

Washington Assessment of Student Learning
Washington Alternate Assessment System (WAAS)

2001

Technical Report

Prepared by
The Riverside Publishing Company

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

Technical Report on WAAS for 2000-2001

Part	Title	Page
1	Overview and Background	
	Report on the Washington Alternate Assessment System (WAAS)	
	Introduction	
	Purpose of the Portfolio Assessment	
	Purpose of the Commercially Available Tests	
	Participation Rates	
	Use of Commercially Available Tests	
Part 2	Scoring	
	Commercially Available Scoring	
	Portfolio Scoring	
Part 3	Evidence for Validity of Inferences from Scores	
Part 4	Reliability	
	Inter-Scorer Agreement	
	Internal Consistency	
Part 5	Description of Performance of Students	
Appendix A	WAAS Washington Alternate Assessment System Demographic	Page
Appendix B	WAAS Portfolio Scoring Summary Sheet	
Appendix C	National Technical Advisory Committee Members and Special Education, Alternative Assessment Task Force Members	

Introduction

The *Washington Alternate Assessment System (WAAS)* was administered operationally for the first time during the Spring 2001. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) recommends that test developers and publishers produce a technical manual that provides information documenting the technical quality of an assessment, including evidence for the reliability and validity of test scores. This document contains the technical information for the 2001 WAAS.

State assessment programs provide one method of determining student academic achievement. The Washington State Assessment System provides accountability for program and educational opportunities for all students. Alternate assessment, as part of Washington's assessment program, ensures a unified system, program, and student accountability linked to the common core of learning within the general curriculum.

The Washington Alternate Assessment System (WAAS) process was developed by the Washington Alternate Assessment Task Force and expanded by Advisory Panels in response to the following requirement in the Individuals with Disabilities Education Act 1997: "The State has established goals for the performance of children with disabilities in the state that . . . are consistent, to the maximum extent appropriate, with other goals and standards for children established by the state." It was toward fulfillment of this requirement that alternate assessments are based on Washington's Essential Academic Learning Requirements (EALRs) in the content areas of Communication, Reading, Writing, and Mathematics. In this manner, all students in Washington will be moving toward the same general standards. The inclusion of students with disabilities in the assessment and accountability system is critical to ensure appropriate allocation of resources and learning opportunities for these students.

The Washington Alternate Assessment System was designed for a very small percentage of the total school population for the *Washington Assessment of Student Learning (WASL)*, even with accommodations, would be an inappropriate measure of progress. The two options currently available in the alternate assessment system are commercially available tests and portfolio assessment.

Purpose of the Portfolio Assessment

The Washington Alternate Assessment Task force, made up of administrators, higher education personnel, teachers, and parents, determined the following two-fold purpose of the portfolio assessment:

- To provide an appropriate method of measuring progress on state goals and standards for students who are not able to access the WASL or any commercially available test, even with accommodations and
- To ensure that students will be able to generalize the Individualized Education Program (IEP) skills to the maximum extent possible.

The basic building block of the portfolio assessment is evidence of the student's work. Each of the entries in the portfolio will document two dimensions of learning: progress on IEP skills linked to the EALRs and student generalization of those skills.

Evidence of the student's work should demonstrate participation in and progress toward those IEP goals that are aligned to state standards (EALRs). In this way, evidence of progress on IEP skills linked to the EALRs can measure progress on state goals and standards.

The student generalization of skills evidence should show the extent to which a student can demonstrate the IEP skill linked to EALRs in the following ways:

- using appropriate modifications/adaptations, supports, or assistive technology in order to demonstrate all he or she knows and is able to do;
- in a variety of settings and contexts in which the student is able to use learned skills. These places can include the classroom, other areas of the school, community settings, and home;
- interacting with nondisabled peers and others during IEP activities for the purpose of developing social relationships to enrich his or her life; and
- using self-determination skills in planning, monitoring and evaluating IEP skill activities.

Purpose of the Commercially Available Tests

The Individualized Education Program (IEP) team may select a commercially available test to measure progress toward state standards in listening, reading, writing, or mathematics. This option is available for students whose academic skills can be measured, but whose disability prevents them from participating in one or more content areas of the WASL, even with accommodations. A commercially available test (CAT) should only be administered in content areas for which the student qualifies for specially designed instruction.

