An Examination of the Longitudinal Effect of the Washington Assessment of Student Learning (WASL) on Student Achievement

Donald C. Orlich
Washington State University


Abstract

Linn, Baker and Betebenner (2002) suggested using the effect size statistic as a measure of adequate yearly progress target (AYPT) as is required by PL 107-110. This paper analyzes a four-year data set from the required high-stakes test—Washington Assessment of Student Learning—using effect size as the AYPT metric. Mean scale scores for 4th, 7th and 10th grade reading and mathematics were examined. Nominal descriptors suggested by Cohen (1988) were applied and showed no yearly effect in student achievement as a function of the WASL. Comparing the 1998 scale scores to those of 2001 showed a small effect. However, manipulating the effect size criterion from 0.20 to 0.05 did show small yearly effects in student achievement. Meeting AYPT objectives will be a problem of defining the standard as yearly score fluctuations occur. The educational research community should challenge the statistical logic associated with setting AYPT’s.

A statistical and accountability dilemma has emerged due to the passage of the “No Child Left Behind Act of 2001” (PL 107-110). States are now forced by federal law to show adequate student yearly progress targets, which will be met through
high-stakes testing. Several states have constructed their own accountability systems that feature criterion-referenced assessments. However, student test scores tend to display characteristics of norm-referenced tests, i.e., normal distributions. The use of effect size statistics, which is herein applied to one state, has been suggested as one means of determining the required adequate yearly progress targets (Linn, Baker & Betebenner, 2002).

The Washington State Model (WASL)

The State of Washington established the Washington Assessment of Student Learning (WASL) as its accountability tool. The WASL is keyed to the state’s standards called “Essential Academic Learning Requirements.” The WASL is used to test all 4th, 7th and 10th graders in mathematics, reading, and writing. The 5th, 8th and 10th graders will be assessed in science. Listening is also assessed. Using the data collected from the 1998 through 2001 WASL administrations; the writer calculated effect sizes (see Cohen, 1988) to observe trends.

The purpose of this study is to determine the effect on student achievement as a consequence of the longitudinal administration of the Washington Assessment of Student Learning (WASL) the state mandated high-stakes test. The WASL scale score means and standard deviations were available for the years 1998, 1999, 2000 and 2001 for mathematics and reading and are shown in Table 1. The average number of students taking the WASL math and reading tests during the four year period for grades 4, 7 and 10, respectively, are 70,431; 72,864 and 66,856. These three combined totals account for 21 percent of the state’s total 2001-02 K-12 student population of 1,010,424 (Education Profile, 2002). The number of WASL test-takers is significant.

Table 1
Yearly Means and Standard Deviations for 4th, 7th, and 10th Grade Mathematics and Reading Scale Scores On the Washington Assessment of Student Learning (1998-2001)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 - Mathematics</td>
<td>383.5</td>
<td>32.2</td>
<td>386.5</td>
<td>33.9</td>
</tr>
<tr>
<td>4 - Reading</td>
<td>402.1</td>
<td>19.3</td>
<td>404.2</td>
<td>19.5</td>
</tr>
<tr>
<td>7 - Math</td>
<td>357.4</td>
<td>46.4</td>
<td>364.7</td>
<td>52.0</td>
</tr>
<tr>
<td>7 - Reading</td>
<td>390.1</td>
<td>20.1</td>
<td>393.1</td>
<td>20.2</td>
</tr>
<tr>
<td>10 - Math</td>
<td>N/R</td>
<td>N/R</td>
<td>382.2</td>
<td>42.8</td>
</tr>
<tr>
<td>10 - Reading</td>
<td>N/R</td>
<td>N/R</td>
<td>402.8</td>
<td>29.5</td>
</tr>
</tbody>
</table>

Education Profile, 2002
An inspection of the scale score means shows a rather small incremental increase in most means. However, there is a scale point decline of 0.4 in the mean of grade 7, 2001 math scores compared to 2000. A similar decline is noted in 2001 for grade 4 reading, where the mean scale score dropped by 1.6 points compared to 2000.

These patterns have been praised by state policy makers as showing evidence of student progress. However, are the scores truly reflective of student achievement? To answer that question, I use the statistical test of effect size.

**Effect size**

The *effect size* is a method by which to judge the relative learning worth from independent samples (see Bloom, 1984; Cohen, 1988; Glass, 1980; Marzano et al., 2001; & Walberg, 1999). In this case, what evidence is there that administering and teaching to the WASL has a positive impact on student achievement?

