

**Validity Evidence for WASL Reading and Mathematics
Performance Standards**

Duncan MacQuarrie
Manager of Student Assessment
Joe Willhoft
Executive Director, Research & Evaluation
Tacoma Public Schools

Paper Presented at the Annual Spring Conference of the
Washington Educational Research Association (WERA)

March 14, 2002
Seattle Airport Hilton

Abstract

The performances of schools and districts on the Washington Assessment of Student Achievement (WASL) have become the primary achievement indicators in the state accountability system. In addition, the WASL scores will become the indicators under the Title I accountability requirements of the recently reauthorized Elementary and Secondary Education Act (ESEA). An important dimension of the use of test scores for such purposes is the validity of inferences about student learning that are based on performance standards (cut-scores). In order for interpretations of scores based on such standards to be valid, the standards must be reasonable and fair. This study reports data that raise questions about the reasonableness of the performance standards in reading and mathematics at the elementary, middle, and high school levels. The reading standard at 7th grade deviates markedly from those at 4th and 10th grade. The mathematics performance standards are consistently difficult across the grade levels. However, the difficulty of the mathematics standards at grades 4 and 10 is much greater than is the difficulty of the standards for reading at those grades. These patterns raise additional questions about the reasonableness of the WASL performance standards.

Validity Evidence for WASL Reading and Mathematics Performance Standards

The Washington Assessment of Student Learning (WASL) is a standards-based assessment program designed to measure important aspects of the Essential Academic Learning Requirements (EALR), Washington's curriculum content standards. The EALR and the corresponding WASL assessments at grades 4, 7, and 10 have been implemented in compliance with state legislation originally enacted in 1992 and substantially revised in 1993. The EALR and the WASL have been reviewed by the United States Department of Education (USED) and certified as complying with federal requirements under Title I of the Elementary and Secondary Education Act (ESEA).

The WASL assessments, administered in the spring, were introduced successively at grades 4, 7, and 10 over a three-year period (1997 to 1999). As standards-based or criterion-referenced tests, performance standards were established for each of the component tests (listening, reading, writing, and mathematics) in the respective initial operational year. In reading and mathematics, four performance categories were established using a modified form of the "bookmark procedure" developed by staff from CTB/McGraw-Hill (Lewis, Mitzel, & Green, 1996).

In 1999 the state established elementary school goal-setting criteria based on the 4th grade WASL reading test. This was expanded in 2000 to include requiring school and district improvement goals in grades 4, 7, and 10 for WASL reading and mathematics.

It has been nearly five years since the first operational form of the WASL was administered to 4th grade students statewide. There now exists multiple years of data regarding the performance of students on the various WASL tests across years and grade levels. In addition, there are several cohorts of students with both norm-referenced test scores and WASL scores as well as two cohorts with 4th and 7th grade WASL scores and one cohort with 7th and 10th grade WASL scores. These rich data sets provide an opportunity to investigate various characteristics of the reading and mathematics performance levels.

Initial Interest and Questions

Several patterns have emerged in the state data that are generally reflected in district data. First, the percent of students meeting the performance standards in mathematics at grades 4, 7 and 10 is lower than the corresponding percent meeting the reading standard. In grades 4 and 10 the differences are quite large. Second, the rate of growth in the percent of students meeting the performance standards has decelerated dramatically in 4th grade, remained more or less flat at 7th grade, and been modest at 10th grade. (See Table 1.)

Table 1. Percent of students meeting WASL standard statewide.

	1997	1998	1999	2000	2001
4th Grade					
Reading	47.9	55.6	59.1	65.8	66.1
Math	21.4	31.2	37.3	41.8	43.4
7th Grade					
Reading		38.4	40.8	41.5	39.8
Math		20.1	24.2	28.2	27.4
10th Grade					
Reading			51.4	59.8	62.4
Math			33.0	35.0	38.9

The WASL performance has become the primary indicator in the public reporting of the health of the public schools (local media) and in the more formal reporting recommended by the Academic Achievement and Accountability Commission (A+ Commission). It has been acknowledged that the performance standards are challenging, and that their degree of difficulty reflects the significant changes in the school curriculum that they are intended to promote. In addition, considerable emphasis has been placed on the importance of making continuous improvement in the percent of students meeting the standards rather than the absolute performance in any particular year. It also has been recognized that changes in the educational system are complex and that only a sustained and long-term commitment will accomplish this goal.

One characteristic of a standards-based reporting system is the use of a limited number of performance categories. Students' performances on the WASL reading and mathematics tests are classified into two primary categories: Met Standard and Did Not Meet

Standard. Each of these primary categories is further divided into two categories. Students meeting the standard are classified as either “at the standard” (level 3) or “well above the standard” (level 4). Likewise, students who do not meet the standard are classified as either “below the standard” (level 2) or “well below the standard” (level 1).

Such systems inevitably produce groups of students scoring just one or two points below the critical score that would result in their being classified in the next higher category. School staff frequently are encouraged to review frequency distributions and to take solace that they really did better than the reports indicate because a fair number of their students were just one or two points away from reaching the next level. Such observations are frequently combined with encouragement to work a little harder next year with those students “close to the standard” so as to help them get that extra point or two. Of course hindsight is so much clearer. We can see exactly which students could have been the focus of our extra effort. However, is there a way of knowing which students might be right on that bubble next year?

