

RELIABILITY, VALIDITY, AND RELATED ISSUES PERTAINING TO THE WASL

Revised January, 2003

The reviews of WASL research presented in this report were compiled by John Brickell and Doris Lyon, researchers at the Washington Education Association. The final report was written by John Brickell.

John Brickell, Ph. D.

Professor of Research & Statistical Methods/Measurement Theory, Illinois State University
Testing and Research Analyst, OSPI
Dean of the College of Education, Pacific Lutheran University
Research Specialist, WEA

Doris Lyon, Ph. D.

Assistant Professor of Sociology/Anthropology/Research Methods, Denison University, Ohio
Planner, Port Gamble S'Klallam Tribe, WA
Research Analyst, WEA

RELIABILITY, VALIDITY, AND RELATED ISSUES PERTAINING TO THE WASL

The following report is based upon the review of seventeen technical reports and related studies pertaining to the psychometric properties of the 4th, 7th, and 10th grade Washington Assessment of Student Learning (WASL) assessments. The technical reports cover the years, 1997, 1998, 1999, and 2000. The other studies included in this review were selected on the basis of their meeting the criteria of, 1) using statewide data, and 2) being studies that dealt with the issues of the reliability and validity of the WASL assessments. The studies reviewed for this report were:

“Evaluation of the Washington State EALRs for Mathematics, Reading, and Science, ” (three separate reports) Mid-continent Research for Education (McREL), December 2002

“A Review of the Washington Assessment of Student Learning in Mathematics: Grades 7 and 10,” SRI International, November 30, 2002

“Study of Grade 4 Mathematics Assessment,” Northwest Regional Educational Lab, Sept., 2000

“Evidence for the Validity and Reliability of the WASL,” Washington PTA presentation, Cathy Taylor, April, 2002

“Validity Evidence for the WASL Reading and Mathematics,” Joe Willhoft and Duncan MacQuarrie, Tacoma School District, March, 2002

“The Relationship Between WASL Scores and UW Performance,” Debbie McGhee, Office of Educational Assessment – UW, July 2002

“The Relationship Among Achievement, Low Income, and Ethnicity Across Six Groups of Washington Students,” Martin Abbott and Jeff Joireman, Washington School Research Center, July, 2001

“The Relationship Between the Iowa Test of Basic Skills and the Washington Assessment of Student Learning in the State of Washington,” Jeff Joireman and Martin Abbott, Washington School Research Center, Sept., 2001.

SUMMARY OF FINDINGS

Technical reports and independent studies were reviewed and analyzed for the purpose of determining the reliability and validity characteristics of the Washington Assessment of Student Learning (WASL) tests. The following is a summary of the key findings of this review. The full report is attached.

- Reliabilities for the WASL test have generally reached traditionally reported levels found for standardized achievement tests. (see p. 3 of report)
- The reliabilities for some of the WASL tests, e.g., the 2000 10th grade Writing and Listening tests, are NOT high enough to warrant decisions regarding individual student achievement levels. (see p. 3 of report)
- Cut-scores and standard settings were guided by the directive of what, “a well taught, hard working student should be able to do in the spring of the 4th, 7th, or 10th grade.” Cut-scores were not guided by what a student should “be able to do” to earn a high school diploma. (see p. 5 of report)
- One independent study recommended that the WASL performance standards be re-set and that more recently developed and currently accepted procedures for establishing performance standards be implemented in the process of re-setting the WASL cut-scores. (see p. 10 of report)
- Reported correlations in the technical reports and other studies reviewed suggest that both the ITBS/ITED tests and the corresponding WASL tests are measuring similar content areas. (see p. 7 of report)
- The Standard Error of Measurement for the WASL tests is too large to warrant the use of the reported WASL score alone when making decisions regarding a student’s academic achievement level(s). (see p. 4 of report)
- In the area of test validity, the reports and studies conducted or commissioned by the OSPI have been directed to providing evidence of content and construct validity. See pp.4 & 5, and pp. 10 & 11 or report).
- There have been no studies of the predictive validity of the WASL tests conducted or commissioned and reported by the OSPI. There appears to be no plans by the OSPI to conduct such studies. (see p. 10 of report)
- An independent investigation conducted by the University of Washington on the predictive validity of the WASL found that although the WASL was the least predictive of Freshman GPA among such factors as high school GPA, SAT and ACT scores, the magnitude of the predictive validity of the WASL was similar to the SAT and ACT (see p. 8 & 9 of report)