For the 2001 administration of commercially available tests, no additional guidelines or list of acceptable tests were provided. IEP teams chose an appropriate test for the student

that measured the student's skills in the content area. No alignment to specific standards was required.

Participation Rates

The participation by district varied (Table 1). District participation depended on the district's approach to providing programming and assessment for student with special needs. Some districts did not submit any commercially available scores but did submit a number of portfolios, other districts submitted a number of portfolios but did not report any commercially available scores. Some district did not report either commercially available test results or portfolios.

Table 1: Number of District Participating

	Washington Assessment of Student Learning	Commercially Available Test Only	Portfolio Submitted
Number of Districts	296	150	77

Approximately 1-2% of the student population is probably eligible for an alternate assessment in each given year. One would expect between 2,100 and 4,200 of the students in grade 4, 7 and 10 to be assessed using either a commercially available test or portfolio. As can be seen in Table 2 the number of portfolios submitted together with the number of commercially available tests submitted is nearer the lower end of this estimate.

Table 2: Number of Students in Grades 4, 7 and 10 by Type of Assessment

	Washington Assessment of Student Learning	Commercially Available Test Only	Portfolio Submitted
Total Number of Students – Listening	218,510	1,290	417
Total Number of Students – Reading	216,999	2,244	428
Total Number of Students - Writing	209,908	2,131	435
Total Number of Students - Math	217,341	2,090	426

Use of Commercially Available Tests

Table 3 indicates the frequency of commercially available tests that were used at least 20 times in the state. There were a number of other tests that were also used across the state but the frequency of use was less than 20.

Table 3: Frequency of Use of Commercially Available Test by Subject

Name of Instrument	Frequency			
	Listening	Reading	Writing	Mathematics
Woodcock Johnson	153	964	967	845
Brigance	340	459	435	442
Wechsler Individual Achievement Test WIAT	299	267	255	258
Wide Range Achievement Test (WRAT)		72	57	71
ITBS	56	71	59	68
Woodcock Johnson -Mini Battery		65	65	61
Oral Written Language Scales (OWLS)	110			
STAR		41		36
Vineland Adaptive Behavior Scale	46	35		30
Peabody Picture Vocabulary Test (PPVT)	36			
Key Math				26
Test of Math Ability (TOMA) 2				24

Part 2 Scoring

Commercially Available Scoring

School personnel scored the commercially available tests according to the publishers' instructions. The scorer entered these scores onto the demographic sheet for WAAS. When applicable, the portfolio scores were also entered onto this sheet. That sheet was then scanned and the reports were generated from that information. During the course of this year, it was found that many of the sheets had contradictory or unclear information entered. Therefore, changes have been made in the sheet itself and the training so that the data generated will be more accurate.

Portfolio Scoring

The portfolios were scored over a two-week period in July. For the first week, a small group of teachers and representatives from the Riverside Publishing Company were led by OSPI in range-finding. Teachers and RPC personnel were trained by OSPI so that

they all had a common understanding of dimension definitions and score points for each dimension in the portfolio.

OSPI pre-selected a number of portfolios that exemplified score points for each dimension. First, two of the aforementioned portfolios were used as tools to train teachers and RPC personnel. Teachers and RPC personnel were given one portfolio to score. When all were finished scoring, OSPI discussed each score given and consensus was achieved. This step was repeated three times.

Once teachers and RPC were trained to OSPI's standards, the group was divided into three groups of two teachers and one RPC person. Each group scored portfolios and then all groups met to come to consensus. Fourteen portfolios were scored in this manner.

The second week, additional teachers and a scoring staff member from NCS Pearson used as scorers. The teachers who had attended the first week served as table leaders and the Riverside Publishing Company and NCS Pearson staff served as assistant table leaders. The first day was used as a full day of training, and scoring started on the second day. Teachers were trained by OSPI so that they all had a common understanding of dimension definitions and score points for each dimension in the portfolio. OSPI led the training on definitions of each dimension and its rubric. OSPI pre-selected two portfolios that exemplified all ranges of score points for each dimension. OSPI facilitated discussion of these portfolios. Teachers were given two portfolios to score independently. OSPI and RPC facilitated discussion upon completion of scoring. When OSPI and RPC concluded that all teachers were properly trained, scoring procedures were reviewed.