At last year’s annual spring conference of the Washington Educational Research Association (WERA) there was a session titled “Some Statistical Techniques for Predicting WASL from ITBS.” In that session a description of the overlap in the coverage of the EALR by the Iowa Tests of Basic Skills (ITBS) and the WASL was given. In addition, the statistical relationship between the 3rd grade ITBS scores and the subsequent 4th grade WASL scores was described and a method was identified for evaluating the utility of predicting WASL scores from the prior year’s ITBS scores. During the discussion period following the presentations several members of the audience described how they were providing classroom teachers information about the expected performance of their fourth grade students on the WASL based on those students prior year’s ITBS performance. It was reported that teachers found such information to be helpful.

Following the conference, staff from the Research and Evaluation Office of the Tacoma Public Schools merged district student files containing the ITBS scores from 1999 and

WASL scores from 2000. The resulting database contained over 2,000 matched cases with both 3rd grade ITBS scores and 4th grade WASL scores and these data were used to investigate various statistical prediction models that might be potentially useful for providing teachers with some idea of their students' expected WASL scores. In addition, using a methodology similar to that described at last year's conference for evaluating the utility of score prediction systems, a more direct portrayal was also constructed. In this method, for each value or range of values for the ITBS national percentile rank (NPR) scores, the percent of the students within that score range who subsequently met the standard was calculated. Then a bivariate plot was constructed in a coordinate system where the prior year's ITBS scores were represented along the X or horizontal axis and the percent of students meeting the standard (or any other performance category) were represented along the Y or vertical axis. The percent of students meeting the standard within each ITBS score category was plotted to complete the portrayal.

Consider the simple example data in Table 2 and the following explanation.

Table 2. Select ranges in an example data set.

ITBS NPR Range	Frequency	N Met WASL Standard	Percent Met WASL Standard
1 – 4	129	16	12.4
5 – 9	129	14	10.9
10 – 14	95	20	21.1
.	.	.	.
.	.	.	.
.	.	.	.
50 – 54	90	65	72.2
55 – 59	145	113	77.9
60 – 64	188	153	81.4
.	.	.	.
.	.	.	.
.	.	.	.
80 – 84	71	69	97.2
85 – 89	71	67	94.4
.	.	.	.
.	.	.	.

The construction of the type of table shown above is straightforward. Each row pertains to an ITBS NPR score range. The second column contains the total number of students who scored in that NPR range as 3rd graders and the third column reflects the number of those students who subsequently met the WASL standard in 4th grade. The fourth column expresses the number meeting the standard as a percent of all students in that NPR range (or, the third column divided by the second column, expressed as a percent). Once such a table has been completed a simple histogram can be developed by plotting the values from column four (percent meeting standard) along the vertical axis for each corresponding value in column one (NPR band) along the horizontal axis.

The initial idea was to use these analyses as the basis for developing very general criteria related to 3rd grade ITBS performance that teachers and principals could use to identify students that might potentially be close to the WASL “meets standard” cut score of 400 in reading and mathematics. Looking at the data it appeared something reasonable could be developed for WASL reading, but the possibilities for mathematics were not as clear. The data suggest that it isn’t until you look at students scoring above the 60th NPR on the ITBS total mathematics that you even find half of them subsequently meeting the WASL standard. That is, the students likely to be very close to the math standard, and who with a little extra help could get over the cut score, were likely to be those students with above average basic skills performance coming into 4th grade. Such a message did not seem to have the potential to be helpful, particularly when the comparable reading basic skills score was closer to the 35th NRP.

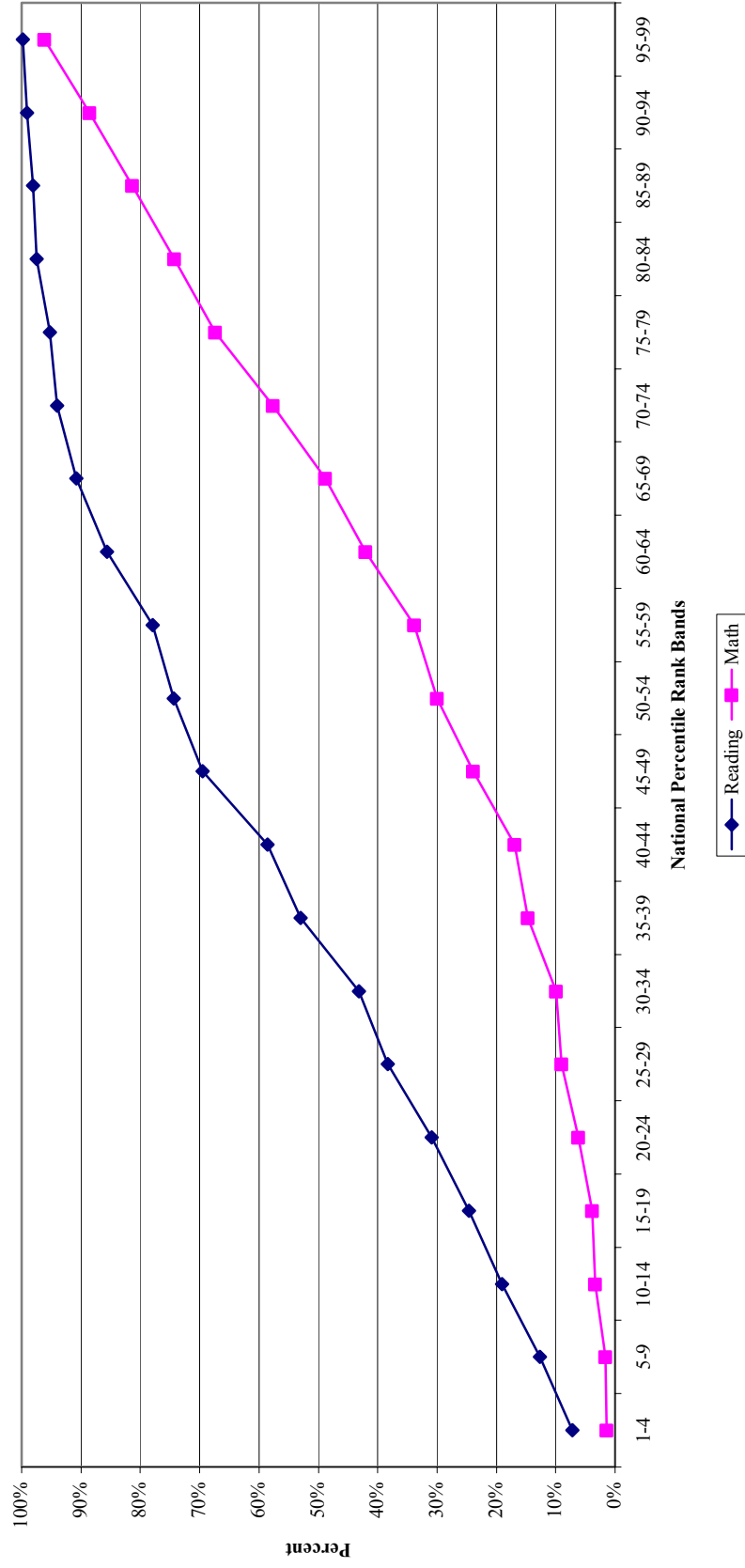
This pattern of considerable difference in the difficulty of the performance standards for reading and mathematics at grade 4 prompted a desire to investigate these relationships at grades 7 and 10. However, it was not until the 2001 WASL results were released in late summer that the two years of data necessary for the analyses were available. Norm-referenced tests were first administered in the grades preceding the 7th and 10th grade WASL grades in the spring of 2000 when 6th grade students were given another level of the ITBS and 9th graders were administered the Iowa Tests of Educational Development (ITED).

