

**A Longitudinal Analysis of the Oregon
State Assessment Tests 1991-2001**

Gregory M. Perry¹

March 11, 2003

¹Dr. Perry is a co-founder of the Franklin School, the first school of choice in the Corvallis School District. This research project is an outgrowth of earlier work to use student assessments to evaluate student progress and school performance at Franklin.

Executive Summary

Oregon schools have used standardized assessment tools for over 50 years. Until the early 1970's, schools chose from among several major assessment instruments developed in the United States. Beginning in the early 1970's, the Oregon Department of Education (ODE) gradually began developing its own set of standardized assessments in reading, mathematics, and writing. This effort culminated in the Oregon State Assessment Tests (OSAT). Since 1991, the state has required that the OSAT be given to every student in grades 3, 5, 8, and 10/11.

The decision by the legislature to allow the ODE to conduct its own assessment system seems problematic from two perspectives. First, developing assessment instruments involves a high amount of fixed costs for it to be done well. If other tests generate essentially the same results as the OSAT, it seems to be a poor use of public resources to have the ODE “reinvent the wheel” in the assessment area. Second, the ODE has been given responsibility to implement many of the educational reforms passed by the state legislature some 10 years ago. Consequently, they have a vested interest in making sure that implementation is successful, that student learning has improved. It seems very problematic, therefore, to also give the ODE responsibility to develop the very tests used to measure progress in implementation.

An examination of the state average third and fifth scale scores for both reading and mathematics reveals a similar pattern over the last decade. Scores from 1992 to 1996 remained essentially flat, then increased every year from 1996 to 2001 by an average of one point or more per year. It seems odd that there would be so little change in student learning for the first five years after passage of the state educational reform act, followed by a relatively rapid and consistent increase in learning thereafter. There are four possible reasons why this rapid increase in scores may have occurred: (1) Student learning has gone up in the last five years, (2) the ODE has done a poor job of equating tests between years, causing the tests to vary in rigor each year; (3) teachers have been teaching to the test over the last five years, so educational gains are illusionary; or (4) the scoring scales/testing items have been intentionally “dumbed down” in the last five years and this has caused an increase in scores. It is highly unlikely that poor equating is responsible for these results because the scores do not vary randomly, but consistently increase every year.

A definitive way to rule out the possibility of “dumbing down” the assessments is to obtain copies of the exams during the period of interest (1996-2001) and analyze them to see if the rigor has changed over that period. The ODE has never released its exams to the public. A request for copies under the Freedom of Information Act was rejected, with ODE arguing that they reuse exam questions and would face increased development costs if old tests were made public. A compromise was proposed whereby the state would only release the exams to outside assessment experts for examination, and these experts would sign a legal agreement subjecting them to legal action if copies were released to the public. This proposal was also rejected.

Another way to investigate this issue is to examine test data for students who sit for the OSAT and another assessment that correlates well with the OSAT scores. This provides an indirect way to evaluate the rigor of a test. There are a number of assessment programs used in the United States, including the Terra Nova, Iowa Test of Basic Skills, and Stanford-9 exams. If the Terra Nova correlates well with the OSAT and the OSAT has not changed in rigor during the last five years, a particular Terra Nova score should reasonably predict the corresponding OSAT score regardless of the year (with some margin of error). If the rigor of the OSAT declined over time, the Terra Nova score would consistently under predict the OSAT, requiring that some adjustment be made to the predicted value to better predict the actual OSAT score.

In order for this analysis to be valid, students would have to sit for both the OSAT and the Terra Nova (or other test) at about the same time and test results would be needed over multiple years to establish any statistically significant pattern. Franklin School in Corvallis is perhaps the only school in the state that requires students in grades 2-8 to sit for the Terra Nova exam, as well as administering the OSAT to grades three, five, and eight. Franklin School has been in operation since 1995, so OSAT and Terra Nova test data were available for 1996-2001. The test data were analyzed using multiple regression analysis for both third and fifth grade students.

The results suggest a mixed bag for the reading assessments in both grades. There are some statistically significant shifts in the OSAT scores from year to year, but the results are somewhat inconsistent. By contrast, results from the mathematics regression models are both clear cut and statistically significant. The estimates suggest the third grade OSAT multiple choice mathematics test has experienced a 5 point jump in scores that cannot be explained by increases in the Terra Nova results. Assuming this result is indicative of results statewide, it suggests that the average third grade mathematics score in the state declined by about 1 point from 1997-2001, rather than increasing 4 points as suggested in the state OSAT results. At the fifth grade level, the Franklin results suggest a 15 point jump in scores that cannot be accounted for by increases in Terra Nova results. Again, if these results are indicative of performance statewide, this means that state mathematics scores actually declined by 8 points, rather than increasing by 7 points.

Given that both the OSAT and Terra Nova mathematics assessments ask essentially the same types of questions, it seems unlikely that students making broad educational progress in mathematics would do much better over time on the OSAT versus the Terra Nova, especially when the differences between the two are statistically significant and increasing over time. Again, the consistent upward trend rules out poor test calibration by the ODE. Another possibility is that teachers are teaching to the OSAT. Franklin only has one third grade and one fifth grade teacher. The third grade teacher has been at Franklin since it opened and the fifth grade teacher has been there since 1997. The third grade teacher is philosophically opposed to teaching to tests. The fifth grade teacher uses the practice OSAT mathematics exam in preparation, but does so in the belief that it will help students on the OSAT and Terra Nova. Thus teaching to the test does not explain the third grade results and is likely a secondary factor for the fifth grade.

With other possibilities ruled out or deemed unlikely, the most plausible explanation for the rapid rise from 1996 to 2001 in the OSAT mathematics (and perhaps reading) multiple choice exams in the third and fifth grade is a “dumbing down” of the exams by the ODE. This explanation could be definitively proved or disproved if the ODE were willing to release the exams to assessment experts for evaluation purposes. A more fundamental question arises out of this research, however. The integrity of the ODE’s assessment program would never be questioned if they contracted with an outside group or company to handle student assessment. Why are the citizens of Oregon paying so much money to operate an assessment system that really has no mechanism for public oversight and scrutiny?

Introduction

Every state K-12 educational system in the United States uses some kind of assessment program to evaluate student and school performance. The Oregon Department of Education (ODE) was one of the first state departments in the country to develop and administer its own assessment program, the Oregon State Assessment Test (OSAT). Testing is an enterprise that has significant fixed costs but relatively low variable costs. Fixed costs include (1) the cost of developing test questions, (2) calibrating the tests to make sure they measure academic performance correctly (test validity), (3) checking and rechecking questions to make sure they are written so as to be clear in what they are requesting of the test taker, (4) using test questions and format in a manner such that a student will earn the same score taking the same test at two different times (test reliability), (5) putting together tests each year that will generate the same score (test equating) and so forth. Variable costs include the cost of materials and scoring expenses. When many students sit for the same exam, the overall exam cost is reduced. On the other hand, if large numbers take the tests, more money can be spent to improve validity, reliability and equality. In other words, having a large population base to test is more efficient and should yield better quality results. In this, the Oregon Department of Education operates at a tremendous disadvantage. For example, the Terra Nova, Stanford-9 and Iowa Test of Basic Skills are used in several states with much larger populations than Oregon.

There is another, more serious problem with ODE’s approach to assessment. The ODE has been given responsibility in the last decade to implement a number of educational reforms adopted by the state legislature. The department frequently points to the gains made in students meeting or exceeding standards as evidence that the reforms have been successful. These gains are based on increases in scale scores on the OSAT. There is a serious potential conflict of interest problem when an agency that has responsibility for implementing a reform program is also given responsibility to develop and administer the tests used to evaluate the progress of implementation. How can one be sure that test scores are valid indicators of performance, not manipulated for political purposes?

An example is the assessment system in Texas, where the state has been conducting its own statewide assessments since 1994. In 1998, Stotsky conducted an analysis of the Texas reading

test for grades 4, 8 and 10. She found that the level of difficulty for the 1998 test was lower than that for the tests in 1995-1997. Similar results were also found for the mathematics assessment.

The major objective of this research project is to better understand the current assessment system used in Oregon and see if there are problems with the OSAT program. Specific questions to be addressed in the study include:

1. How did the ODE come to have responsibility for K-12 assessment in Oregon?
2. Has the ODE made any attempts to correlate or calibrate its assessment program with other commonly used assessment programs?
3. Is there any statistical evidence to suggest that the OSAT results are truly reflecting student progress over the last 10 years?
4. Is there any evidence that ODE has “dumbed down” their assessments over time, i.e., placed easier questions on the exam so student scores will improve.

History of Assessment in Oregon

Assessment of student academic achievement has been around as long as organized educational programs have existed. The earliest standardized, paper-and-pencil tests were administered in Massachusetts in 1845. Other experiments followed throughout the remainder of the 19th Century. By the end of World War I, many tests and scales had been developed and used among K-12 students across the United States (Ahmann, Glock, and Wardeberg).

The major objectives in assessment in those early years were largely student-centered. A 1945 bulletin by the California Test Bureau identified the following uses for “skills” tests:

1. Make a profile for each pupil so that his weak and strong areas may be identified and contrasted.
2. Make a diagnostic analysis of learning difficulties in the weak areas, so that the specific difficulties and needs of each pupil may be identified.
3. Combine the diagnostic analyses for all pupils so that a class diagnostic analysis will result. Handle the results of this combination as group or class problems.
4. Have pupils make personal lists of their difficulties as guides to their activities. Have them check off these difficulties as they are eliminated.