On the first day every portfolio was scored twice and the table leader (or assistant table leader) score was used as the final score. When clarification was needed, or discrepancies were found, OSPI staff served as the final arbiter.

Four tables were established for scoring purposes. At each table there was a table leader (RPC person or teacher returning from range-finding) and four teacher scorers. There was a lead scorer (OSPI) table, as well. Scorer reliability was calculated at this table. When clarification was needed, or discrepancies found, OSPI was the final arbiter.

Scorers chose one portfolio randomly. Portfolios were arranged according to district. Scorers were told not to choose a portfolio from their district or their table leader's district. Scorers signed for one portfolio with its unique number. At each table was a sheet on which scorers were required to check in and out with their initials. Next, scorers scored the portfolios and then recorded their scores on content and dimension sheets. Scorers gave portfolio and paper work to table leader.

Table leaders initialed and scored each portfolio without looking at teacher scorers' results. Table leaders scored portfolio "blind." Table leaders filled out an entry form with each student's name and portfolio number. Table leaders filled in and transferred all

scores onto bubble sheet. Table leaders handed each portfolio to lead scorer's table and scores were entered into a database.

Part 3: Evidence for Validity of Inferences from Scores

Discussion of validity and reliability in this report is limited to the scoring of the portfolio assessments. Due to the range of commercially developed tests that were used and the lack of consistency of how the scores were reported no attempt was made to comment on the validity or reliability of the commercially available tests as an alternative assessment.

Evelyn S. Johnson from the University of Washington conducted a study titled the "Testing the Validity of an Alternate Assessment" (in press). This study examined the validity of Washington state's alternate assessment portfolio system using Messick's (1996) six aspects of validity evidence as a unified concept. These aspects are; content, substantive, structural, generalizability, external and consequential. Johnson notes that "To date, scant research evidence exists that alternate assessments contain the psychometric qualities required for state-level decision making." (page 4).

Results of Johnson's analysis indicates some shortcomings in the evidence for content, substantive and structural validity. However, since the portfolio system was in its first year of implementation at the time of this study, results are interpreted with caution. Johnson also presents options for improving the validity of the assessment systems together with the implications of these options.

Part 4: Reliability

The reliability of assessment scores is a measure of the degree to which the scores on the test are a "true" measure of the examinees' knowledge and skill relevant to the tested knowledge and skills. There are several ways to obtain estimates of score reliability: test-retest, alternate forms, internal consistency, and generalizability analysis are the most common. Test-retest estimates require administration of the same instrument at different times. In a sense a portfolio system is a collection of evidence from a full school year and as such should increase the reliability of the student's ability. However, no evidence was collected to confirm this speculation. Alternate forms reliability estimates require administration of two parallel assessments. These tests must be created in such a way that we have confidence that they measure the same domain of knowledge and skills using different items. Unfortunately at this time there is only set of evidence collected

The scoring design for the 2001 assessment did not readily allow for estimating the rater variance component. However, options for a special data collection for subsequent assessments could be determined. In the future, additional reliability evidence can be reported through the use of generalizability theory to analyze portfolio scores. In preparation for such an analysis, preliminary discussions would be needed. Some of the discussion topics would be related to identifying the facets of the measurement process (e.g., raters, items, etc.) and their characteristics (e.g. fixed or random, nested or crossed). More generally, the universe of admissible observations would need to be defined. For

example, does the state feel the progress scores in each content area and/or the holistic (Part 2) scores are exchangeable? A key outcome of such a study would be the variance component estimate for the rater facet (i.e., the amount of variance in the portfolio scores that can be directly attributed to the raters). A fully crossed design would allow for this estimate.

For 2001, inter-score agreement and coefficient alpha were two internal consistency measures used to estimate score reliability.

Inter-Scorer Agreement

Inter-scorer agreement is an important source of evidence for the reliability of test 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. To determine the degree to which judges gave equivalent scores to the same student work the percent of agreement between scorers was examined. Reliability of scoring was determined by looking at the difference between the score from the teacher scorers and the table leader scorers (Table 4).