Analyses and Findings

Through the cooperation of the Assessment unit of the Office of Superintendent of Public Instruction (OSPI), matched sets of student records containing WASL scores and prior year ITBS scores were obtained. These files represented data for students who had taken the WASL as 4th graders in the spring of 2000 and 2001, and for 7th graders who had taken the WASL in 2001. Matched sets of student records containing WASL scores and 9th grade ITED scores were also made available for students who had taken the 10th grade WASL in 2001. In addition, OPSI research staff provided matched sets of student records for students with both 4th and 7th grade WASL scores (7th graders of 2000 and 2001) and students with both 7th and 10th grade WASL scores (10th graders of 2001). These sets of matched records have allowed various relationships between norm-referenced and standards-based scores for the same students to be investigated and also the relationships between different grade levels of WASL scores for the same students.

Using the approach applied to the Tacoma data, the relationships between the prior year's norm-references scores and the WASL scores for reading and mathematics were tabulated and graphed. Figures 1 through 3 show the patterns found for reading and mathematics at grade 4, 7, and 10 respectively. The number of matched cases for grade 4 was 57,348, for grade 7 it was 56,430, and at grade 10 it was 53,372. The relationship between the 2001 4th grade WASL reading and mathematics scores and the corresponding 3rd grade ITBS scores for the state data (Figure 1) was very similar to that found for Tacoma students the year before. The 7th grade data (Figure 2) was different than that found for 4th grade. The relationship between the 2001 7th grade WASL reading scores and the prior year's ITBS reading scores looked almost identical to the pattern found for the 4th grade WASL mathematics scores. (See Figure 4.) The 10th grade pattern (Figure 3) looked very similar to that for 4th grade. The relationship between the percent of students meeting the WASL mathematics standard and the prior year's norm-referenced mathematics scores was fairly similar across the grade levels (Figure 5). However, for reading (Figure 6) the patterns were similar for 4th and 10th grade but the 7th

Figure 1. Percent of Students Scoring in 2000 3rd Grade ITBS Reading and Math NPR Bands Subsequently Meeting 2001 4th Grade WASL Reading and Math Standards.



**Figure 2. Percent of Students Scoring in 6th Grade ITBS NPR Bands Subsequently Meeting 2001
7th Grade WASL Reading and Math Standard**