Oregon school districts began experimenting with a number of assessment programs in the 1940's, but no statewide coordination existed for several years. Districts began asking the State Department of Education to help them in identifying assessment programs that met the needs of their students. In 1952, the state began the process that led to the creation of the Oregon Cooperative Testing Services (OCTS). This organization helped design school or district-wide evaluation programs, aided in selecting the appropriate tests, helped in collective purchasing of test materials, scored tests and developed norms for Oregon students, and aided in interpretation of results.

Consistent with public education at that time, the focus was providing advice and leaving control to local school districts.

“Each school should have its own testing program. This program should be based upon the school’s own needs. Tests give important information if the information is used wisely to understand the children and youth who are tested and if used to provide them with a better educational opportunity. General programs of testing are administered to assemble information of the widest usefulness and tests are given on an individual basis to supplement the information available from the general programs.” (Oregon Cooperative Testing Services, p. 3)

By the 1970's, assessments were also being viewed as a valuable tool in planning and evaluating educational programs. At that time, the State Legislature felt there was a need for a comprehensive assessment program at the state level. Policymakers evaluated existing standardized tests then available in the United States and determined that none were totally appropriate in measuring the educational outcomes that Oregonians felt were important. Consequently, in 1973 the legislature allocated state and local funds to develop assessment tools in reading and mathematics. In 1974 a pilot assessment of the reading test was conducted and in 1975 the first full-scale, statewide assessment was completed (Duncan, Hall and Impara). In 1976 the first mathematics statewide assessment was conducted. It is important to note that these assessments were not conducted for all students in all grades. Rather, the tests were administered to a sample (about 8,000) of fourth grade students in the state. In 1978, a statewide assessment of writing was conducted for grades 4, 7 and 11. This test was the first conducted using writing samples, rather than by means of a multiple choice exam format. The writing assessment was repeated in 1982, again using a sample of students in grades 4, 7 and 11.

In 1985, 1987, and 1989 the State Department of Education conducted statewide assessments in Reading, Writing and Mathematics for eighth grade students in Oregon. The objectives of these assessments (as noted in the 1985 report) were to

1. Provide information to parents, students and teachers regarding strengths and weaknesses in reading, writing and mathematics;
2. Give direction to the improvement of curriculum and instruction in participating schools and the state as a whole;
3. Provide an overall indication of how well Oregon students are achieving in reading and mathematics, relative to the national norm;
4. Determine the feasibility of using locally-selected standardized tests to obtain statewide achievement data.

(See Duncan, p. 1)

The assessments were considered to be more challenging than those given in prior years. Again, the tests were limited to a sample of fewer than 6,000 students statewide.

Despite the mention of objective #4, none of the summary reports provided any indication as to how commonly used standardized tests could be used to obtain statewide achievement data. There were a couple of attempts to compare the performance of Oregon students to national or international standards. In 1985, 278 students in Newburg were given the Oregon exam and the SRA Achievement Series to permit a national comparison in reading. Based on these results, it was determined that the average reading score on the Oregon exam equated to the 62nd percentile on the SRA exam. In 1989, the mathematics assessment included 15 items from the International Mathematics Study. On these questions, Oregon students performed slightly above the United States and international medians.

In 1989, the Oregon legislature approved House Bill 2132, designed to “...define by rule a basic education program to be available to all elementary and secondary students in public schools in the state...” (Erickson, p. 1) The ODE was given responsibility to outline common curriculum goals for Oregon students. Although the legislature did not mandate statewide assessment of students in HB 2132, the State Board of Education determined that statewide assessment was a logical part of teaching the curriculum goals. The board mandated that, beginning in 1991, the ODE was to assess students in grades 3, 5, 8 and 11 in basic skills, i.e., reading, writing, mathematics, listening, reasoning and study skills. They also directed that students be tested in specific subject areas, including literature, health, science, physical education, social studies, art, music and personal finance.

In 1991, the Oregon legislature passed the Education Act for the 21st Century (House Bill 3565), considered at the time to be one of the most sweeping educational reform laws in the country. Among other things, HB 3565 mandated a more demanding educational curriculum, a more extensive assessment system, and higher performance standards.

The first statewide assessment was conducted in 1991 and covered reading, literature, mathematics, listening, study skills and writing of students in grades 3, 5, 8 and 11. All but the writing assessment were scored on a scale developed for the Oregon assessments in 1980 by Holmes. No attempt was reported by ODE to show how the Oregon results compared to results of other commonly used assessment instruments.

A less ambitious testing schedule was implemented in 1992, in part because of budgetary limitations arising from the passage of Measure 5. Students in grades 3, 5, 8 and 11 were assessed in reading, mathematics and health, with the writing assessment conducted for only grades 3 and 8. In 1993, the reading and mathematics assessments were given to grades 3, 5, 8 and 11, with the writing assessment given to grades 5 and 11 students. In 1994, all students in grades 3, 5, 8 and 11 were assessed in reading, mathematics, and physical education, with the writing assessment given to students in grades 3 and 8. In 1995, students in grades 3, 5, 8 and 11 were assessed in reading, mathematics, and science.

In 1995 the legislature refined HB 3565 to focus on academic standards in specific content areas and expanded the assessment system from strictly multiple choice tests to include performance tests for writing and problem solving and classroom work samples. The high school assessment was moved from the 11th to the 10th grade. In addition, ODE began segmenting student scores into three categories: Those exceeding the performance standard, those meeting the performance standard, and those who fail to meet the standard. This segmentation of results is similar to the interpretive methods used in the National Assessment of Educational Progress (NAEP). The 1996 standards for the reading and mathematics multiple choice tests were²:

	Meets Standard	Exceeds Standard
Grade 3	202	215
Grade 5	215	231
Grade 8	231	239
Grade 10	239	249

In addition, students taking the open-ended mathematics assessment and the writing assessment had to score 4 out of 6 on all four categories. The rationale for selecting these scores as standards is not clear from ODE documents.

Since 1996, the assessment process has been done more or less the same each year. Reading and mathematics multiple choice exams are given to grades 3, 5, 8 and 10 each year. The writing and open-ended mathematics assessments are given to grades 5, 8 and 10 each year. A revised science assessment was given to grades 5, 8 and 10 in 2000 and 2001. The ODE is also working on a social studies test to be given in the near future.

A few changes have been made in the standards reported above. The scale score to meet the third grade standard was dropped from 202 to 201. The open-ended mathematics and writing assessments now need to average 4.0 to meet the standard, rather than meeting the 4.0 standard in all categories.

Over time the ODE has continued to de-emphasize scale scores and place more focus on percent of students meeting the scale standard. Currently, one can download the percentage of students meeting and exceeding the scale score standards for all schools in the state for the last four years. Scale scores are available by request from ODE, but are not available on their web site. In particular, the 1998 and 1999 scale scores were, at one time, provided on the web site but have since been removed.

²It's noteworthy that the standards used for the eighth grade assessment in 1989 was 220 for the first standard and 241 for the second standard.

Statistical Summary of OSAT Data

The ODE collects and summarizes test data by school and district each year. As noted before, the department has de-emphasized the reporting of scale scores in favor of reporting the percentage of students meeting and exceeding the state standards. Consequently, it was necessary to request the scale scores from the ODE³.

The data were provided in a variety of formats and configurations. In some years scale scores were not reported for some schools, apparently because of the small number of students at the school. In some years, scores for home school children in some districts were reported, as well as scores for middle school students taking courses at the local high school, and many more special case situations. For this reason, the number of schools used to calculate average scale scores differs from year to year. One change made to the data sets to make them more comparable from year to year was to drop scores from all schools with 5 or fewer students sitting for the OSAT. Note also that these scale scores are averages of school averages, rather than averages of all test scores recorded in the state. Consequently, the scores may be a bit different than the actual averages using student data, although any difference is likely to be more or less the same each year.

Table 1 provides a summary of mathematics and reading scale scores for third grade students from 1991 to 2001. In addition, the table provides a summary of characteristics used by ODE to calculate socioeconomic status (free and reduced lunch, average daily attendance, and percent student mobility). The percent of students on reduced lunch increased throughout the 1991-2001 time period, from 27% in 1991 to 41% in 2001. Average daily attendance and student mobility remained largely constant.

Both reading and mathematics average scores exhibited a similar pattern over the 1991-2001 period. Scores jumped markedly from 1991 to 1992, then remained essentially flat from 1992 to 1996. In 1997, the data exhibited another relatively large increase, then increased every year thereafter by an average of one point per year.

Table 2 contains a summary of OSAT scores for the fifth grade, again covering the 1991-2001 period. Fifth grade reading scores stayed relatively constant from 1991 to 1996, fluctuating from year to year and, on average, increasing by about 0.6 points per year. Mathematics scores exhibited an even flatter pattern, increasing from 1991 to 1996 by only 0.2 points per year. As was the case with the third grade exams, the scores from 1996 to 1997 jumped by 1.5 points for

³ The process to obtain these scores required several months and much persistence. In July a request for some of the test scores was ignored. Several telephone messages left in August were not returned. In October a letter was sent to the Attorney General's office requesting the data under the Freedom of Information Act. Contact was made and the data were promised to be delivered. It required a second contact with the Attorney General's office before the data were sent.

reading and by almost three points for mathematics. Thereafter, reading and mathematics increased every year by an average of one point per year.