- There have been no studies of consequent validity conducted or commissioned and reported by the OSPI, specifically for such student characteristics as, ethnicity, race, gender, primary language, or economic status, as well as for students with special learning needs. (see p. 12 of report)
- Content validity of the WASL has been established through the professional judgment by members of such groups as the “Content Committees” , matching the test item to the Essential Academic Learning Requirements (EALRs). (see p. 5 of report) **However,**
- An independent study identified several WASL test items on the 4th grade math test that did not align with the grade 4 level EALRs, while another independent study by found that there was an uneven representation of test items linked to specific EALRs, e.g., 78% to 88% of the test items, and 70% of the student’s test score, were attributable to EALR 1 for the 7th and 10th grade 2000 Mathematics WASL, while there were no test items linked to EALR 5. (see pp. 5-6 of report)
- Consistently high correlations between the WASL Mathematics test and the WASL Reading test, across years and grade levels, suggest that WASL mathematics tests have a high reading component. (see pp. 7 & 8 of report)
- There have been no studies conducted or commissioned and reported by the OSPI regarding the validity of the cut-scores (passing scores), particularly for the purpose of using the passing of the WASL as a requirement for high school graduation.
- One independent study concluded that 7th and 10th grade WASL mathematics tests and the corresponding passing scores were at a level that was “...2 or 3 grades above the intended test grade level.” (see p. 9 of report)
- There is not a consistent level of performance expectations across WASL grade-levels and content areas. (see p. 10 of report)
- The percentage of 10th grade students “passing” all four WASL tests has leveled at 30%.
- The percentage of 10th grade students “passing” 2000 WASL Mathematics test was: Males = 35.5%, Females = 33.0%, African American = 10.8%, Native American = 16.2%, Asian/Pacific Islander = 40.4%, Latino/Hispanic = 11.8%, White/Caucasian = 38.7%.
- The percentage of students performing at Level 1 (the lowest level of achievement) on the 10th grade 2000 Mathematics WASL was: African American = 56.3%, Native American = 48.5%, Asian/Pacific Islander = 29.6% , Latino/Hispanic = 58.1%, White/Caucasian = 30.2%.l

UNDERSTANDING THE TERMS USED IN RELIABILITY AND VALIDITY STUDIES

Technical terms such as reliability, validity, norm-referenced and criterion referenced tests, together with statistical analyses like Factor Analysis, Correlation, and Regression, can be confusing to the non-technically trained or experienced. The discussion that follows will, hopefully, assist the non-technically trained to have a better understanding of these terms and how they relate to the qualities and characteristics of tests like the WASL.

CRITERION AND NORM REFERENCED TESTS

Criterion-referenced tests specify an *a priori* level of performance (cut-score) that indicates that the test-taker has either reached “criterion” or has not. Therefore, with criterion-referenced tests we are typically interested in only two groups of students: those who achieve “criterion,” i.e., “pass,” and those who do not. And the follow-up question for those who do not “pass” is what do we do to get them to criterion.

In contrast, **norm-referenced** achievement tests are purposefully designed to maximize the differences of achievement between all test takers (i.e., maximizing the statistical variance); individuals are compared to each other and not to a predetermined criterion performance level. Therefore, the selection of items for criterion- and norm-referenced tests, as well as the intended purposes for these two types of test, can differ significantly.