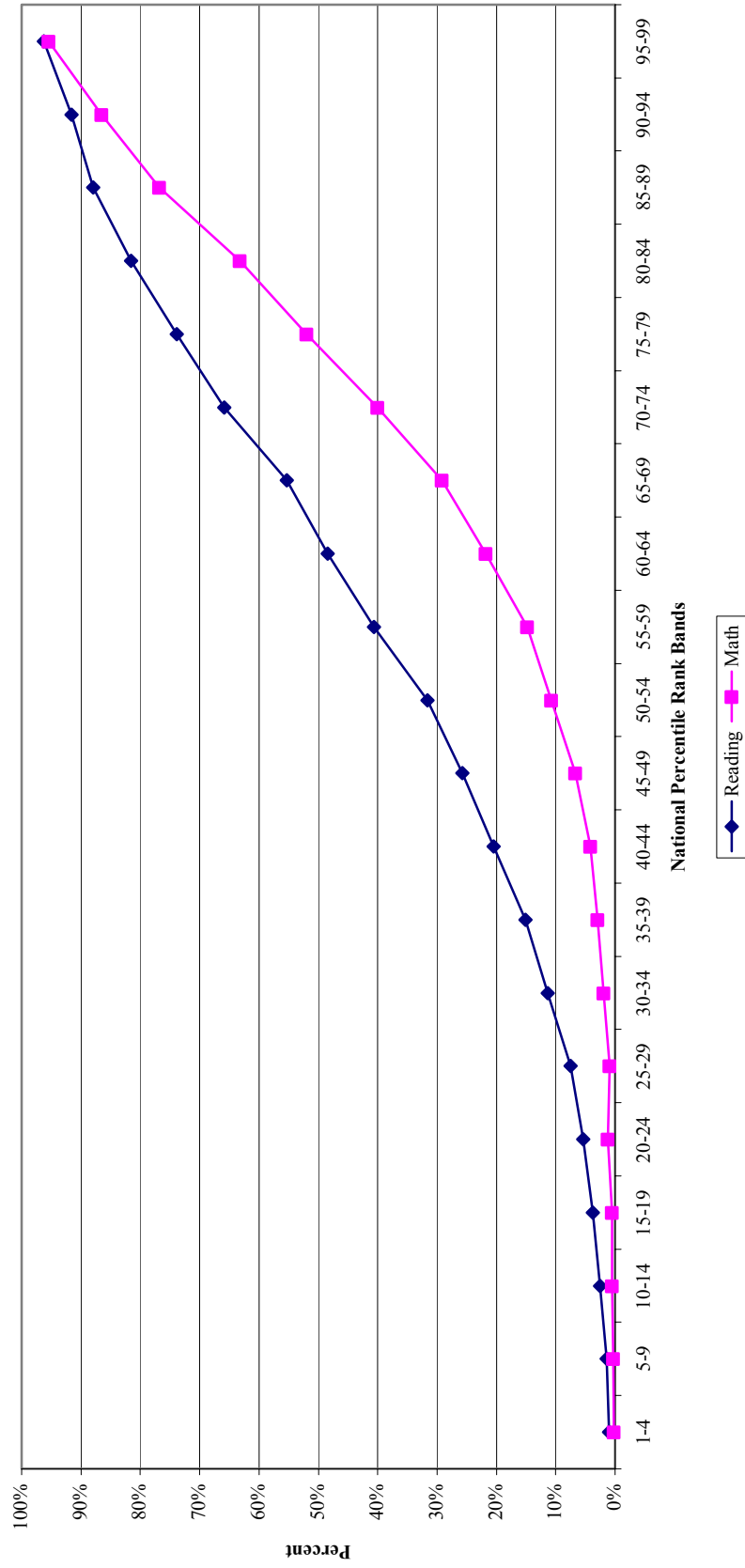


Figure 3. Percent of Students Scoring in 9th Grade ITED Reading and Math NPR Bands Subsequently Meeting 2001 10th Grade WASL Reading and Math Standard

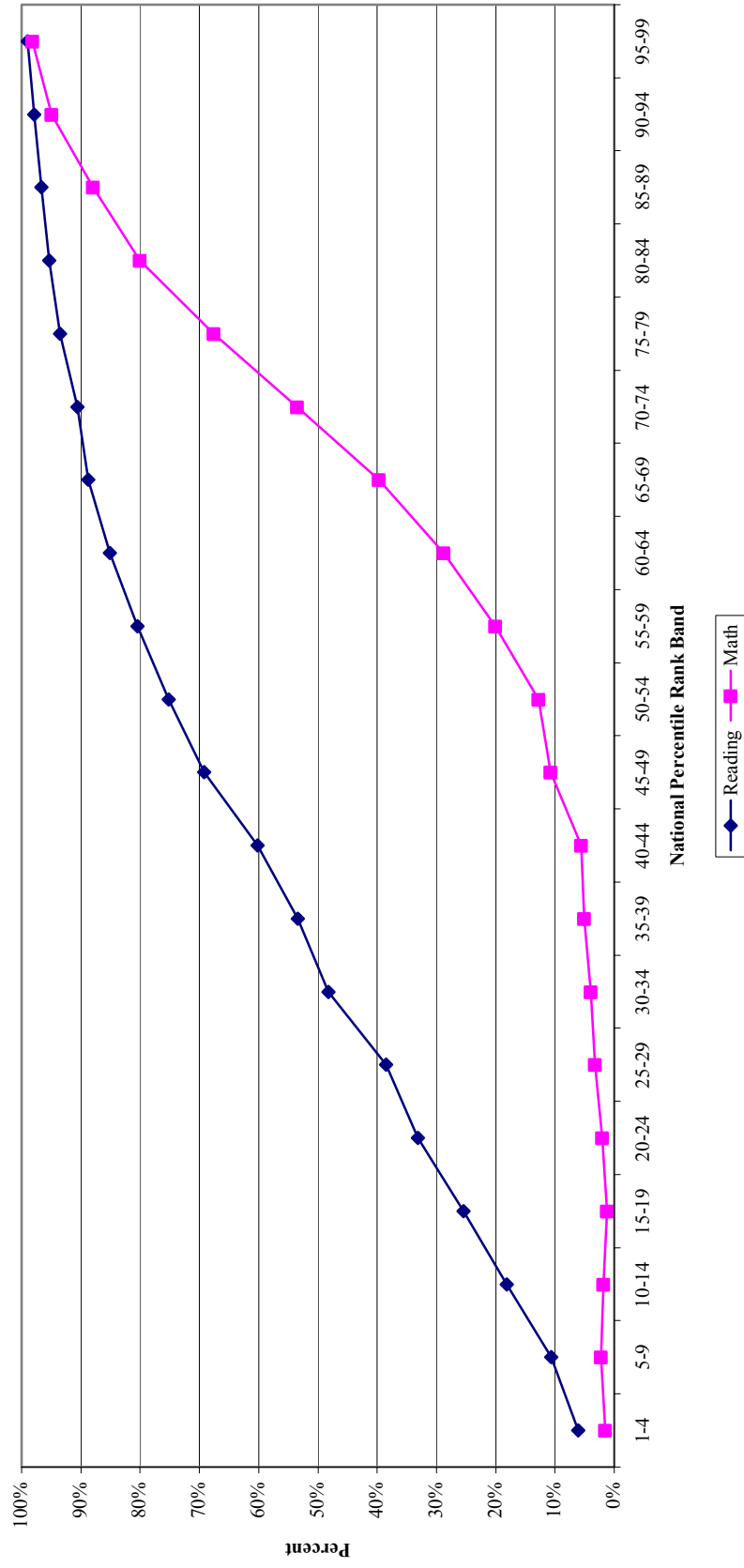


Figure 4. Percent of Students Scoring in Prior Year's ITBS Mathematics and Reading NPR Bands Subsequently Meeting 2001 4th Grade Mathematics and 7th Grade Reading WASL Standard