Results for the eighth grade are provided in Table 3. The patterns for the eighth grade are different from those exhibited for the third and fifth grade. There is a gradual increase in scores over time, but substantial fluctuation from year to year. The increase occurs at about the same rate in the first five years as the second five years, with several declines in scale scores occurring throughout the 10 year period.

In summary, the third and fifth grade results are similar in that average scale scores remain relatively flat until about 1996, then increase every year thereafter at a healthy rate. The eighth grade results are much more consistent throughout the period, with periodic declines occurring throughout the 1991-2001 period.

Explaining Increases in the OSAT

The reason for these marked increases in scale scores is of great importance, because they directly impact the percent of children meeting and exceeding academic standards at each grade level, as well as on the results for the Oregon School and District Report Cards. They are also used to support ODE's contention that Oregon students are improving in their academic performance. For example, consider the following statement by State Schools Superintendent Stan Bunn after publication of the 2001 fifth grade test scores:

“We continue to see steady progress in student achievement. Oregonians should take pride that higher percentages of their public school students are reaching higher academic standards.”

There is something unsettling about the results summarized in Tables 1-3, however. If the Oregon educational reforms were put into place in the early 1990's, why did they seem to have so little impact on test scores up to 1997? And why have such rapid gains taken place since then? Consistent with technology adoption patterns, one would expect a period of gradual adoption for the new educational reforms, with some teachers embracing the changes immediately but many others being more cautious in changing what and how they teach. In addition, there likely is some trial and error period, where some teachers identify methods that work and gradually pass the good ideas on to others. Changes in the knowledge base of students are also likely to be gradual, given the gradual adoption of the reforms by teachers. All of this leads one to expect that OSAT scores should increase gradually (assuming the reforms are in fact beneficial). The results do not exhibit this kind of pattern at all. This point can be better understood by examining scores for other assessment exams. Figure 1, for example, shows average SAT scores for Oregon and the United States over the 1991-2001 period. With the Oregon and U.S. SAT scores, there is an upward trend over time, but the scores show substantial fluctuation from year to year. Figure 2 illustrates the OSAT state scores for mathematics in third and grades for the same 1991-2001

period. There is a marked jump from 1991 to 1992 for third grade, followed by essentially no change for 3-4 years, after which the scores increase every year for all tests.

Table 1. Summary Statistics for Oregon State Assessment Tests 1991-2001 - Third Grade

	1991	1992	1993	1994	1995	1996	1997	1998	1999
Number Schools	723	720	762	714	722	727	717	714	721
Avg. Students per School	NA	51.5	49	51.8	NA	50	51.1	52.8	53.4
Reading Score									
Avg	201.34	204.1	202.84	203.94	203.5	205.69	208.87	209.41	210.4
Std Dev	3.61	4.56	4.3	4.15	4.02	4.44	4.78	4.54	4.36
Math Score									
Avg	197.23	201	201.55	201.44	201.43	201.46	204.59	205.75	206.5
Std Dev	3.55	8.29	3.97	3.77	3.78	3.93	4.32	4.32	4.43
Percent Reduced Lunch									
Avg	27.23	30.71	32.3	34.58	37.37	37.6	NA	39.53	40.66
Std Dev	15.9	17.62	19.41	18.45	19.66	20.77	NA	21.53	21.77
Percent Attendance									
Avg	94.46	94.6	94.89	94.33	94.14	94.17	NA	93.93	93.96
Std Dev	1.26	1.2	1.22	1.22	1.32	1.41	NA	1.46	1.6
Percent Mobility									
Avg	16.54	15.51	15.48	14.83	15.59	17.17	NA	17.19	16.93
Std Dev	8.17	7.82	8.99	8.63	8.02	10.63	NA	15.64	11.94

Table 2. Summary Statistics for Oregon State Assessment Tests 1991-2001 - Fifth Grade

	1991	1992	1993	1994	1995	1996	1997	1998	1999
Number Schools	697	700	751	704	708	700	705	712	728
Avg. Students per School	NA	52.5	51.91	55.1	NA	55	55.1	53	53.5
Reading Score									
Avg	213.47	214.70	214.22	215.93	215.63	216.38	217.75	218.48	219.7
Std Dev	8.75	3.80	3.71	3.62	3.70	3.99	4.17	4.12	4.30
Math Score									
Avg	213.47	214.27	214.25	214.01	214.00	214.35	217.11	217.71	218.4
Std Dev	9.00	3.68	4.10	3.89	4.19	4.17	4.24	3.98	9.17
Percent Reduced Lunch									
Avg	27.77	31.13	32.37	34.58	37.60	37.53	NA	39.13	40.49
Std Dev	16.01	17.54	19.48	18.52	19.60	20.48	NA	21.68	21.61
Percent Attendance									
Avg	94.47	94.60	94.84	94.31	94.13	94.17	NA	93.93	93.97
Std Dev	1.25	1.21	1.27	1.24	1.34	1.45	NA	1.49	1.46
Percent Mobility									
Avg	16.35	15.58	15.61	14.83	15.63	17.41	NA	17.21	16.52
Std Dev	7.95	7.97	9.83	8.66	8.06	11.83	NA	16.19	8.21

Table 3. Summary Statistics for Oregon State Assessment Tests 1991-2001 - Eighth Grade

	1991	1992	1993	1994	1995	1996	1997	1998	1999
Number Schools	315	310	339	310	308	313	317	308	338
Avg. Students per School	NA	110.4	108.1	114.8	NA	121.5	122.5	123.5	117.6
Reading Score									
Avg	226.81	226.42	225.25	227.95	227.86	229.32	230.47	229.87	229.5
Std Dev	3.24	3.18	3.56	3.39	3.71	3.71	4.33	3.68	4.63
Math Score									
Avg	227.84	230.01	229.42	229.78	230.06	230.09	230.1	230.66	230.4
Std Dev	3.48	3.32	3.48	3.6	3.93	3.7	4.47	3.92	4.5
Percent Reduced Lunch									
Avg	24.54	28.11	28.59	30.95	33.55	33.76	NA	34.84	36.59
Std Dev	13.68	14.96	16.89	16.21	16.13	18.18	NA	19.72	18.25
Percent Attendance									
Avg	93.15	93.35	93.49	92.94	92.78	92.69	NA	92.19	92.45
Std Dev	1.76	1.73	1.75	1.82	1.91	2.53	NA	3.7	2.62
Percent Mobility									
Avg	15.65	15.27	15.08	14.12	15.03	17.9	NA	22.46	21.77
Std Dev	9.71	7.97	8.18	6.72	6.01	15.27	NA	42.13	31.09
Parent Education									
Avg	3.18	3.17	3.14	3.22	3.2	3.15	NA	3.08	3.1
Std Dev	0.44	0.43	0.46	0.41	0.44	0.47	NA	0.47	0.48

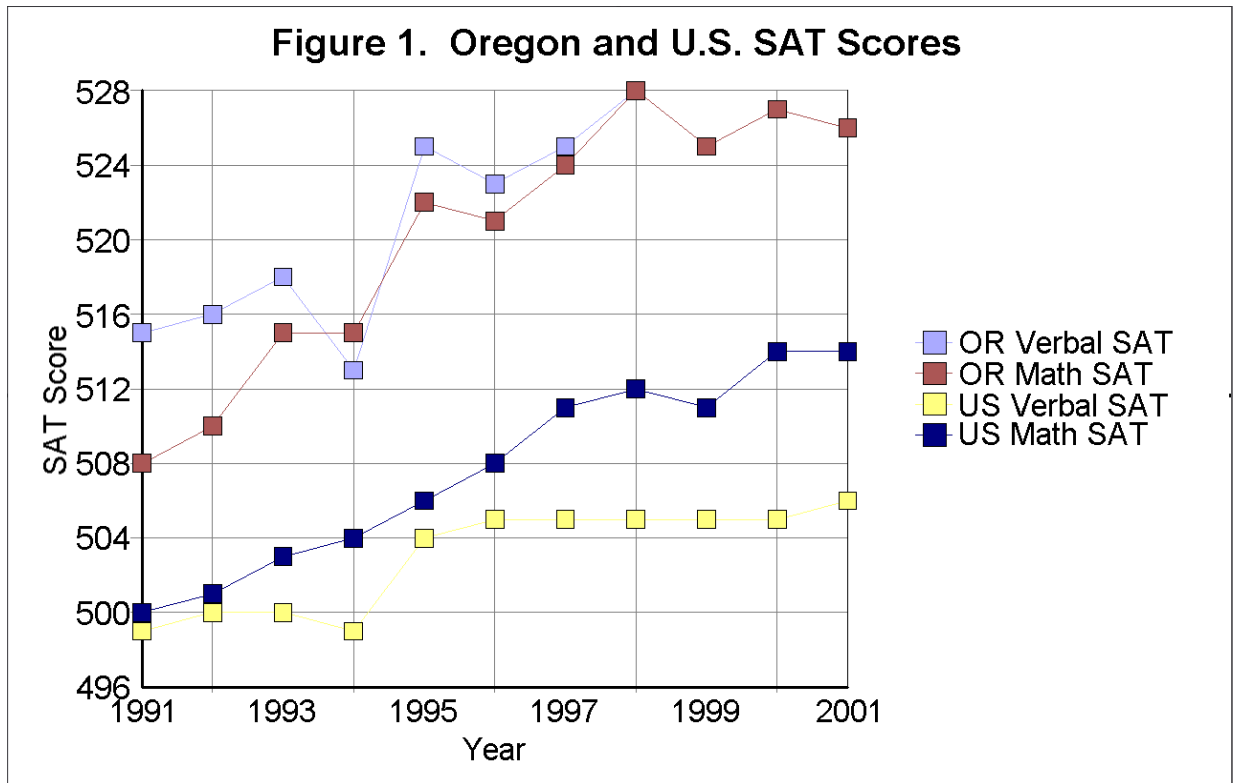
One could argue that comparing the OSAT to the SAT is unfair because the SAT has been around for many years, whereas the OSAT has been in place only a decade. Nevertheless, the patterns illustrated in the tables and Figure 2 do not seem consistent with adoption of new educational curriculum and methods at the state level.