For **norm-referenced** tests the difficulty level of the test items will tend to be at the .5 level or close to that value. (A difficulty level of .5 means that 50% of the test takers answered the test item correctly.) Achievement tests composed of items at the .5 difficulty level will maximize the variability (statistical variance) among test takers, which is one of the objectives of a norm-referenced achievement test.

For **criterion-referenced** tests, the criterion performance level is determined by standard setting “experts” and there are a variety of techniques to arrive at this level, i.e., the cut-score; for the WASL a modified form of the “bookmark procedure” developed by the staff from CTB/McGraw-Hill was used. Clearly, the selection of the items for a criterion-referenced test will be influenced by the standard setting process that eventually results in the determination of the cut score for attaining criterion or “passing”. The difficulty level of the items will be either above, below, or at the .5 level depending, at least in part, on the performance level selected by the “experts,” i.e., the level of performance expected of students who can meet the predetermined criterion level of achievement.

For a **norm-referenced** achievement test, the “norm” is determined by the resultant performance of those taking the test (not by a panel of standard setting experts as for

criterion-referenced tests) and it is the statistical mean of the student scores that becomes the “norm,” so to speak. For achievement tests like the ITBS and ITED that are used in Washington’s assessment of student academic performance, these norms are “national” norms based upon a national sample of students at the appropriate grade levels. Washington state’s students are then compared to these national norms, or the “average” (statistical mean) math, reading, etc. level of those students in the national sample. Percentile ranks, as well as grade-equivalents, are typically used to indicate a student’s level of performance relative to other students, be they local, state, or national. In contrast, a student’s level of performance on a **criterion-referenced** test is compared to the pre-determined criterion level for “passing,” or cut-score, and not to the performance of the other students (national or otherwise) who take the test.

As discussed above, the “criterion,” or cut-score, is determined *a priori*, by a group of “experts” who have determined that a performance at or above a given level represents attainment of a desired/important level of performance. **This level of performance is generally viewed as being sufficient for successful accomplishment of a future task(s), i.e., sufficient mastery of a given content of knowledge and/or skill that is necessary for success on some future outcome(s) that is (are) also deemed as important.** This speaks to the importance of establishing the predictive validity for a criterion-referenced test, especially if the test is to be used in making high stakes decisions like awarding a high school diploma.

RELIABILITY

Reliability refers to the “consistency” or accuracy (lack of error) in the measurement of a characteristic, e.g., student academic achievement. This “accuracy” also includes the inter- and intra-rater consistency of scoring open-ended/essay type items.

What reduces reliability is the extent to which errors enter into performance scores. For example, the extent to which guessing can be a factor in taking a test (e.g., multiple choice items) is the extent to which reliability will be reduced, as is the lack of consistency in scoring open-ended/essay type items.

Reliability is not an “all or nothing” characteristic; so we speak of the level/degree of reliability that a test evidences. Statistically, reliability can range from “zero” to “one”. A test with a calculated reliability coefficient of $r = 0.0$, would be indicative of a test with a complete lack of reliability. At the other extreme, a test with a calculated reliability coefficient of “1.0” would be indicative of a test that is completely reliable, i.e., a test that measures a characteristic, like academic achievement, without error.

One question that is often asked about a test’s reliability is, “How reliable should a test be, before it is useful.” One testing and measurement text author has reported the following regarding the **minimum level** of reliability for decision making at the **individual level**:

“To evaluate level of **individual accomplishment**, reliability should be **at least .94**. In contrast, the minimum level of reliability to evaluate the level of **group accomplishment**, would be **only .50**.” [bold added for emphasis] (Helmstadter, Principles of Psychological Measurement, p. 84)

WASL ISSUES

From the 1998, 1999 and 2000 OSPI Technical Reports the **reliability estimates** for the Listening, Reading, Mathematics, and Writing WASL tests at the 4th, 7th, and 10th grades, have ranged from a low of **r = .56** for the 1999 7th grade Listening test, to a high of **r = .93** for the 1999 10th grade Mathematics test, with most of the reliability estimates being in the .7 to .9 range.