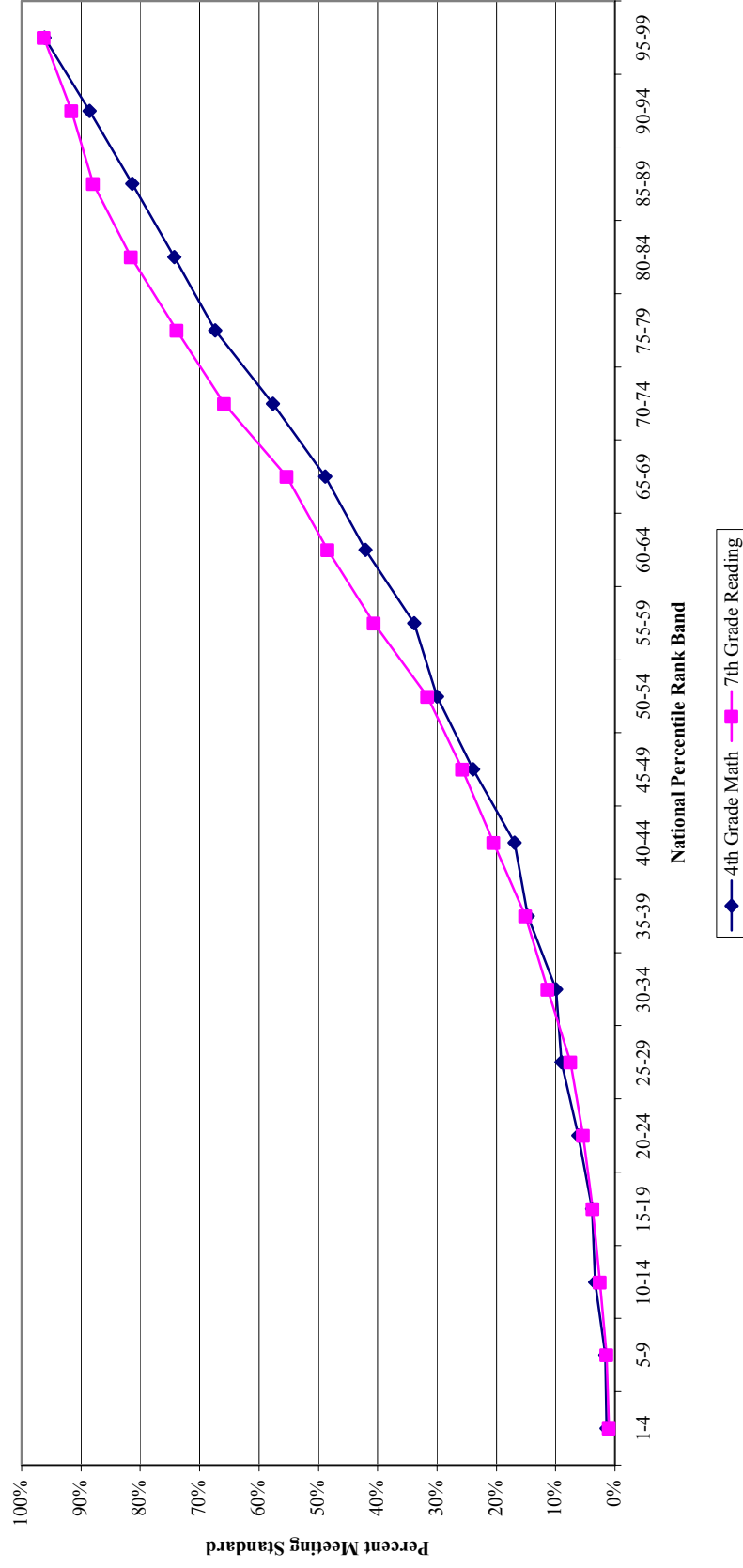


Figure 5. Percent of Students Scoring in ITBS/ITED Math NPR Bands in 2000 Subsequently Meeting 2001 WASL Math Standard