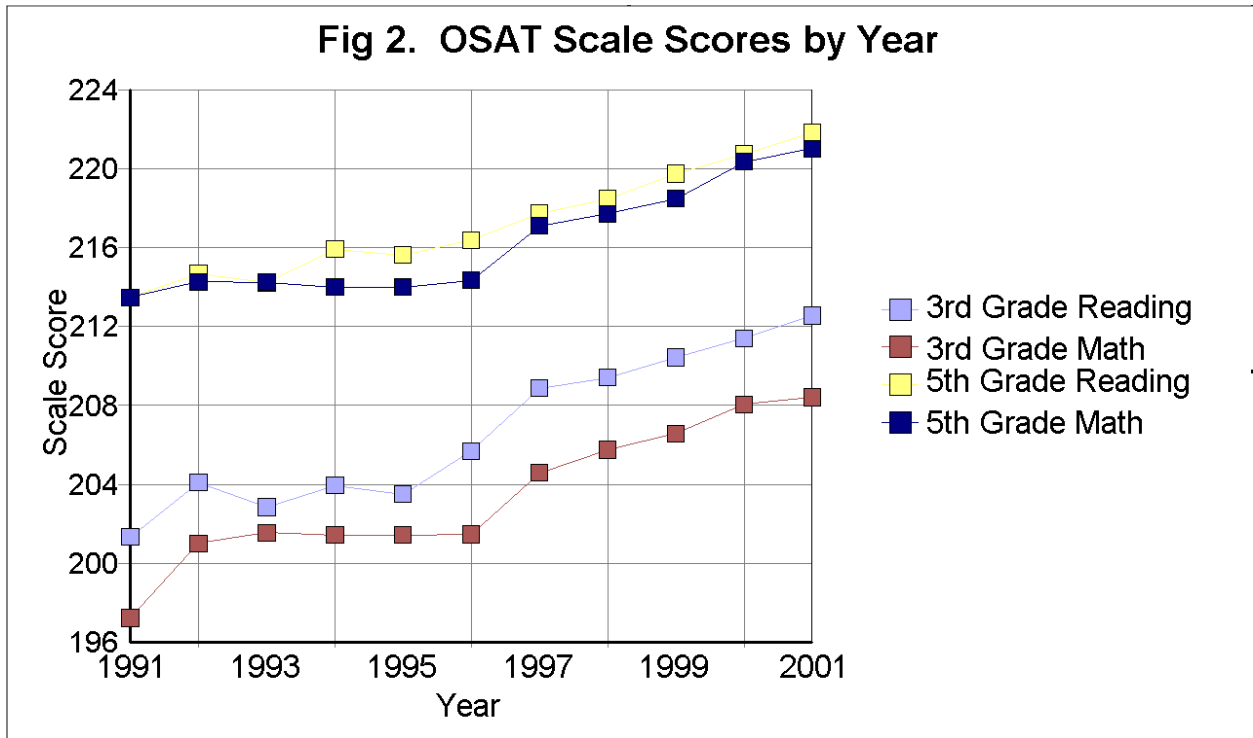
There are, in fact, 3 alternative explanations for this rapid rise in scores. They are (1) the ODE has done a poor job of equating tests between years, causing the tests to vary in rigor each year; (2) teachers have been teaching to the test over the last five years; or (3) the scoring scales/testing items have been intentionally “dumbed down” in the last five years and this has caused an increase in scores. None of these explanations place the ODE’s Assessment Division in a favorable light, so each needs to be given careful consideration and analysis.

The issue of problems with test equating seems unlikely. The ODE has staff who are trained to develop assessments that are well equated. More telling, a poor job of equating should introduce randomness into the average scores, causing them to be high one year and low the next. An examination of the scores shows steady increases each year at the third and fifth grade level. Consequently, test equating was ruled out as a possible explanation.

The explanation of teachers teaching to the tests has come up in a number of states. Poorly designed, trivial questions over a relatively narrow subject set can weaken the meaningfulness of

tests. In these situations teachers can focus classwork on drills designed to prepare students to do well on assessments. In Texas, for example, some believe that many of the gains in the Texas Assessment of Academic Skills can be attributed to teaching to the test (Patterson). In larger states a small industry has developed to provide materials to teachers that will help them improve the assessment scores of their students.





The hypothesis that the OSAT scores increased because teachers teach to the tests seems consistent with the patterns exhibited in Tables 1 and 2. When this occurs in other states, however, the scores usually increase in the first few years of the test before leveling off (Dietel). In Oregon, the scores were level, then began increasing rapidly. Perhaps the creation of formal standards in 1995 motivated these sudden surges in scores. Again, however, the question is why three years, so what incentive was introduced at this point to increase test scores? Teaching to the test doesn't seem to fit these results, but one cannot entirely rule it out either.

Alternative (3) is perhaps the most serious of these explanations, because it suggests the tests are being manipulated to make it appear that students are learning at a higher level when in fact they are not. Much as one might like to believe otherwise, this explanation cannot be automatically dismissed. As the citation by Superintendent Bunn suggests, the ODE uses the OSAT scores as evidence that their efforts are having a positive impact on the learning of students in Oregon. Because they develop their own tests and make no attempt to equate the tests to any other assessment exam used in the United States, the ODE certainly has the power to “dumb down” the exams without it being easily detected.

There are a couple of effective methods to rule out the issue of “dumbing down”. One approach is that taken by Stotsky. It involves an expert analysis of the exams over time to see if the language and content have changed. A second approach is more straightforward and less

dependent on expert opinion. It requires that the ODE set up a special experiment involving 500-1000 students in grades 3, 5, 7 and 10. In this experiment students are given the 1996 OSAT on one day, then given the 2001 OSAT within the next couple of days. Both tests should be graded using the exact same system utilized in 1996 and 2001, respectively. The population of students involved in these experiments should be representative of all Oregon students. The grading process should be overseen by outside individuals with no ties to the ODE, to ensure that the grading process is not manipulated in any way. If the scores on the 1996 and 2001 tests are the same (within acceptable statistical limits), then it can be definitively stated that ODE has not “dumbed down” the OSAT. Of course, this kind of experiment will cost the ODE time and resources to carry out. Based on the evidence presented thus far, it is questionable whether this experiment is worth carrying out.

Statistical Summary of Franklin School Data

Fortunately, another data set from Franklin School in Corvallis can provide some additional insight into the rapid increases exhibited in the third and fifth grade OSAT scores. Franklin School opened in the Fall of 1995 as the first School of Choice in Corvallis. In its first year, Franklin had 96 students in grades 3-6. Over the next two years the school expanded to serve grades K-8, with one classroom per grade. Franklin School was created by the Corvallis School Board to counteract the loss of Corvallis parents from the public school system. The parents and teachers who created the school formulated a number of guiding principles around which the school was formulated. The first was the importance of identifying a spiral, or coordinated, curriculum to be adopted by all teachers in the school. The Core Knowledge curriculum was adopted by the teachers in the Spring of 1995 and has been in use since that time.

A second principle was the use of standardized tests to monitor student progress each year at Franklin. The teachers and administration, working with Corvallis School District 509J, determined that the Comprehensive Test of Basic Skills (CTBS) was the best tool to evaluate the breadth of subject material taught at Franklin School. Since the first year, the CTBS (now called the Terra Nova) has been administered in the spring of each year to every child enrolled at Franklin. Students in the third, fifth and eighth grade have also sat for the OSAT each year. Consequently, Franklin is one of the few schools in the state that have test scores for another standardized test taken essentially at the same time as the OSAT.

The OSAT and Terra Nova are different tests developed from a different testing theory framework, so understanding the differences is important before proceeding further. The OSAT is a criterion-referenced test, although Bracey terms this kind of exam “content-referenced”. The criteria being tested are the educational benchmarks outlined by the ODE for each age level. The criteria might be thought of as a continuum from essentially no knowledge to very advanced understanding in a particular area. Each level of knowledge has associated with it a scale score. States (like Oregon) use certain scale scores to indicate a student meets or exceeds a “criterion”. The scale scores can be set up to be comparable across age levels. In Oregon, for example, a

third grade student and a fifth grade student who both score 225 on the OSATs for their respective grades exhibit roughly the same level of knowledge relative to the state benchmarks (Wolfe).