The reliability estimates for the 2000 **10th grade** WASL tests were: Listening $r = .62$, Reading $r = .90$, Mathematics $r = .92$, and **Writing $r = .76$** . The Reading and Mathematics reliabilities approach the level ($r = .94$) suggested in the text by Helmstadter for individual decision making, while the reliabilities for the Listening and Writing tests are considerably lower.

Conclusion: Reliabilities for the WASL tests have reached traditionally reported levels found for standardized achievement tests. However, it can NOT be concluded that the reliability of the 2000 10th grade Writing test is sufficient for decisions regarding individual student achievement levels for high-stakes decisions such as the awarding of a high school diploma.

STANDARD ERROR OF MEASUREMENT

The standard error of measurement (SEM) is a function of the reliability of the test. If the reliability of a test is, $r = 1.0$, then the SEM would be “zero,” i.e., measurement without error. Test theory assumes that a student’s reported achievement score (what is called the “observed score”), for example, is composed of two components, the student’s “true score” and an error component (measurement error).

We generally speak of a student’s “true score” as being the student’s observed score, “plus or minus” the estimated SEM. [Note: More accurate, however, would be the statement, “There is a **68% probability** that a student’s ‘true score’ is his/her observed score plus or minus the SEM.]

For example, the reported SEM for the 7th grade Mathematics WASL (1999) is 15.59. Therefore, if a student’s reported score (what we call the “observed score”) was 280 for the 1999 WASL Mathematics test, we could say that there is a 68% probability that the student’s true score (“real” performance level) is 280 plus or minus 15.59, or between

295.59 – 264.41. [Note: If we wanted a 95% probability, then we would have to use “plus or minus” two SEMs, or for our example above we would have, 280 plus or minus 31.18, 311.18 – 248.82].

WASL ISSUES

The magnitude of the SEMs combined with the lower than suggested reliabilities for making individual student decisions regarding achievement levels, strongly argues for the use of the SEM in estimating an **individual student’s** level of achievement as determined by his/her WASL score.

The reported SEMs have been fairly consistent across the testing years and grade levels. The SEM for the WASL Listening has ranged from 27.6 to 35.84; for Reading, from 6.15 to 9.6; for Mathematics, from 11.14 to 14.19; and for Writing, 22.13 to 24.0.

In an independent study of the **4th grade WASL Mathematics** tests, researchers/reviewers from the Northwest Regional Educational Laboratory concluded, “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.**” [bold and underlining added for emphasis] (p. 74)

Conclusion: The Standard Error of Measurement for the WASL is sufficiently large to suggest **caution** when a student’s reported WASL scores **alone** are used to make decisions regarding a student’s measured academic achievement levels.

VALIDITY

Questions concerning issues of whether the test measures what it purports to measure and whether it is appropriate for an intended purpose or use are questions of validity. Since there can be different purposes or intended uses of a test, it should not be surprising to find that there is more than one type of validity evidence that can be established and reported for any specific test.

Although there has not been uniform agreement among test specialists regarding the different types of validity that can and/or should be reported to establish test validity, types of validity have been historically classified into three, four, or for some authors, even more categories. A typical classification scheme is: content validity, concurrent validity, empirical/predictive validity, and construct validity. Even with these four types, some authors have viewed some of these categories as subsets of the others. Recognizing the potential for over simplification, the following definitions are offered:

Content validity	Evidence (via professional judgment) that the contents of a test is composed of, and thus, represents items from
-------------------------	--

the intended content area. One critical issue in establishing content validity is in the sample of items selected for the test, relative to all items that could have been selected and included in the test. Are the items selected for the test **representative** of the achievement domain being tested, both in content as well as level of difficulty?

WASL ISSUES

Selecting items and setting standards were guided by the directive of what a “well taught, hard working student” should be able to do in the spring of the 4th, 7th, or 10th grade.