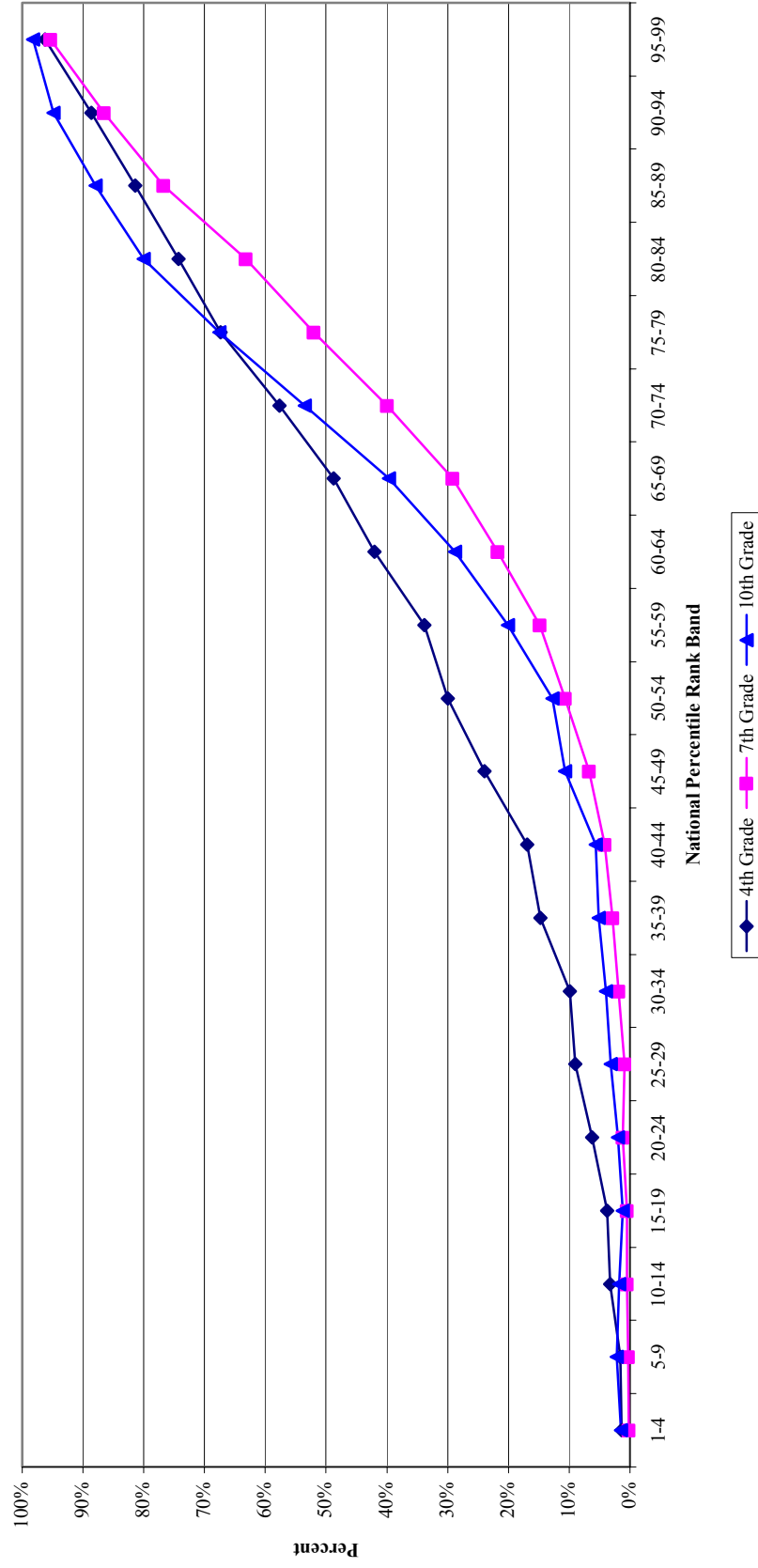
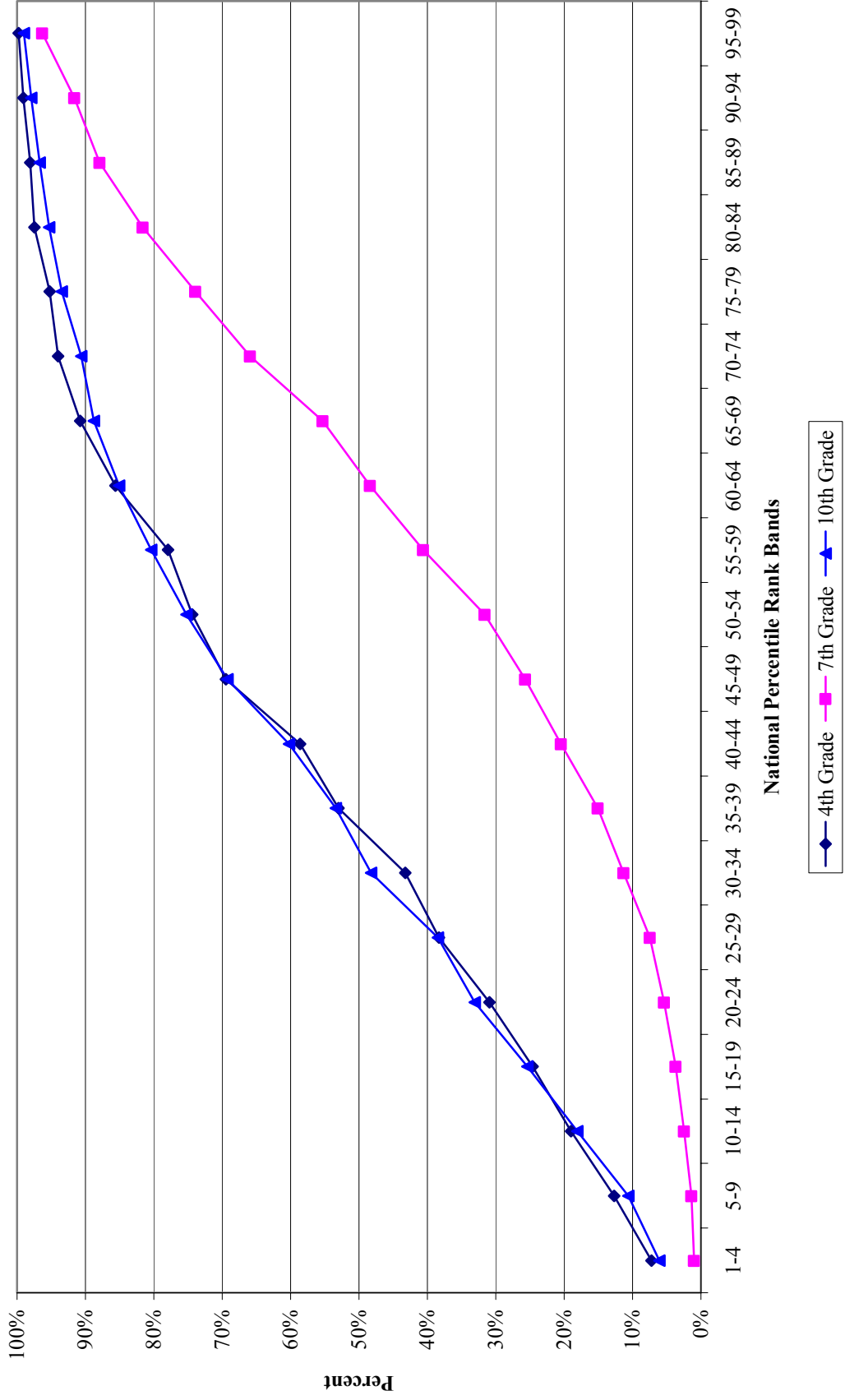