The Terra Nova, by contrast, is a norm-referenced test. A norm referenced test is developed by identifying a set of test questions that reflect the content of knowledge students at each level are commonly expected to know. Most questions can be answered by 30% to 70% of the students, with a few questions added that are harder and easier to help test administrators determine where each student is in his/her subject knowledge. This test is given to a sample population and the results scored to determine the knowledge level of the average student. This average is then assigned a scale score and all other scores are assigned in reference to this norm, hence the term “norm-referenced test.” The Terra Nova, Iowa Test of Basic Skills, Stanford-9, and other similar exams are norm-referenced.

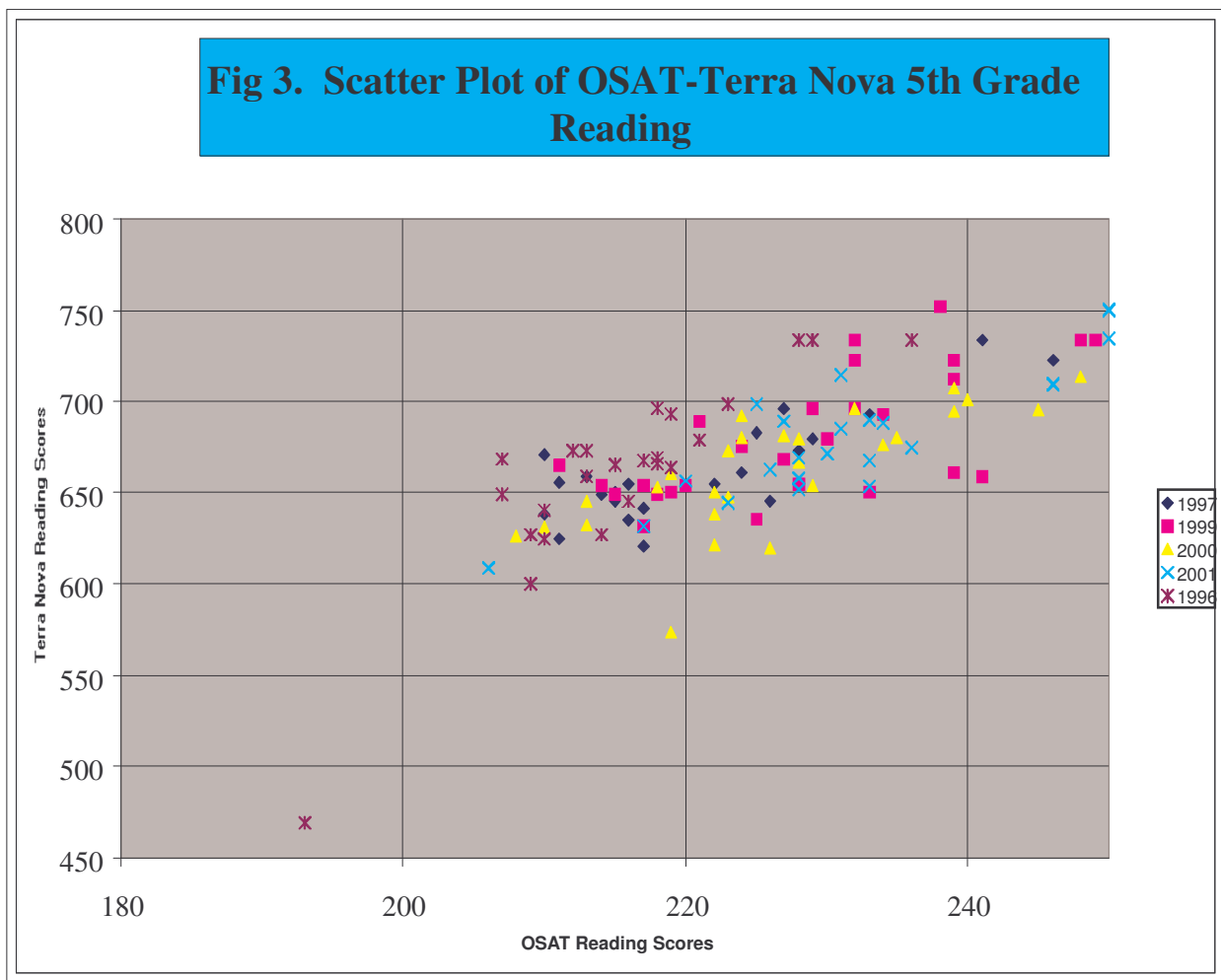
Most test experts are cautious when attempting to equate scores from one test to another, because each evaluates a somewhat different set of skills and knowledge. They are even more cautious when comparing norm-referenced and criterion-referenced tests, because the scale scores in each are derived in fundamentally different ways.

Although from a theoretical perspective the Terra Nova is quite different from the OSAT in its creation and what it attempts to measure, in a practical sense the tests are quite similar. The ODE provides teachers and the public with guidance on the major subject matter areas covered in the OSAT for reading and mathematics. For example, the ODE identifies 5 subject matter areas covered by the Benchmark 2 (fifth grade) mathematics multiple choice exam: (1) Calculations and Estimations, (2) Measurement, (3) Statistics and Probability, (4) Algebraic Relationships, and (5) Geometry (Oregon Department of Education, 2000). An examination of the Terra Nova Level 15 (fifth grade) test reveals that all five of these topics are also on their assessment exam.

Comparing a 1997 fifth grade sample OSAT provides additional insights. The OSAT is less than half as long as the Terra Nova, but has many of the same kinds of questions. It also seems to have much of the same emphasis on calculations and measurement, with less time spent on statistics, algebra and geometry. Consequently, one expects that the results for these two tests should track each other moderately to very well. The major concern would be that the shorter OSAT probably introduces more variability in the final scores because each question weighs heavier in the overall assessment.

A graphical representation of the Terra Nova and OSAT scores for each student illustrates the strong relationship between these two exams (Figure 3). Despite the inherent randomness that will occur because students are sitting for these exams on different days and the fact that the tests pose a different set of questions, there is still a clear relationship between how the students do on

the OSAT and the composite mathematics score from the Terra Nova⁴. Note that scores for 1998 are not provided in the graph. In that year the school leadership decided to not test third, fifth and eighth grades on the Terra Nova because they were already sitting for the OSAT. Third grade OSAT scores in 1996 were lost and so were also not considered in the analysis. The relationship in Figure 3 appears to be linear across the range of observations. So, while its not acceptable to completely equate the results of the two tests, it seems clear that the Terra Nova is going to provide a reasonable estimate of the corresponding OSAT for most students. Scatter plots for the other third and fifth grade reading and mathematics tests are provided in the appendix.



⁴It should be noted that Franklin students actually sat for the old version of the Terra Nova (CTBS 4) from 1996-1999. The CTBS 4 scores were converted to the Terra Nova equivalents using an equating program provided by CTB McGraw-Hill.

A simple average of the OSAT and Terra Nova scores by year for the Franklin third and fifth grade students provides a beginning point to explore the idea of score irregularities. These data are provided in Tables 4 and 5.

Table 4. Average Terra Nova and OSAT Scores by Year for Third Grade Students at Franklin School, Corvallis

Year	Reading		Mathematics		Sample Size
	Terra Nova	OSAT	Terra Nova	OSAT	
1997	635.2	210.2	603.5	205.6	26
1999	661.1	217.8	637.5	214.8	26
2000	643.6	216.3	609.1	210.1	25
2001	662.8	218.3	631.1	216.2	27

Table 5. Average Terra Nova and OSAT Scores by Year for Fifth Grade Students at Franklin School, Corvallis

Year	Reading		Mathematics		Sample Size
	Terra Nova	OSAT	Terra Nova	OSAT	
1996	657.8	221.0	660.6	215.6	24
1997	686.5	225.2	663.1	221.2	22
1999	678.9	226.9	680.7	228.6	27
2000	681.0	227.2	662.8	226.3	27
2001	703.4	234.7	685.7	235.6	27

At the third grade levels, scores for both reading and mathematics are tending to trend upward for both Terra Nova and the OSAT. In 1999 and 2001, the Terra Nova reading scores are very close, as are the OSAT scores. This and other comparisons suggest nothing unusual is occurring with the third grade reading OSAT. The average score at Franklin has increased from just above the state average (209) to well above the state average by 2001 (213). Possible reasons for this rapid increase will be discussed later in the report.

For third grade mathematics, the patterns seem to suggest something unusual has occurred. In particular, the 1999 Terra Nova score was actually higher than the 2001 Terra Nova score, yet the 2001 OSAT was higher than the 1999 OSAT.