Evidence for content validity for the WASL tests has been approached through a comparison of test items against the Essential Academic Learning Requirements. Items were reviewed and revised by the **Content Committees** (20 – 25 persons), “...to ensure that they measured Washington’s Essential Academic Learning Requirements both accurately and comprehensively.” (OSPI’s 1997 Technical Report, p.3)

In reviewing the 1998, 1999, and 2000 **4th grade Mathematics WASL** test items, an independent review team found that 10 of the 120 items on the tests were not aligned with the 4th grade EALR benchmarks. (“Study of the Grade 4 Mathematics Assessment,” p. 30)

A study conducted by SRI International of the 7th and 10th grade Mathematics WASL tests found that between 78% and 88% of the test items, and 70% of the student’s test score, were attributable to EARL 1 (Concepts and Procedures of Mathematics). The remaining 12% to 22% of test items were spread across EALRs 2, 3, & 4. They also noted that there were no items linked to EARL 5 (mathematical connections). (SRI International report, p.31)

The SRI study also found that there were EALR “Components” that were not linked to any test items. For example, EALR Component - 3.2, “Uses mathematical reasoning to predict results”, was not linked to any items on the 7th and 10th grade Mathematics WASL. (SRI International report, p. 34)

There has been a decrease in the number of mathematics items on the 7th and 10th grade WASL tests that employ a context within the item question. The SRI researchers recommend that, “OSPI and the test developer pursue further study on the relationship between item context and item difficulty.” (SRI International report, p.44)

A study by Tacoma School District’s researchers concluded, “The findings from the present investigation raise fundamental questions about the reasonableness of the performance standards associated with the WASL reading and mathematics tests.” They also assert the **7th and 10th grade WASL** passing standards are “...more like that of students 2 or 3 grades above the intended grade level.” (“Validity Evidence for WASL Reading and Mathematics Performance Standards,” pp.16 - 17). Their findings raise questions about the appropriateness of the **level** of the content of these WASL tests.

Conclusion: Although OSPI’s technical reports refer to the process of linking test items to the content strands of the EALRs, which addresses the issue of content validity of the WASL tests, there has been very little, and for some WASL tests no, independent analysis of the linking of test items to EALR content strands. An **independent analysis of the content validity** of the various WASL tests seems reasonable and advisable. The study by SRI International revealed a very uneven representation of test items linked to specific EALRs for the 7th and 10th grade mathematics WASL. (SRI International report, pp.30 – 38)

Additionally, the difficulty-level of test items and the validity of the cut scores appear to be legitimate areas of concern, and speak to the need for further research. More on these areas of concern are addressed later in this document.

Concurrent validity Evidence that a test measures what it purports to measure by correlating it with an existing test that has been used (and accepted) to measure the same or similar content (using the same students for each test). And conversely, evidence that a test does **not** correlate (or correlates only moderately) with a test that measures a different content area.

WASL ISSUES

The WASL tests are given to 4th, 7th, and 10th grade students in the spring of their academic year, while the ITBS is given to 3rd and 6th grade students, and the ITED to 9th grade students. Thus, students do not take the WASL tests and the ITBS or ITED during the same grade level, so that when correlations are calculated between these tests, they are not technically “concurrent” measurements of academic achievement. However, the results of such analyses are included here, and in OSPI’s technical reports, since they represent the best and closest approximation to such “concurrent” testing as is available.

These correlations have been reported for both between and within comparisons of the WASL tests and the ITBS and ITED, for each of the separate content area tests, as well as for the sub-sections contained within each of these content area tests. The results of these analyses generally reveal that these WASL tests and the ITBS and ITED are moderately to highly correlated between each matched content area, i.e., math to math, reading to reading. For example, the correlation between 1998 7th grade WASL-Math and the 1999 8th grade ITBS-Math (for the same students) was $r = .703$. The corresponding correlation for the reading tests was $r = .678$. When the 2001 10th grade WASL – Math results were correlated with the 9th grade ITED – Math results, the correlation was $r = .796$, while the corresponding correlation for Reading achievement was $r = .744$. **Correlations of this size are consistent with the expectations that both the ITBS/ITED tests and the corresponding WASL tests are measuring similar content areas.**

These correlations have been fairly consistent, which suggests that relative success, or lack thereof, on the WASL is indicative of, or predictive of, relative success, or lack thereof, on the corresponding nationally normed achievement test, and conversely.