Cohen (1988) defined an effect size as the difference between two means divided by the standard deviation of either group. With independent samples, such as the WASL, one can determine the effect sizes by comparing the means of two different years. Cohen also suggested that the relative efficacy of an effect could be stated in nominal terms. If an effect size (ES) were at least a 0.2, it was labeled as *small*. An ES of at least 0.5 was labeled as *medium*; while and ES of 0.8 or greater was *large*. Thus, an effect size of 0.2 is required to show efficacy of learning. Table 2 shows the effect size calculations and nominal descriptors for this study. In all calculations a uniform method was used. The earlier year is the control, while the latter year of the pair is the experimental group. Standard deviations are from the control years.

### Table 2
**Effect Size Calculations for 4th, 7th, and 10th Grade Mathematics and Reading Scores**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 - Math</td>
<td>0.09</td>
<td>None</td>
<td>0.14</td>
<td>None</td>
<td>0.06</td>
<td>None</td>
</tr>
<tr>
<td>4 - Reading</td>
<td>0.11</td>
<td>None</td>
<td>0.16</td>
<td>None</td>
<td>-0.08</td>
<td>Negative</td>
</tr>
<tr>
<td>7 - Math</td>
<td>0.16</td>
<td>None</td>
<td>0.08</td>
<td>None</td>
<td>-0.01</td>
<td>Negative</td>
</tr>
<tr>
<td>7 - Reading</td>
<td>0.15</td>
<td>None</td>
<td>0.03</td>
<td>None</td>
<td>0.05</td>
<td>None</td>
</tr>
<tr>
<td>10 - Math</td>
<td>N/R</td>
<td>N/R</td>
<td>0.13</td>
<td>None</td>
<td>0.08</td>
<td>None</td>
</tr>
<tr>
<td>10 - Reading</td>
<td>N/R</td>
<td>N/R</td>
<td>0.15</td>
<td>None</td>
<td>0.09</td>
<td>None</td>
</tr>
</tbody>
</table>

The effect is described in nominal terms as per Jacob Cohen’s (1988) definitions.

**Discussion of Data Sets**

Table 2 shows the effect sizes for the 4th, 7th and 10th grade mathematics and reading scores yearly from 1998 to 2001. Examining Table 2, you may note that at the 4th grade level, five scores show no effect in achievement, while there is one negative learning effect on grade 4 reading in 2001, that is, a decline in achievement.

The grade 7 pattern is similar showing no effect on five of the six scores and one negative effect in mathematics for 2001. The grade 10 results show no effect on mathematics and reading scores in all cases.

If one were to use an effect size of 0.05, which would account for a two percentile gain, as the Adequate Yearly Progress Target (AYPT), suggested by Linn, Baker, and Betebenner (2002), then 13 of the 16 scores would meet that target. However, using Cohen’s (1988) definitions the 16 scores would show no effect and not meet the target. Setting the criterion measure of an adequate yearly progress target (AYPT) may become a major problem of definition. This situation may further complicate the implementation of AYPT policy. It appears that further analysis of setting AYPT’s and field-studies are essential.

The average percentile gain during the four-year period of this study (1998-2001) for grades 4, 7 and 10 in math and reading was 3.3. That number would correspond to an effect size of 0.08. That effect size would exceed the AYPT suggested by Linn et al. (2002), but not the nominal ES suggested by Cohen. These differences will be explored further.

Now examine Tables 3 and 4, which display the data and effect sizes by comparing the 1998 WASL scores for grades 4 and 7 to those of 2001; and grade 10 for 1999 and 2001. Using Cohen’s classifications, five of the six comparisons show a small effect and only one with no effect. It appears to take several years to show any effect. Using WASL math results released in the SRI International report (2002) showed that for grade 4 from 1997-2002, the ES was 0.79. The ES for grade 7 from 1998-2002 was 0.31; while the ES for grade 10, 1999-2002, was 0.14. These data are inconclusive, but do suggest that like national norm-referenced tests, state
criterion-referenced tests may take several years to show a positive impact on student achievement.