There was also a positive upward trend for the Terra Nova and OSAT fifth grade exams. Both subject areas exhibited irregularities in the test scores. From 1997 to 1999 the Terra Nova reading average actually dropped by 8 points, but the OSAT increased by 2 points during the same period. From 1996 to 2000 the average Terra Nova mathematics score was essentially the

same, yet the OSAT average went up almost 11 points. From 1999 to 2001 the Terra Nova increased by 5 points, but the OSAT increased by 7 points. Keep in mind that the OSAT scale scores operate on a much smaller, narrower range than does the Terra Nova scale scores. In short, something seems very wrong with the changes in the OSAT scale scores. This phenomenon can be seen in Figure 3. Each year's scores tends to cluster to the right of the previous year's scores, suggesting a gradual shift toward higher OSAT scores for the same level of performance on Terra Nova.

Regression Models

Of course, each class cohort varies from the others in the number of strong and weak students in each class. In addition, the newer groups have generally been at Franklin longer and so should have benefitted from a well articulated curriculum. It becomes important, then, to try and account for differences in student knowledge level. If a group of students in 1996 earn a 650 on the Terra Nova math exam and a second group in 2001 also earn 650 on the same exam, one would expect that the average OSAT scores for the two students are also quite close. If a consistent shift in test scores is occurring, it should become more apparent using multiple regression analysis. Results for regression analysis are reported for the third and fifth grade math and reading assessments.

Third Grade Reading Assessment

The first data set to be examined was the 1997-2001 third grade OSAT Reading Assessment.⁵ The resulting regression model is as follows:

$$\begin{aligned} \text{OSAT Reading Score} = & 49.26 + 0.253 \cong \text{Terra Nova Reading Composite Score} \\ & (11.78)^{***} \quad (0.019)^{***} \\ & +1.092 \cong 1999 \text{ Year} \quad +4.037 \cong 2000 \text{ Year} \quad +1.136 \cong 2001 \text{ Year} \\ & (2.14) \quad (2.12)^* \quad (2.13) \end{aligned}$$

F Statistic = 52.23
 Sample Size = 104
 $R^2 = 0.6785$

The number in parentheses is the estimated standard error for the coefficient above it. An asterisk means the coefficient is significantly different from zero at a 90% confidence level, i.e., there is only a 10% chance that the true coefficient is actually zero. Two asterisks mean the coefficient is significant at the 95% confidence level and three asterisks mean the coefficient is significant at the 99% confidence level.

Stated verbally, the estimated model indicates that the Franklin third grade reading assessments for 1997 for each student can be predicted by adding the intercept (49.26) to 0.253 times the weighted CTBS 1997 Vocabulary and Reading Comprehension scores for each student. Both these coefficients are significant from zero at the 99% confidence level. To predict the 1999 OSAT score, one would follow the same process (using the 1999 CTBS results), but add 1.092 to the score. This 1999 coefficient had a standard error of 2.14, meaning that substantial uncertainty existed whether this coefficient was in fact zero, i.e., the OSAT score had apparently not been inflated from 1997 to 1999. However, in 2000 the coefficient is large (4.037) and significantly different from zero. We can interpret this result to mean that students performing at a particular level in 2000 would score on average about 4 points higher than the same performance in 1997. In other words, it seems pretty likely that the ODE inflated the 2000 scores by 4 points or so above the previous year. In 2001, the estimated coefficient of 1.136 was not significantly different from zero. This result suggests the 2000 inflation by ODE may have been at least partially unintentional and they attempted to correct the scoring on the 2001 exam.

Third Grade Mathematics Assessment

The second data set to be examined was the 1997-2001 third grade OSAT Mathematics Assessment. The resulting regression model is as follows:

$$\begin{aligned} \text{OSAT Math Score} &= 83.13 && + 0.203 \cong \text{Terra Nova Math Composite Score} \\ &(11.42)^{***} && (0.019)^{***} \\ &+ 2.29 \cong 1999 \text{ Year} && + 3.36 \cong 2000 \text{ Year} && + 5.00 \cong 2001 \text{ Year} \\ &(2.09) && (2.01) && (2.04)^{***} \end{aligned}$$

F Statistic = 38.08
 Sample Size = 104
 $R^2 = 0.6061$

The results for the mathematics assessment are much more striking and consistent. A small increase from 1997 to 1999 was identified, but the variability around this estimate was such that one cannot say with much certainty that this coefficient is different from zero. In 2000, the coefficient was larger but still not statistically significant. The 2001 coefficient was larger than that for 2000 and was significant at the 99% level. These results strongly suggest something unusual is occurring with this test. Either the ODE is “dumbing down” the mathematics multiple choice test in the third grade or the teacher involved is teaching to the test in a way that causes the OSAT scores to increase without increasing the Terra Nova. If the test is being “dumbed down” and these relationships hold statewide, they suggest that third grade mathematics scores actually declined by 1.2 points from 1997 to 2001, rather than increasing by 3.8 points as reported in Table 1.

Fifth Grade Reading Assessment

The third data set to be examined was the 1996-2001 fifth grade OSAT Reading Assessment. The regression model follows:

$$\text{OSAT Reading Score} = 64.34 + 0.238 \cong \text{Terra Nova Reading Composite Score}$$

(9.67)*** (0.015)***

$$\begin{array}{cccc} -2.57 \cong 1997 \text{ Year} & +0.93 \cong 1999 \text{ Year} & +0.73 \cong 2000 \text{ Year} & +2.87 \cong 2001 \text{ Year} \\ (1.78) & (1.67) & (1.68) & (1.77) \end{array}$$

F Statistic = 68.109

Sample Size = 127

$R^2 = 0.7378$

Like the third grade, the results for the fifth grade reading assessment are not very conclusive. The difference between 1996 and 1997 is actually negative, although statistically insignificant. In 1999 and 2000 there is a small increase from the 1996 level, but the estimates are very insignificant. In 2001, however, the estimated coefficient (2.87) is significant at the 89% confidence level. Based on these results, it appears that the fifth grade reading scores have not behaved unexpectedly except perhaps in 2001. If the 2.87 adjustment is correct statewide, it suggests reading scores have increased less than 3 points from 1996 to 2001, not 5.8 points as suggested in Table 2.

Fifth Grade Mathematics Assessment

The final data set to be examined was the 1996-2001 fifth grade OSAT Mathematics Assessment. The resulting estimation was as follows:

$$\text{OSAT Math Score} = 76.77 + 0.210 \cong \text{Terra Nova Math Composite Score}$$

(11,58)*** (0.017)***

$$\begin{array}{cccc} +5.07 \cong 1997 \text{ Year} & +8.76 \cong 1999 \text{ Year} & +10.26 \cong 2000 \text{ Year} & +14.74 \cong 2001 \text{ Year} \\ (2.16)*** & (2.08)*** & (2.05)*** & (2.10)*** \end{array}$$

F Statistic = 50.84

Sample Size = 127

$R^2 = 0.6775$

These results are the most startling of all, as they suggest a major shift in the fifth grade OSAT scores relative to the Terra Nova over the 1996-2001 period. Each year the scores increased relative to the Terra Nova, so that by 2001 the OSAT scale score was almost 15 points higher than the amounts reported in 1996. If this estimate is an accurate reflection of the state test scores, it means that the average OSAT mathematics score did not increase by 7 points as reported in Table 2, but actually fell by 8 points during that same period.

The regression results reported here involve very simple linear models, using the composite Terra Nova mathematics and reading scores to predict the respective OSAT scores. In fact the Terra Nova exam in mathematics consists of a mathematics component and a second component focused on computation. The Terra Nova for reading is also divided into two components. Given the relative weighting of these two components may differ between the OSAT and the Terra Nova, one might expect a better predictive model could be estimated by using the component scores rather than the overall average. This is in fact the case. The variables generated for each year in each model remain equally significant or actually become more significant using component Terra Nova scores.

Test Preparation by Teachers

It is apparent from the Franklin results that something unusual is occurring with the OSAT scores, particularly for mathematics. Students are showing increased scores on the OSAT that aren't being reflected in the equivalent Terra Nova tests. If students are really gaining additional knowledge, both tests should be moving in tandem because the OSAT and Terra Nova mathematics tests are so similar. The hypothesis of poor test equating was already ruled out and the results from Franklin only confirm the rejection of that hypothesis (especially for mathematics). That leaves the other two alternatives: Either the state is "dumbing down" the OSAT over time (especially for mathematics), or the Franklin teachers are teaching to the test. Specifically, Franklin third and fifth grade teachers are prepping students to do well on the OSAT.

The issue of Franklin teachers teaching to the test is very easy to resolve by interviewing them to learn what efforts they make to prepare students for these assessments. At Franklin School, there is only one teacher at each grade level, so the preparation process is the same for all students. Both teachers adhere quite closely to the Core Knowledge curriculum. This curriculum is used in over 1000 schools nationwide and is designed to provides teachers an outline of curriculum to be covered at each age level. In teaching math, both teachers (as well as all teachers at Franklin) utilize the *Saxon Math* series.

The third grade teacher has been at Franklin since it opened in Fall 1995. She has a general idea of the topics covered on both the Terra Nova and OSAT, and makes sure that these items are covered as part of the curriculum. She firmly believes that students should not be prepped for assessments, so she does not use old exams or any other special method to prepare students. She does, however, spend some time teaching test taking skills, but believes these skills would be equally useful when taking the OSAT versus the Terra Nova. The curriculum calls for reading

particular novels, working on vocabulary words and practicing spelling. Based on her observation of students since 1995, she believes the OSAT has indeed been “dumbed down” during that time.