One interesting observation, however, is the **consistently high correlation between the WASL Mathematics test and the WASL Reading test**. These correlations have tended to range in the mid .70s, across years and grade levels, e.g., $r = .745$ between the 2001 10th grade Math and Reading tests. Results such as these have led Cathy Taylor, the lead analyst and author for OSPI's technical reports, to state, "These results show a stronger than expected relationship between reading and mathematics within ITBS/ITED, within WASL, and sometimes, between WASL and ITBS/ITED. This requires follow-up studies to investigate two competing explanations for results: a) Total scores are really measures of general ability [or] b) Both the ITBS/ITED and WASL demand a great deal of reading in the mathematics tests." ("Evidence for the Validity and Reliability of the Washington Assessment of Student Learning," Catherine Taylor, April 26, 2002).

Speaking to the reading level required of the 4th grade mathematics test, an independent study of the 4th grade mathematics WASL test concluded, **"...reliable data on the readability of the mathematics test is lacking."** And continues with the observation, **"If the reading level is too high, the test may confuse reading ability with mathematics ability."** [bold added for emphasis] ("Study of the Grade 4 Mathematics Assessment," September 2000, p. 46)

Note: Using student performance on the WASL as an accountability measure for either teachers specifically, or an entire school's staff, is called into

question regardless of which of the two hypotheses proposed by Taylor is accepted: If “Total scores are really measures of general ability” , then the tests are measures of academic ability, rather than academic achievement that can be more directly linked to instruction. If, on the other hand, the “...WASL demand[s] a great deal of reading in the mathematics tests,” then the WASL mathematics tests should not be considered “valid” measures of mathematics knowledge and skills **tied directly and independently to mathematics instruction.**

**Empirical/
Predictive Validity**

Evidence that test scores can be used to predict some outcome of predetermined importance. One of the most common examples of predictive validity is using SAT or ACT scores to predict the GPA of first year college students. As one measurement expert has stated, “...test results, if they are to be useful (as contrasted with just being interesting) must, in the final analysis, be predictive.” Helmstadter, p.130; and, **“It is the principle for making inferences from scores and not the test itself which is validated.”**

WASL ISSUES

To date, there have been no predictive validity studies conducted, reported, and then released by OSPI. The information, contained in the technical reports produced by OSPI, has been **limited to providing evidence of Content and Construct validity for the various WASL tests.**

One independent study that specifically looked at the predictive validity of the WASL was conducted by the Office of Educational Assessment at the University of Washington (“The Relationship between WASL scores and UW Performance in the First Year,” July 17, 2002). This study used 2,682 first year UW students who had taken the WASL test in 1999 in the 10th grade. This study investigated the relationships between and among, High School GPA, UW first year GPA, WASL Listening, Reading, Writing, and Math scores, SAT Verbal and Math scores, ACT scores, and grades in selected UW courses.

High School GPA had the highest correlation with UW GPA, $r = .452$. The correlations for the WASL tests with UW GPA were: Listening, $r = .21$; Writing, $r = .25$; Reading, $r = .32$; and Math, $r = .35$. When WASL Math and Reading were combined into a “Total” score, the correlation with UW

GPA was, $r = .38$, which was practically identical to the correlation for SAT Total with UW GPA, which was, $r = .39$.