Figure 6. Percent of Students Scoring in ITBS/ITED Reading NPR Bands in 2000 Subsequently Meeting 2001 WASL Reading Standard



grade pattern was dramatically different.

The differences in the relationships between the percent of students meeting the WASL performance standard and those students' norm-referenced scores suggested looking more closely at the cut-points for the various performance levels. Recent work by Schwarz, Yen, and Schafer (2001) and Shepard and Peressini (2002) suggest that using existing normative data can help make clearer the degree of difficulty associated with the different performance levels of standards-based assessments. Shepard and her colleagues were asked by the Denver Area School Superintendents' Council to "provide a more complete understanding of the content and difficulty" of the 10th grade mathematics assessment that is a component of the Colorado State Assessment Program (CSAP). In addition to the 10th grade standards-based assessment, the CSAP also incorporates the American College Testing (ACT) Program's 10th grade PLAN assessment, a nationally-normed test administered earlier in the year. Scores from the PLAN and the CSAP are highly correlated ($r = .81$) and thus they rank individuals in about the same order. When tests order individuals in a similar manner, score points on one assessment can be linked to score points on the other. These researchers related the two sets of scores using an equipercentile linking approach and used this linking to add meaning to the various cut points of the CSAP.

Since the correlations between the WASL standard scores and the prior year's ITBS/ITED scale scores are also strongly correlated (Table 3) equipercentile linking can also be used to add meaning to the cut scores that define the various levels of WASL reading and mathematics tests.

Table 3. Correlations between reading and mathematics WASL standard scores and prior year's ITBS/ITED scale scores.

	3 rd /4 th Grade	6 th /7 th Grade	9 th /10 th Grade
Reading	.72	.76	.75
Mathematics	.77	.83	.80

At each grade level, for the reading and mathematics tests, cumulative percent distributions were constructed for both the norm-referenced scale scores (ITBS or ITED)

and for the WASL standard scores. These cumulative percent distributions actually represent “state percentile rank” distributions. The WASL standard scores (and their corresponding state percentile ranks) representing the first active score in each of the performance levels 2, 3, and 4 were linked to the ITBS/ITED scale scores that represented those same state percentile ranks. Then, using the NPR scores associated with the ITBS/ITED scale scores, a cautious interpretation of the difficulty of the WASL cut scores can be made in terms of the corresponding national normative performance. Table 4 shows the prior year’s ITBS/ITED NPR scores associated with the WASL cut score that defines “meets standard” (standard score of 400).

Table 4. Prior year’s NPR associated with WASL reading and mathematics performance standards.

Grade	Subject	NPR	Grade Level Difference*
4th	Reading	40 th	About a month below
	Mathematics	61 st	Half a year above
7th	Reading	63 rd	One year above
	Mathematics	72 nd	Two years above
10th	Reading	43 rd	Half a year below
	Mathematics	72 nd	Three years above

*The grade level for which the NPR would be expected to be about average.

This analysis confirms the patterns found in the first analysis. That is, the mathematics performance standard is difficult and very similar at 7th and 10th grades and only slightly less difficult in 4th grade. The reading standard is much easier at 4th and 10th grades, while the 7th grade reading standard is much more difficult and looks to be very similar to the 4th grade math standard in degree of difficulty. The cut scores at the standard vary in difficulty. In addition, the cut scores for levels 2 and 4 (between “well below standard” and “below standard;” and between “well above standard” and “above standard”) also are quite different across the grades and between reading and mathematics. Table 5 shows the various cut scores, their linked NPR scores based on the prior year’s norm-referenced testing, and the estimated deviation from grade level for a student for whom that NPR would be typical. These linked values also demonstrate the considerable variability in all the cut scores across grade levels for both reading and mathematics.

The normative benchmarks for level 2 vary from the 6th percentile for 4th grade reading to the 20th percentile for 7th grade mathematics. Level 4 benchmarks range from a low of the 56th percentile for 10th grade reading to a high of the 86th percentile for 7th grade mathematics, although 5 of the 6 level 4 benchmarks are at the 80th percentile or above.

Table 5. Prior year’s NPR associated with WASL reading and mathematics cut scores at different performance levels.