The Franklin fifth grade teacher has been at Franklin since Fall 1997. He does spend some time in his class prepping for the assessments. He indicated that there is little that one can do to prepare for the reading assessment, other than practice answering questions like those on the reading exam. In preparing for the math assessment, he uses 1997 practice OSAT mathematics exams provided by ODE. He also spends time teaching test taking skills. He has never closely examined the Terra Nova because he wants it to use it as an accurate assessment of student performance. He believes the preparation described above will also be useful in preparing students for the Terra Nova. This teacher’s observation regarding the OSAT mathematics exam was that the students found most of the questions to be quite easy after spending a year in the *Saxon* textbooks.

From these interviews, it is very clear that the leaps in the third grade math assessments were not caused by the teacher teaching to the OSAT test. The much larger leaps in the fifth grade math assessment may be partially attributed to using old exams as practice, but not completely. Again, there is so much overlap in the questions asked on the Terra Nova and OSAT mathematics exams that focused preparation for one should also help students on the other. A 15 point increase in the fifth grade mathematics OSAT score means that the average student was correct on roughly 20% more of the OSAT questions in 2001 than in 1996, given the same performance on the Terra Nova. It seems unlikely that using OSAT sample tests as practice could generate this kind of increase on average for Franklin fifth grade students.

Inflation Effects on Meeting Standards

One obvious reason why the ODE would want to “dumb down” the tests is because it increases scale scores and, in turn, the percentage of students scoring at or above the standards mentioned earlier in this paper. A higher percentage of students meeting the state benchmarks would provide compelling evidence that the educational reforms begun 10 years ago have improved the quality of education for Oregon students. The impact of “dumbing down” the exams can be clearly illustrated using the Franklin test results. Assuming that the estimates reported for Franklin School are accurate indicators of the relationship between OSAT and Terra Nova scores, one can estimate what the 2001 third and fifth grade assessment scores would have been if the 1996 or 1997 exams had been used. From these estimates one can calculate the percentage of students meeting and exceeding the third and fifth grade standards.

-----Scale Scores-----			
Below			
	202 (Below)	202-214 (Meets)	215+ (Exceeds)
Third Grade			
Reading (2001 Scale)	7%	22%	71%
Reading (1997 Scale)	0%	33%	67%
Math (2001 Scale)	11%	33%	56%
Math (1997 Scale)	4%	67%	29%
-----Scale Scores-----			
Below			
	215 (Below)	215-230 (Meets)	231+ (Exceeds)
Fifth Grade			
Reading (2001 Scale)	4%	22%	74%
Reading (1996 Scale)	3%	41%	56%
Math (2001 Scale)	4%	36%	60%
Math (1996 Scale)	22%	59%	19%

Keep in mind that these are estimates, that the actual score a student receives on the OSAT may well differ from the predicted values. However, the error occurs in both directions around the estimated value, meaning that the actual OSAT score may be higher or lower than the predicted value. This summary illustrates well why the ODE might be motivated to “dumb down” the state assessments. Even small scale increases, such as those for reading in the third grade, can shift the percentage of students in each category. In the case of fifth grade mathematics, the inflation of scores completely distorts these results. Its important to note that Franklin students, on average, score well above average on both the OSAT and CTBS/Terra Nova exams. For this reason, deflating the 2001 scores had relatively little impact on the percentage of students not meeting the state benchmarks. For schools with student that are at or below the state averages, one can expect major shifts would occur if the 2001 test scores were deflated to the 1996/1997 levels in the manner predicted by the regression models.

Another area where artificially high test scores have a major impact is on the Oregon School and District Report Cards. According to the 2002 Report Card Manual, a major portion (80 percent) of the school rating is determined by student performance on the OSAT. Student performance, in turn, is based on two criteria: (1) Performance in the current testing year and (2) improvement in test performance. Seventy percent of the current year performance is determined by the scores for the mathematics and reading multiple choice assessments. Improvement in test performance is calculated using the change in mathematics and reading exams over the last 3 years. In short, the reading and mathematics assessments (shown by the Franklin results to be inflated) are by far the most influential data used to calculate the Oregon Report Card. It should not be too surprising, therefore, that the percentage of schools in the exceptional and strong categories has increased:

Percent of Schools by Rating Category:	2000 Report Card	2002 Report Card
Exceptional	3.5%	4.5%
Strong	35.3%	50.6%
Satisfactory	56.8%	43.6%
Low	3.9%	1.3%
Unacceptable	0.4%	0.0%

Again, if the tests are indeed being “dumbed down”, these increases could well be an illusion. If the scores in reading have remained flat in third and fifth grade, while scores in math have declined, then it seems likely that the number of exceptional and strong schools has declined while those in the satisfactory, low and unacceptable categories has increased.

Summary and Conclusions

This study provides some significant insights into the history of the ODE’s state assessment system and the possible “dumbing down” of the OSAT.

- Until the 1970's, Oregon schools relied on the Iowa Test of Basic Skills or other similar tests to conduct student assessments.
- In 1973, the legislature funded a ODE-developed assessment program because they were persuaded that Oregon’s academic objectives were sufficiently different from other states that the assessment tools then in existence could not meet the needs of Oregon schools.
- Over time ODE added more tests and increased the frequency of assessment. In 1991, they began the OSAT program to assess each Oregon student.
- No attempt has been reported to date to show how the OSAT compares or differs from other assessments used at the K-12 level in the United States.
- Although in the 1980's the ODE expressed a desire to identify other tests that could be used locally to meet state standards, apparently no effort was made in this area.
- The statewide assessments were initially very ambitious in terms of subject matter tested. In recent years, however, the assessments have narrowed to multiple choice tests in reading, mathematics and science, and the open ended mathematics and writing exams. An assessment in social studies may yet be developed, but assessments in subjects like language arts, geography, spelling, health, physical education and study skills do not seem to be in the mix anytime soon.
- There has been a definite shift in emphasis away from reporting of scale scores and toward reporting the percentage of students meeting and exceeding statewide standards.
- There has been an upward trend in statewide scale scores for mathematics and reading, particularly since 1996 and for third and fifth grade students. The sudden upward turn around 1996 for all third and fifth grade tests, along with consistent growth every year since then, seems puzzling. Two possible explanations are (1) the OSAT has been “dumbed down” over time or (2) teachers are putting more effort into teaching to the assessments.

- A comparison of the fifth grade OSAT and Terra Nova mathematics exams reveals that both cover the same major topic areas and contain similar types of questions. Although derived from a different theoretical base, both assign scale scores that permit identification of student performance from very low to very high levels.
- A plot of Terra Nova and OSAT third and fifth grade test scale scores for Franklin School reveals that the two test results are consistent with one another and exhibit a linear relationship.
- A statistical analysis of OSAT and CTBS/Terra Nova for Franklin School provides compelling evidence that third and fifth grade math scale scores from 1997 to the present have been increasing for the OSAT in a manner inconsistent with increases in the CTBS/Terra Nova. Reading scores may also be increasing in a similar manner, although the evidence is less clear-cut on this point.
- The third grade teacher at Franklin is opposed to efforts to teach to the assessments, other than teaching general test taking skills that should be equally valuable regardless of the assessment taken. The fifth grade teacher uses old OSAT practice exams to help prepare his students for the OSAT and Terra Nova, but believes these help students for both exams. Consequently, it seems unlikely that the unexplained increases in the OSAT mathematics scores at Franklin are solely the result of teachers teaching to the test. Based on the statewide and Franklin results, after considering the possible explanations for increases in OSAT scale scores, the most likely explanation at this point is that the OSAT has been systematically “dumbed down” over the last 5-6 years.
- Inflating scores can cause major differences in the percentages of students meeting and exceeding state benchmarks in these key subject areas. It also has a major impact on the ratings given in the Oregon Report Card.

Although this study provides some very compelling evidence that the OSAT is being “dumbed down” in some cases, the methods here simply cannot provide definitive proof. Scores statewide may be increasing because teachers are teaching to the test. The sample data set from Franklin School is relatively small and tends to be skewed toward the high performance end. Nevertheless, it is remarkable that statistical significance was achieved with such a small data set. Larger samples tend to only strengthen the tests of statistical significance in these kinds of studies.

Despite the inability to provide definitive proof that the OSAT is being “dumbed down”, the information provided here, taken together, is significant and worthy of further investigation. The state has the motivation and the means to “dumb down” the assessments, and the data suggest that this could well be happening. If ODE is indeed “dumbing down” any of its assessments, it means that many of the claims being made are simply not true. For example, consider this statement in a report entitled *An Assessment of Oregon’s K-12 Educational Reform*:

“Fifth Grade: The share of Oregon fifth-graders meeting the state reading standards rose from 51 percent in 1991 to 69 percent in 1999, and the share reaching state math standards rose from 47 percent in 1991 to 66 in 1999.”

In point of fact, the percent meeting the fifth grade standards in 2001 was 77 % for reading and 73% for mathematics. If the Franklin results are a valid indicator of the results statewide, then certainly the percentage of students meeting the math standards has *declined* over the last decade and the percentage meeting the reading standards has probably not increased as much as was claimed. Instead of 73% meeting the fifth grade standard in mathematics, the actual percentage may well be below 40%. And of course, this throws into question the validity of scores on assessments in eighth and tenth grades, as well as the science, writing and mathematics problem-solving tests.