Thus, the WASL had similar predictability of first year UW GPA as did the SAT, but was less predictive than a student's high school GPA. If high school GPA, SAT Total, and WASL Reading+Math are used to predict first year UW GPA, the most predictive measure was HS GPA, the "best" two predictors in combination were HS GPA and SAT-Total; and knowing a student's WASL Reading+Math score did not appreciably improve predictability of first year UW GPA when a student's HS GPA and SAT scores were known.

Validity of the Cut-Scores

To date, there have been no validity studies of the performance standards (cut-scores) conducted, reported, and then released by OSPI.

An independent study by MacQuarrie and Willhoft (2002), of the Research and Evaluation Office of the Tacoma Public Schools, using statewide student data, investigated the reasonableness of the performance standards in reading and mathematics at the elementary, middle, and high school levels.

MacQuarrie and Willhoft addressed the issue of the difficulty of the WASL cut score by using **3rd grade ITBS** scores from 1999 and **4th grade WASL** scores from 2000. Their investigation found, "...that it isn't until you get to students scoring above the 60th [percentile rank] on the ITBS total mathematics that you even find half of them subsequently meeting the WASL [passing standard]." Students performing at the 60th percentile rank represent student performance at or above 60% of those students used to set the national norms for the ITBS.

They concluded that, "The findings from the present investigation raise fundamental questions about the reasonableness of the performance standards associated with the WASL reading and mathematics tests." (p. 16); and continue, "Across the three WASL testing levels the mathematics performance standards are associated with normative performances that are more like that of **students well above what we typically expect**. This is particularly true at grade 7 and 10 where the linked normative performance is **more like that of students 2 or 3 grades above the intended test grade level**." [bold added for emphasis] (p.17)

They further noted, "Of the 29,164 students in the 4th grade data set who met the WASL reading standard in 1998, 10,101, or 35%, failed to meet

the reading standard three years later as 7th graders.” ... “This pattern does not seem to be reasonable and raises questions about the [appropriateness of the] 7th grade reading standard.” (MacQuarrie and Willhoft, p,16)

An independent study, commissioned by OSPI, conducted by SRI International found that there were uneven performance expectations between the 7th and 10th grade mathematics WASL tests: 15% of the 7th grade items were judged to be at a “higher-grade-level” benchmark, while 22% of the 10th grade test items were judged to be at a “lower-grade-level” benchmark. (SRI International report, p.39)

The SRI study noted that, “...the recognition that performance standards on high-stakes tests are going to be challenged, so that the compilation of validity evidence [for the performance standards] is critical.” (SRI International report, p.72)

The SRI study recommends the re-setting of the cut-scores and makes very specific recommendations for the procedures to be used in the re-setting of the cut-scores (performance standards) for the WASL assessments. (SRI International report, pp.72 – 78)

Conclusion: The validity of the cut scores, particularly when used as part of the requirements for a high school graduation requirement, has not been established, although one independent study has strongly suggested that such validity evidence be provided (see SRI International report, pp. 74 – 76). There is evidence that the performance standards/cut scores may be inappropriately high for some of the WASL grade level and subject matter tests, and the WASL Mathematics tests, in particular, have been judged to have higher performance standards than the other WASL tests.

Construct validity Evidence that is not just used in an evaluation of the test itself, but also of the theory and concept of the trait being tested as well. Evidence of construct validity generally includes the establishment of the other types of validities listed above, and may include the use of the statistical procedure called Factor Analysis.

With respect to the WASL, it is instructive to note that, “Excuses that no really adequate criterion for validation purposes exists are not acceptable, and rationalizations about a trait which fail to result in observable consequents cannot be considered construct validation.” (Helmstadter, p. 137)

WASL ISSUES

Correlations between and among the WASL tests and sub-tests, ITBS, ITED, and CTBS tests, together with the statistical procedure of Factor Analysis, have been used to provide evidence for Construct validity by OSPI in their technical reports. The intercorrelations among these tests and their respective sub-tests have been discussed above under the category of Content validity. As stated previously, establishing Content validity is part of establishing Construct validity.