Table 4: Percentage Agreement Between Teacher and Table Leader Scores

Amount of Agreement	Percentage			
	Day 1	Day 2	Day 3	Day 4
Scores exact the same	68.6%	71.9%	65.0%	63.6%
Scores are different by 1	27.6%	24.4%	30.3%	29.1%
Scores are different by 2	3.7%	3.3%	4.5%	7.3%
Scores are different by 3	0.2%	0.4%	0.2%	0.0%

Periodically during each day, and at the end of each day, the scores reliabilities for each teacher and for the group were calculated by each dimension/trait. On day 2, 3 and 4 portfolios scored by teachers who had 70% or above exact agreement received one score. However, every third portfolio was still double-scored by a table leader or assistant table leader to check on reliability. In these cases the table leader scores were still used as the final score. If teachers fell below 70% exact agreement, the portfolios they scored were double-scored on every portfolio. When reliabilities were low by trait, then the scorer would be re-trained.

On the second day one scorer grew more uncomfortable as the activities progressed. He stated that he had philosophical differences with assessing the children. His reliabilities decreased as the week went by, and his scores were never used for scoring the children, but they were included in the overall reliability figures.

Scoring was completed by noon the fourth day.

One further analysis was done to confirm consistency of scorer judgment. One of the OSPI staff who was not a scorer did a “polish” rating of the portfolios. Unlike a standardized test, the portfolios were extremely individualized, and highly dependent on the person (typically the teacher) who had assembled the portfolio. The portfolios were rated on a polish score of 1-3, with 3 being the highest score. This was used to represent the portfolios that were put together with the most care, while a 1 ranking went to the portfolios that were less well-assembled. Unfortunately, there was a clear difference in the scores received.

A total of 41 portfolios received a 3, 164 received a 2, and 210 received a 1.

Table 5: Polish Score by Grade

	Polish 1	Polish 2	Polish 3
Grade 4	83	65	16
Grade 7	72	50	19
Grade 10	55	49	6

Table 6: Polish by Total Score

Polish	Dimension Score			
	Total Communication	Total Reading	Total Writing	Total Math
1	6.11	6.21	6.21	6.20
2	8.16	8.30	8.15	8.25
3	12.34	12.32	12.10	12.07

In a traditional large-scale assessment, there is a chance that a student’s handwriting or neatness could influence a scorer’s impression of the student’s work. In the case of the portfolios, the skill, training and ability of the person putting the portfolio together could influence the scorer’s impression and score. As teachers gain more experience in this activity, it is likely that the polish of the portfolios will even out, and it will be possible to ensure that the scores received are not unduly influenced by the presentation of the portfolio.

Coefficient Alpha

Coefficient Alpha is a score reliability index of internal scale consistency/homogeneity. Alpha can be estimated from scores obtained on one occasion and is appropriate when a score is intended to measure a single trait. Table 7 provides the Coefficient Alpha for the Total scores and Part II scores. As indicated in the associated formula, the value of Alpha is affected by the number of components making up a score, the variance of the individual components, and the total score variance. In the context of the WAAS Total scores and Part II scores, relatively higher values of Alpha will tend to result when the total scores have greater variability and/or the scores across the individual components are very similar (i.e., internally consistent). This reliability index is only sensitive to random errors associated with this source of score variability. It does not incorporate

temporal errors (as would a test-retest reliability index) or random error associated with rater variance (addressed elsewhere in this document). Systematic sources of variance, such as rater effects, might artificially increase these values.

$$\text{Alpha} = (N/N-1) * (1 - \Sigma\text{Var}(\text{part})/\text{Var}(\text{total}))$$

Where N = Number of components combined to form total
 $\Sigma\text{Var}(\text{part})$ = Sum of the variance for the individual components
 $\Sigma\text{Var}(\text{total})$ = Variance of the total scores

Table 7: Coefficient Alpha for Total Scores and Part II Scores

	Total Score				Part II
	Math	Reading	Communication	Writing	
Grade 4	.77	.79	.79	.77	.81
Grade 7	.81	.82	.83	.82	.86
Grade 10	.72	.75	.75	.72	.73
All Grades	.78	.80	.80	.79	.82

Part 5: Description of Performance of Students

Table 8 provides a summary of the percentage of students obtaining each of the scale scores in each dimension that was scored plus the mean and standard deviation for each dimension. As can be seen from this table the majority of students were given a scale score of 1 for most dimensions except for the dimension modifications. At grade 10, no students were awarded a score of 4 on six of the eight dimensions. The mean for modifications is considerably higher than for the other dimensions. The scores awarded for modifications are more spread out than for the other dimensions as well.