Although the alleged problems identified with the state assessment process are very serious, there is a clear solution. The legislature has given the ODE responsibility for implementing the educational reforms passed in 1989 and 1991. Therefore, *the state legislature should require that all assessment of K-12 students be done using tests developed by entities operating outside the state.* Corporations are not allowed to audit their own financial records. Similarly, it seems to be poor public policy to give ODE responsibility to conduct an educational “audit” of Oregon students, when they are accountable for the success of the educational reforms. Parents deserve to know how their children are performing in school. The public deserves to know how well their schools are carrying out the mission they have been assigned. The legislature deserves to know how the ODE is doing in carrying out the educational mandates given them a decade ago. Student assessment by an independent entity is the key to meeting the needs of parents, the public and the legislature.

References

- Ahmann, J. Stanley, Marvin D. Glock, and Helen L. Wardeberg. *Evaluating Elementary School Pupils*. Boston: Allyn and Bacon. 1960.
- Applegarth, Boyd L. *Defining a Standard Education for Oregon Students*. Oregon Department of Education, Salem. Spring 1990.
- Bracey, Gerald W. *Thinking About Tests and Testing: A Short Primer on "Assessment Literacy."* American Youth Policy Forum, Washington D.C. 2000.
- Bunn, Stan. "Oregon Students Score Higher on Three of Four Math, Writing Tests." News release from the Oregon Department of Education, July 2, 2001.
- California Test Bureau. "Teacher Uses of Test Results." *Educational Bulletin No. 5*. Los Angeles, 1945.
- Dietel, Ron. Specialist, National Center for Research on Evaluation, Standards and Student Testing. Email communication, February 6, 2002.
- Duncan, Verne A. *1985 Oregon Assessment: Reading*. Oregon Department of Education
- Duncan, Verne A., Mary Hall, and James Impara. *IMPACT of Oregon Education: an Assessment of Reading, 1975*. Oregon Statewide Assessment Program, General Report. Oregon Department of Education, Salem. December 1975.
- Holmes, S.E. *ESEA Title I Evaluation and Reporting: The Linking Project, Final Report*. Oregon Department of Education, ERIC #TM800153. Salem. 1980.
- Neuburger, Wayne. *Eighth Grade Assessment of Mathematics and Reading: 1989 Summary Report*. Oregon Department of Education, Salem. June 1990.
- Oregon Business Council, K-12 Education Task Force. *An Assessment of Oregon's K-12 Education Reform*. Portland, OR. June 2000.
- Oregon Cooperative Testing Services. *OCTS Bulletin*. Number 1, February 1954. University of Oregon, Eugene.
- Oregon Department of Education. *Mathematics Test Specifications, Benchmark 2*. Office of Assessment and Evaluation, Version 4.0. Salem, July 2000.
- Oregon Department of Education. *Oregon School Report Card, 2002 Manual*. Salem, OR. Web URL <http://reportcard.ode.state.or.us/docs/2002RCMan.pdf>.

Patterson, Chris. *From TAAS to TAKS: A Progress Report on New Assessments for Texas Public Schools*. Texas Public Policy Foundation, Austin. January 2002.

Perry, Gregory M. "Comparison of OSAT and Terra Nova Test Scores at Franklin School, Corvallis, Oregon." Unpublished manuscript in possession of the author. December 27, 2001.

Stotsky, Sandra. *Analysis of the Texas Reading Tests, Grades 4, 8, and 10, 1995-1998*. November 1998. Available on the internet at <http://www.educationnews.org>.

Wolfe, Barbara. Oregon Department of Education, Assessment Office. Email communication, dated July 28, 2000.

Appendix
Scatter Plots for Franklin Test Data

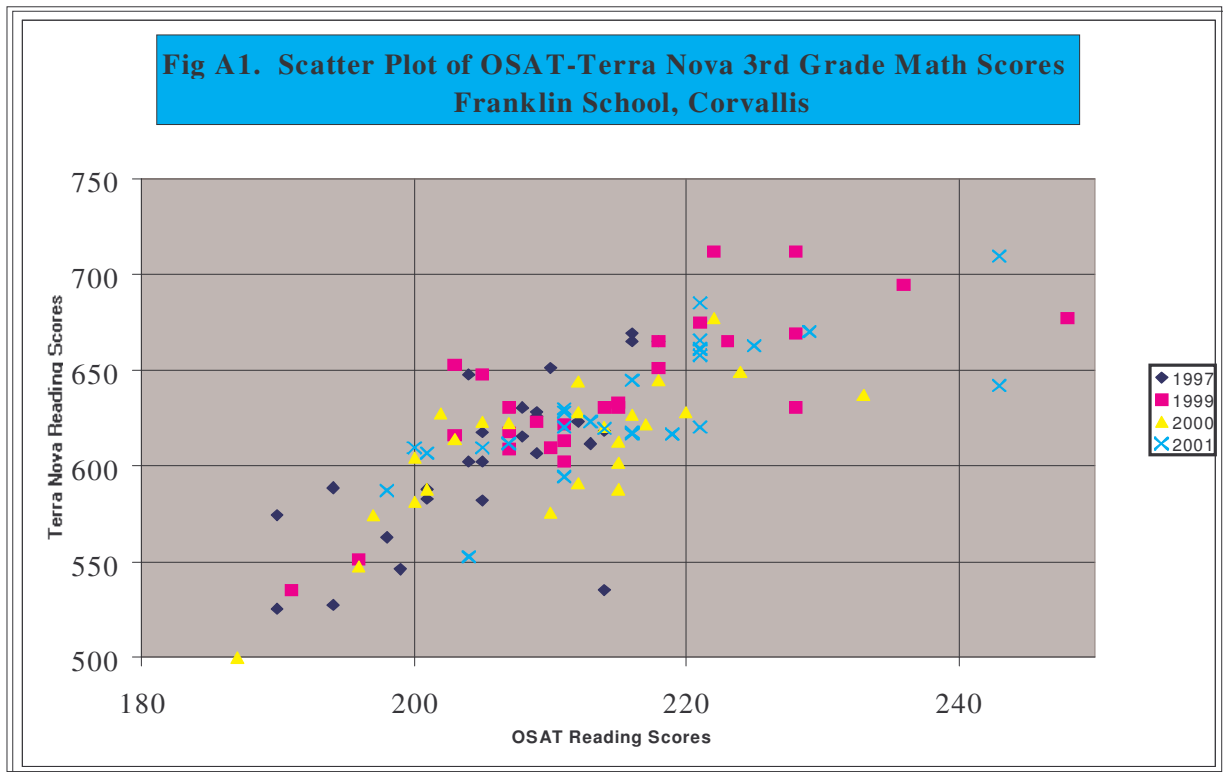


Fig A2. Scatter Plot of OSAT-Terra Nova 3rd Grade Reading

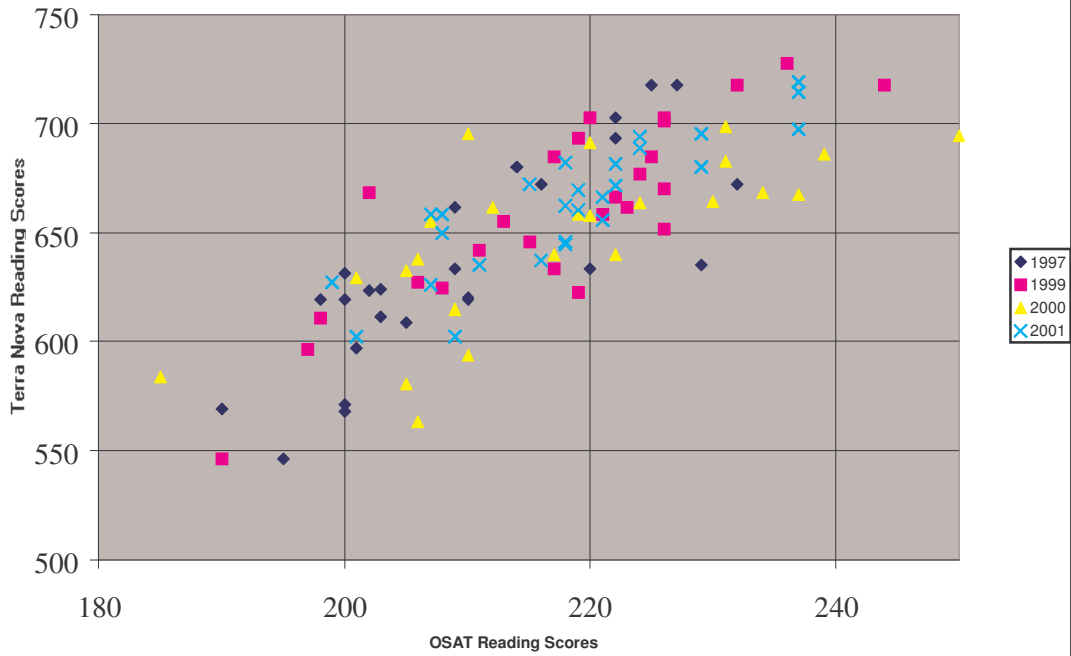


Fig 3. Scatter Plot of OSAT-Terra Nova 5th Grade Reading