Camara and Powers (1999) concluded that coaching students for the SAT does in fact increase a student’s SAT score. Washington State teachers have had at least four years experience in preparing for and teaching to the WASL. Further, between June 21 and June 30, 2001, the State Superintendent of Public Instruction selected 175 classroom teachers to attend a special, all expenses paid, WASL assessment training program in Mesa, Arizona (Brown, 2001, May 17). The impact of that training and student familiarity with the WASL appear to be similar to the SAT coaching findings reported by Camara and Powers (1999), especially at grade 4 where teachers have had over six years of “practice.” Additionally, the state superintendent of public instruction initiated a “School Improvement Specialist” program in 2001-02. About 200 selected individuals are being paid $30,000 a school year to work 1.5 days per week (or up to $90,000 for 4.5 days) to help teachers and schools “improve student performance.” No independent evaluation of this multi-million dollar expenditure has yet been conducted.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 - Mathematics</td>
<td>383.5</td>
<td>32.2</td>
<td>393.3</td>
<td>34.9</td>
<td>0.30</td>
<td>Small</td>
</tr>
<tr>
<td>4 - Reading</td>
<td>402.1</td>
<td>19.3</td>
<td>405.7</td>
<td>18.6</td>
<td>0.19</td>
<td>None</td>
</tr>
<tr>
<td>7 - Math</td>
<td>357.4</td>
<td>46.4</td>
<td>368.7</td>
<td>51.6</td>
<td>0.24</td>
<td>Small</td>
</tr>
<tr>
<td>7 - Reading</td>
<td>390.1</td>
<td>20.1</td>
<td>394.5</td>
<td>20.6</td>
<td>0.22</td>
<td>Small</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - Math</td>
<td>382.2</td>
<td>42.8</td>
<td>390.8</td>
<td>41.1</td>
<td>0.20</td>
<td>Small</td>
</tr>
</tbody>
</table>
It appears that nearly four years would be required to show a small effect on student test performance; thus AYPTs may be elusive and troublesome. And how does a state account for a negative effect size year as the Washington data illustrate? The Washington State reform movement was legislated in existence in 1993. Thus, one could argue that during an eight year period little impact on student achievement is shown for the estimated $1 Billion cost for the state’s total school reform package.

A question needs resolution: “Is the small effect shown on the four year comparisons a function of increased teacher knowledge and possible compromise of WASL questions, or is the effect a function of growth in student academic achievement?” Obviously that question is not answered in this paper, but must be explored.

**Further Implications**

This article only examines the effect size of student achievement as a consequence of longitudinal high-stakes testing in one state. To illustrate the gravity of the problem let us also examine the percent of students meeting the arbitrarily set standard scale score in math and reading for grades 4, 7 and 10 from 1998-2002. (See Table 5.)

**Table 5**

Percent of Students Meeting and Not Meeting Standard for 4th, 7th, and 10th Grade Mathematics and Reading on the Washington Assessment of Student Learning (1998—2002)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 - Math</td>
<td>37.3</td>
<td>41.8</td>
<td>43.4</td>
<td>51.8</td>
</tr>
<tr>
<td>4 - Reading</td>
<td>59.1</td>
<td>65.8</td>
<td>66.1</td>
<td>65.6</td>
</tr>
<tr>
<td>7 - Math</td>
<td>24.2</td>
<td>28.2</td>
<td>27.4</td>
<td>30.4</td>
</tr>
<tr>
<td>7 - Reading</td>
<td>40.8</td>
<td>41.5</td>
<td>39.8</td>
<td>44.5</td>
</tr>
<tr>
<td>10 - Math</td>
<td>33.0</td>
<td>35.0</td>
<td>38.9</td>
<td>37.3</td>
</tr>
<tr>
<td>10 - Reading</td>
<td>51.4</td>
<td>59.8</td>
<td>62.4</td>
<td>59.2</td>
</tr>
</tbody>
</table>


The range of percentages of students meeting the standard for math in grade 4 is
from 37.3 in 1998-99 to 51.8 percent in 2001-02. For grade 7, the range is 24.2 to 30.4 percent. And in grade 10, the range in math is from 33.0 to 37.3 percent. Considering that yet about one-half of the fourth graders and almost two-thirds of the seventh and tenth graders do not meet the standard (fail) policy makers should be alarmed. In deed, on November 30, 2002 writers of the SRI International report concluded “The analyses further suggest that the grade 7 test was more challenging for the 7th graders than the grade 10 test was for the 10th graders” (page ii). The report also noted that several WASL test items on the 4th grade math test were not aligned with the Essential Academic Learning Requirements.

The state superintendent of public instruction informed the standard-setting groups that the WASL cut scores and standard-setting are guided by the belief that, “In all content areas the standard should reflect what a well taught, hardworking student should know and be able to do near the end of grade [4, 7, or 10]” (SRI, 2002, p. 20). It is obvious that developmental, cognitive or behavioral perspectives are not being reflected in that guiding principle. Is faculty psychology now in vogue?

Orlich (2000) analyzed the 4th grade practice WASL items by using the developmental scales published by Epstein (2002, p. 184), Bloom’s Taxonomy (1954), and the NAEP scales (Campbell et al, 1998) and concluded that the bulk of items on the 4th grade WASL were well beyond their developmental level. Results of the 1998-99 WASL results confirmed that conclusion. Thus, the very high percentages of students “not meeting the WASL standard” may be traced to developmentally inappropriate items.

