendix provides information about these independent experts.

PRIMARY NWREL PROJECT STAFF

Dr. Dean Arrasmith directs NWREL's assessment program and was in charge of their work on this study. He completed his Ed.D. at the University of Massachusetts specializing in psychometrics and research design. Prior to joining the NWREL in 1990, Dr. Arrasmith was the statewide test coordinator for New Mexico, where he was responsible for the development, implementation, and reporting of a comprehensive state assessment program that integrated norm-referenced, criterion-referenced, and performance assessments. This assessment system was designed to include the needs for Title I and bilingual program assessment needs. In addition, he was a Senior Evaluator for the Dallas school district for eight years. He serves as an Advisory Editor for the *Journal of Educational Measurement* and is an active member of the Psychometric Society, the American Education Research Association, and the National Council on Measurement in Education.

Dr. Thomas V. Tinkler, an Associate Evaluator at NWREL, helped design and conduct the various analyses of the 4th grade test items. He completed his Ph.D. at Ohio State University in educational policy and leadership. While at NWREL, he has worked on various evaluation projects in various school districts, community agencies, and state universities. Dr. Tinkler also has eight years of teaching experience at the secondary and college levels and has held various research associate and consulting positions in areas including school finance and vocational education. He recently organized a Rasch/Development Conference at NWREL to explore new ways of assessing students using developmentally appropriate classroom assessments.

INDEPENDENT EXPERTS

Dr. Verna Adams is Associate Professor in the Department of Teaching and Learning at Washington State University in Pullman, Washington. She received her doctorate from the University of Georgia and has degrees in mathematics and mathematics education with a focus on cognitive and affective issues related to learning mathematics. She has been a teacher at the middle and secondary school and college levels, has written K–8 mathematics curriculum materials, and has worked extensively on research projects in K–8 classrooms. Dr. Adams is a

member of the American Educational Research Association and its special interest group in mathematics education research, the National Council of Teachers of Mathematics, and Psychology of Mathematics Education.

Dr. Stanley Pogrow is Associate Professor of Educational Administration at the University of Arizona. He has a Ph.D. in education from Stanford University and specializes in the design, implementation, and dissemination of learning environments for educationally disadvantaged students. Dr. Pogrow developed the Higher Order Thinking Skills (HOTS) Project, a pure thinking skills approach to Title I in grades 4–8. With support from the National Science Foundation, he recently developed SUPERMATH, a new form of middle school mathematics curriculum for all students. Dr. Pogrow has worked at several other universities, the National Science Foundation, the California State Department of Education, and was a public school teacher in New York City for six years.

Dr. Cindy Walker is Assistant Professor in the Department of Educational Psychology at the University of Wisconsin–Milwaukee and until recently was Assistant Professor of Educational Psychology, College of Education at the University of Washington in Seattle. She has a Ph.D. in applied psychometrics and statistical analysis from the University of Illinois and has undergraduate and graduate degrees in mathematics. She has worked as an education researcher and taught mathematics at several colleges. Dr. Walker has also conducted several research projects on the 4th grade mathematics WASL.

Dr. John Woodward is Professor at the School of Education at the University of Puget Sound in Tacoma, Washington. He has a Ph.D. from the University of Oregon in special education and specializes in several educational areas, including mathematics, special education, and technology. He has written and consulted extensively on each of these subjects. Prior to joining the University of Puget Sound, Dr. Woodward conducted education research at the Eugene Research Institute and taught special education students at the elementary and secondary school levels.

Appendix J

LEGISLATIVE MANDATE

ANALYSIS OF FOURTH GRADE MATHEMATICS ASSESSMENT. By August 1, 2000, the superintendent of public instruction shall complete an objective analysis of the fourth grade mathematics assessment. The analysis shall include, but need not be limited to, the student developmental level required to achieve the fourth grade standard successfully and the extent to which the assessment measures a student's computational skills, problem-solving skills, mathematics communications skills, and a breakdown of other skills assessed. The analysis shall include the percentage of items that: Require students to use computational skills without the use of technology; require the use of technology to complete an item; measure mathematics communications skills; measure problem-solving skills; and measure other skills included in the mathematics assessment. The superintendent of public instruction shall consult recognized experts with differing views on the instruction of mathematics, and report the results of the analysis to the governor and the education committees of the house of representatives and the senate by August 15, 2000.

Washington Laws, 1999, Chapter 388, Section 601 (SSB 5418.PL)