Standards have not been established for these scores so it is not possible to judge whether the students are demonstrating satisfactory progress or not. Standards are to be established in the fall of 2002. However, before attempting to establish standards it may be appropriate to review the scoring criteria to see if it is possible to award more scores at the different scale points.

Table 8: Percentage of Students Obtaining Each Score on the Portfolio and Average Score By Grade

Grade 4

		Part I				Part II			
		Communication	Reading	Writing	Math	Modifications	Settings	Social Relations	Self Determination
Percentage of Students Obtaining Each Score	0	12.3%	9.6%	10.6%	9.6%	0.6%	0.6%	0.6%	0.6%
	1	61.0%	55.4%	59.0%	61.8%	23.4%	62.0%	66.0%	71.1%
	2	18.8%	24.8%	24.8%	21.0%	36.1%	24.1%	21.4%	22.0%
	3	7.8%	8.3%	5.0%	7.0%	24.1%	10.1%	10.7%	5.0%
	4	0.0%	1.9%	0.6%	0.6%	15.8%	3.2%	1.3%	1.3%
Mean Score		1.22	1.38	1.26	1.27	2.31	1.53	1.46	1.35
Standard Deviation		0.76	0.84	0.74	0.76	1.02	0.81	0.74	0.64
Number of Portfolios Scored		154	157	161	157	158	158	159	159

Grade 7

		Part I				Part II			
		Communication	Reading	Writing	Math	Modifications	Settings	Social Relations	Self Determination
Percentage of Students Obtaining Each Score	0	15.3%	11.1%	15.0%	12.1%	4.2%	4.2%	4.2%	4.2%
	1	51.8%	55.6%	57.1%	58.2%	23.6%	61.1%	66.0%	66.0%
	2	23.4%	24.3%	20.4%	22.0%	27.8%	23.6%	22.2%	23.6%
	3	8.8%	8.3%	7.5%	7.8%	28.5%	8.3%	5.6%	5.6%
	4	0.7%	0.7%	0.0%	0.0%	16.0%	2.8%	2.1%	0.7%
Mean Score		1.28	1.32	1.20	1.26	2.28	1.44	1.35	1.32
Standard Deviation		0.86	0.81	0.79	0.77	1.12	0.82	0.74	0.68
Number of Portfolios Scored		137	144	147	141	144	144	144	144

Grade 10

		Part I				Part II			
		Communication	Reading	Writing	Math	Modifications	Settings	Social Relations	Self Determination
Percentage of Students Obtaining Each Score	0	19.8%	19.3%	23.3%	16.8%	0.9%	0.9%	0.9%	0.9%
	1	63.1%	61.4%	62.1%	66.4%	34.9%	71.7%	84.0%	81.1%
	2	12.6%	15.8%	12.9%	13.3%	31.1%	20.8%	13.2%	16.0%
	3	4.5%	3.5%	1.7%	3.5%	24.5%	4.7%	1.9%	1.9%
	4	0.0%	0.0%	0.0%	0.0%	8.5%	1.9%	0.0%	0.0%
Mean Score		1.02	1.04	0.93	1.04	2.05	1.35	1.16	1.18
Standard Deviation		0.71	0.70	0.66	0.67	0.99	0.68	0.44	0.46
Number of Portfolios Scored		111	114	116	113	106	106	106	106

References

Johnson, E. S., Arnold, N. and Anderson, D. (in press) Testing the validity of an alternative assessment.

Messick, S. (1996) Validity and washback in language testing. Educational Testing Services: Princeton, NJ (ED 403277)