All children must take the WASL, with very few exceptions. The math effect size differences between grade-four minority children and white/Caucasian children range from 0.75 to 0.80, or about 28 percentile. For Special education the effect size difference between that population and white/Caucasian children is 0.96 or a 33 percentile difference (see Taylor, 2001, Tables 8-10 and 8-12). The latter is nearly one full standard deviation on the WASL. Finally, there is a correlation between reading and math WASL scores of 0.73 (Abbott and Joireman, 2001). Is math or reading being tested?

The cost-effectiveness of the WASL may be determined by the actual contract costs of the WASL with the Riverside Publishing Company. The initial contract was for about $40 Million. The renewal (2001-05) is $61,673,910 (Contract no.120-761, 2001). Thus the WASL cost is about $102 Million per se. That figure does not include teacher salaries for time spent to prepare students or to administer the WASL, plus other bureaucratic costs associated for its administration. With an average 3.3 percentile gain per year, the cost per one percentile gain is about $11 Million per year. Obviously, the cost-benefit calculation is challengeable; but no matter because the cost to meet any AYPT, as is mandated by federal laws (PL107-110), will be staggering. More importantly the money goes only to the test publishers. Not one dollar of the WASL reform expenditure goes to teachers to aid their instructional efforts.

Are the WASL Scores aberrations? Apparently not: Linn and Haug (2002) examined Colorado school buildings test scores over a four-year period and concluded the following:
“The performance of successive cohorts of students is used in a substantial number of states to estimate the improvement of schools for purposes of accountability. The estimates of improvement, however, are quite volatile. This volatility results in some schools being recognized as outstanding and other schools identified as in need of improvement simply as the result of random fluctuations. It also means that strategies of looking to schools that show large gains for clues of what other schools should do to improve student achievement will have little chance of identifying those practices that are most effective. On the other hand, schools that are identified as ‘in need of improvement’ will generally show increases in scores the year after they are identified simply because of the noise in the estimates of improvement and not because of the effectiveness of the special assistance provided to the schools or pressure that is put on them to improve.” (p. 35).

Similarly, Darling-Hammond (2003) reported that a doubt must be cast on state test gain scores because in Texas, students showed gains on the state mandated assessment, but did not make comparable gains on national standardized tests or the Texas college entrance test.

Conclusion

Using an effect size measurement and Cohen’s (1988) nominal definitions, there is no effect, that is, no positive impact on yearly student achievement as a consequence of the longitudinal administration of the Washington Assessment of Student Learning (WASL). However, over a four-year period a small effect size does emerge. The results of this study parallel the findings of Amrein and Berliner (2002a) who analyzed the consequences of 18 states with high-stakes tests. They reported that in 17 of the 18 states, student learning remained at the same level as it was before the policy of high-stakes tests was instituted.

In two separate studies, the first of 28 states with high-stakes tests, Amrein and Beliner (2002b) concluded that these tests do little to improve student achievement. In a second study of 17 states (2002c) they concluded that high-stakes tests may actually worsen academic performance and exacerbate dropout rates. The affective dimensions of high-stakes tests should be of great concern to policy makers and educators alike.

Washington State policy makers must re-examine the intent of the WASL and the empirical data sets that analyze it to determine its educational worthiness and continued fiscal support (Orlich, 2000; Abbott & Joireman, 2001; Basarab, 2001; Fouts, 2002; & Keim, 2002). At the federal level there is need to examine the practicality, reasonableness and statistical logic of setting adequate yearly progress targets. The experience in the state of Washington apparently shows that setting AYP’T’s may not only be an assessment fallacy, but a gross misapplication of adapting the banking practice of applying compound interest calculations to human cognition. Is educational reform anything you can get away with?

Note

The Author expresses appreciation to colleagues at Washington State University,
V. S. Manaranjan and Michael S. Trevisan for their critique, insights, and suggestions relating to this paper; and to the anonymous referees for their constructive feedback.

References


**About the Author**
Donald C. Orlich
Science Mathematics Engineering Education Center
Washington State University
Pullman, Washington 99164-4237
Telephone: (509) 335-4844 FAX: (509) 335-7389
E-mail: dorlich@wsu.edu


The World Wide Web address for the Education Policy Analysis Archives is epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu.

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

David C. Berliner
Arizona State University

Greg Camilli
Rutgers University

Linda Darling-Hammond
Stanford University

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Gustavo E. Fischman
California State University–Los Angeles

Richard Garlikov
Birmingham, Alabama

Thomas F. Green
Syracuse University

Aimee Howley
Ohio University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Ontario Institute of
EPAA is published by the Education Policy Studies Laboratory, Arizona State University