Grade	Subject	Level	NPR	Grade Level Difference*
4th	Reading	2	6 th	-1.6
		3	40 th	-0.1
		4	80 th	1.5
	Mathematics	2	36 th	-0.2
		3	61 st	0.6
		4	82 nd	1.7
7th	Reading	2	20 th	-1.9
		3	63 rd	1.1
		4	84 th	3.1
	Mathematics	2	58 th	0.9
		3	72 nd	2.0
		4	85 th	3.7
10th	Reading	2	15 th	-3.6
		3	43 rd	-0.7
		4	56 th	0.8
	Mathematics	2	55 th	0.9
		3	72 nd	3.2
		4	86 th	4.3

*The grade level for which the NPR would be expected to be about average score.

The last analyses used data from two different cohorts of students. Each cohort represented students with WASL scores separated by three years. The first set of scores came from 48,540 students whose files could be matched from the 4th grade WASL administration in 1998 with those from the 7th grade WASL administration in 2001. The second set came from 47,390 students whose files could be matched from the 7th grade WASL administration in 1998 with those from the 10th grade administration in 2001.

These files are particularly relevant to the WASL reading performance pattern that shows 7th grade reading performance to be consistently lower than that found at either 4th or 10th grades. The lower performance in 7th grade has prompted some districts to institute reading programs specifically aimed at correcting a perceived instructional weakness causing the big drop in performance from 4th grade to 7th grade. However, the apparent recovery of the 7th grade losses by 10th grade suggests the problem may lie more with the standards at 7th grade than with middle level reading instruction. This is not to say that reading instruction at the middle level does not need improvement, but only that the location of the standard confounds our understanding and interpretation of students' academic performance.

Of the 29,164 students in the 4th-7th grade data set who met the WASL reading standard in 1998 as 4th graders, 19,063, or 65%, met the reading standard three years later as 7th graders. Ideally, we would like to know how many of those students subsequently would meet the standard as 10th graders. But we will need to wait until after the 2004 WASL assessment to have a cohort of students with 4th, 7th and 10th grade WASL scores. However, we can use the information from the 47,390 students in the 7th-10th grade data set as a reasonable estimation. Of the 21,379 students who met the reading standard as 7th graders in 1998, 19,703 or nearly 92%, subsequently met the standard as 10th graders. This pattern does not seem to be reasonable and raises additional questions about the 7th grade reading standard.

Conclusions and Suggestions

Curriculum reform and instructional improvement is a complicated and challenging undertaking. Designing and implementing assessment systems that will be complementary and helpful rather than harmful to such an undertaking is full of its own problems. The analyses described above have been undertaken to try and make a basically sound and thoughtful assessment system better. If our assessments are to serve our curriculum improvement efforts then they must be perceived as helpful. An important component of that perception is that the assessment results be seen as reasonable.

The findings from the present investigation raise fundamental questions about the reasonableness of the performance standards associated with the WASL reading and mathematics tests. The magnitude of the differences between reading and mathematics at both 4th and 10 grade do not seem to be reasonable. It would not be unreasonable to find mathematics performance lower than reading performance at all grade levels. But the magnitudes of these differences do not seem rational. The performance standards have been benchmarked by linking them to the norm-referenced performances in related domains that students bring to the WASL testing years. Such benchmarking is not without limits, but it can help provide an understanding of just how demanding the standards are given students' current educational experiences. 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 have expected. This is particularly true at grades 7 and 10 where the linked normative performance is more like that of students 2 or 3 grades above the tested grade level. Standards-based tests are intended to "pull" curriculum and instructional practices to higher levels. But those levels must be perceived as reachable. Asking that all students reach a level that currently only the most advanced students can obtain may actually discourage thoughtful teachers and possibly even make them cynical of the whole reform movement.

The much lower reading performance at grade 7 compared to that at grades 4 and 10 would seem to be more a problem of lack of alignment in the performance standards across the grade levels than one of poor instruction in the middle grades. It just doesn't make sense that large numbers of the students who have met the performance standard as 4th graders subsequently fail to meet the standard in 7th grade. This is particularly troubling when we know that large numbers of students failing to meet that same 7th grade standard subsequently meet the 10th grade standard. It is hard for a reasonable person to believe that these standards are making sense.

So what can be done to address these issues? The first step is to recognize the magnitude of the discrepancies and understand that there is not likely to be any quick fix. The magnitude of the difficulty of the mathematics standards across the grade levels and the difficulty of the reading standard at 7th grade should be acknowledged in our public reporting. We need to help policy makers understand just how challenging are the standards we have adopted and how long and what it will take to achieve them. But we also need to look to the future for opportunities to make a conscious effort to adjust them.

Such an opportunity might naturally occur as the state assessment system is modified to bring it into compliance with the new requirements of the recently reauthorized ESEA. That federal legislation will require that standards-based assessments, or at least assessments linked to performance standards, are added to the current state assessment program so that reading and mathematics are measured at all grades from 3 to 8. This new system must be in place by the spring of 2006. In developing these new assessments performance standards will need to be set at the new grade levels (3, 5, 6, and 8). That would be an ideal time to simultaneously revisit the standards at grades 4, 7, and 10 and institute a standard setting procedure that would incorporate some sort of moderation activity that could address dramatic grade level and/or subject matter differences.

References

- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the meeting of the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, Arizona.
- Schwartz, R. D., Yen, W. M., & Schafer, W. D. (2001). The Challenge and Attainability of Goals for Adequate Yearly Progress. *Educational Measurement: Issues and Practice*, 20 (3), 26-33.
- Shepard, L. A. & Peressini, D. D. (2002). *An analysis of the content and difficulty of the CSAP 10th-grade mathematics test: a report to the Denver Area School Superintendents' Council*. Boulder, CO: University of Colorado, School of Education, Education and the Public Interest Center.