As of this date, Factor Analyses have been conducted and reported by OSPI for the **1998 4th and 7th grade** WASL using Reading and Mathematics WASL tests, together with CTBS, ITBS, and TCS (Test of Cognitive Skills); **1999 WASL 4th, 7th, and 10th grade** Reading, Listening, Writing, and Mathematics tests only (no other achievement test used); and the **2001 WASL 10th grade**, together with the ITED (note: this 10th grade analysis appears in the report by Cathy Taylor for her presentation to the Washington state PTA, and is in narrative form only with no factor analysis tables). Additionally, Joireman and Abbott (2001), of the Seattle Pacific University's Washington School Research Center, conducted a Factor Analysis using the **2000 WASL 4th grade** and **1999 ITBS 3rd grade** test results, and the **1998 WASL 4th grade** and **2000 ITBS 6th grade** test results.

The authors of these technical reports and studies have suggest several "models" of measured academic achievement for the WASL depending on grade level, tests used, and year tested. The resultant components (Factors) have included, a mathematics factor, a logical reasoning factor, a reading factor, a language arts factor, and a general academic achievement/ability factor.

Furthermore, the authors suggest that these Factor Analyses tend to support a two- (e.g., mathematics, and reading or language arts) or three-factor (e.g., mathematics, language arts, and general academic ability) "construct" for the WASL and norm-referenced achievement tests used in these analyses. However, the Joireman and Abbott study found, in part, support for a one factor solution, e.g., general academic achievement/ability. They conclude, **"In sum, results from the replication group [second study] provide weaker support for the claim that Reading and Math are more clearly distinguishable on the WASL, in comparison to the ITBS/ITED"** [bold added for emphasis] (p. 7).

Conclusion: The consistently high correlations between the WASL Mathematics and Reading tests, combined with factor analytic results which have not been consistent in supporting a two-factor theory of WASL content between Mathematics and Reading, suggest that, at this time, it

cannot be concluded that the WASL Mathematics and Reading tests measure distinctly separate areas of academic content, or alternatively, there is a large component of reading in the mathematics tests. As authors of the study of the grade 4 math test observed, “...**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.**” [bold and underlining added for emphasis] (p. 46)

Thus, the most reasonable hypothesis, at this time, is that the WASL Mathematics test is heavily dependent on reading ability and/or the ability to communicate ideas in a narrative form, and should not be treated as an independent test of mathematic knowledge and skill.

A more recent area of validity evidence concern has been:

Consequent validity Cronbach (1988) and Messick (1989) both want evidence regarding consequences considered in the overall evaluation of validity.

Additionally, Robert Linn has stated, “If the argument that validation should include an evaluation of the consequences of the uses and interpretations of assessment results is accepted, then it is **not sufficient** [bold and underling added] to provide evidence that the assessments are measuring the intended constructs.

Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and, at the same time, not producing unintended negative outcomes.” (p. 8)

WASL ISSUES

One of the areas of greatest concern of test validity for the WASL, is for Consequent validity. This is particularly true if the WASL is to be used as a high school graduation requirement. Yet, this is the one area of validity that has received the least, if any, attention by the producers and authors of OSPI’s Technical Reports.

The developers of the WASL tests, from the item development and selection process, to the selection of the cut scores, were **not** instructed to make their decisions on the basis that the test was to be used as a high

school graduation requirement. A review or study of the validity of the cut scores has (to date) neither been attempted nor conducted, or in any case, reported. The decision to use the “passing” of the WASL tests as a high school graduation requirement came after the cut scores had been established for other purposes.

Conclusion: To date, no studies by OSPI, or otherwise, have dealt with the consequential validity of the WASL tests, specifically with respect to student personal and demographic factors of ethnicity, race, gender, primary language, or economic status, as well as for students with special learning needs.

To recap, validity concerns the intended purpose(s) or intended use(s) of a test. One cannot say, “This test is valid” period, but must also specify “valid” for what purpose, including for which group of students, especially in the case of the WASL.