Running head: AN ANALYSIS OF SCANTRON TESTING

An Analysis of Scantron Testing

Jim Davis

National Louis University

TIE 593

Executive Summary

Standardized testing is an important part of student, teacher and school evaluation. Chicago Public Schools recently introduced a new assessment tool, the Scantron Performance Series. The Performance Series is novel for Dvorak Technology Academy in that it is (a) a computer-based assessment, and (b) it is a computer-adaptive assessment. The general consensus of the literature on possible mode effects when switching from paper-and-pencil tests to computerized tests is that mode effects are minimal, although the research does not directly apply to testing of elementary students where the tests have a low consequence for the students.

How should Dvorak use Scantron results? The Scantron results may be used for many possible purposes including guiding instruction, measuring student growth, evaluating teachers and evaluating overall school performance and success. Is Scantron a valid measure for these purposes?

This paper analyzes the results of Dvorak's Scantron experience since the winter of 2010, using the Illinois Standards Achievement Test (ISAT) as a control. The analysis looks at the correlation of Dvorak Scantron scores with ISAT scores at both the grade level and the classroom level, as well as comparing the national percentile rankings from both test series. In addition, the analysis looks at specific Scantron features, including student gain patterns, the Scantron standard error of measure, and test spoilage to see if there are any other issues related to test validity and reliability.

The analysis concludes that Scantron is in general a valid and reliable instrument for most students, but for a sizable number of students it falls short. The test is useful for suggesting areas of instruction emphasis, and providing additional evidence of student growth, but it should not be used for high-stakes purposes for teachers.

The report concludes with recommendations for maximizing useful test results, and guidelines for the responsible use of Scantron test results. Recommendations include training students in the new testing regime; gaining student support for the new test regime; providing assistance to teachers to maximize the usefulness of test results; and avoiding the use of Scantron results to measure teacher performance.

Table of Contents

An Analysis of Scantron Testing

Introduction

Dvorak Technology Academy is a Chicago public school of about 600 pre-K through 8th grade students in the North Lawndale community on Chicago's west side. Its student body is overwhelmingly African-American (99.5%) and poor (94% low income). Dvorak has failed to make Adequate Yearly Progress (AYP), as defined by the No Child Left Behind Act (NCLB), for several years now, and is in its fourth year of Academic Watch Status. It qualifies for "restructuring implementation" under the Department of Education guidelines, which means one of several harsh steps may be taken to attempt to alter school performance (Northern Illinois University, 2010; Office of Performance, 2010c). The primary measuring stick for AYP in Illinois for K - 8 schools is the Illinois Standards Achievement Test (ISAT), a battery of reading, math and science tests, primarily consisting of multiple-choice items.

In addition to the requirements set forth by NCLB, Dvorak and most other CPS schools are the subject of a "performance management" program initiated by recently departed CPS Chief Executive Office Ron Huberman. The CPS implementation of "performance management" relies primarily on student standardized test data, including ISAT, but supplemented with a new computerized testing regime from Scantron Corporation, as well as "five-week assessments", still another set of standardized tests whose administration and format varies across the CPS administrative areas. Performance management data feeds into a formula that CPS uses to determine which schools are "failing" and should be closed or become a "turnaround" school. Test scores may also become an important part of teacher evaluations, on which employment will hang (Illinois Government News Network, 2010).

Standardized tests then fit prominently into the life of Dvorak, as it does with most other public schools in the United States. Dvorak's success or failure with standardized tests mark the school as a success or failure.

Standardized testing represents a specific facet of technology in education. The traditional paper-and-pencil tests that have been in widespread use for several decades are the tip of a technological iceberg -- their widespread use would not be economically feasible were it not for the optical scanners and computer processing and storage used to score tests and report results. Likewise, the form of the assessment, the ubiquitous multiple-choice tests, is in part due to the relative ease with which computers can score the tests. The prevalence of personal computers connected to wide area networks, and in particular, the Internet, allow data to be shared at all levels of a large organization like CPS.

Over the past year, CPS has begun taking steps to have students take tests on a computer. Computer-based test-taking has additional economic and educational benefits (Wang, 2007). Tests can be scored immediately upon completion, and the test results made available to educators. This in turn has led to a trend towards "data-driven instruction" (see, e.g., Love, 2009). To provide computer-based test-taking capability to 600 some elementary schools itself is a sizable technology challenge. When these tests become "high-stakes", because promotion, employment and the existence of a neighborhood school depends on them, the question of how well the tests reflect actual student ability and learning becomes very important. And do computers affect the test-taking process? What mode effects might a school expect, and plan for, to ensure that test results are representative of a students' ability? After three computer-based test sessions, what lessons can be learned? These perhaps are the wrong questions, because they assume too much. What is the proper role for these tests?

These are questions this paper will explore in the context of a CPS elementary school.

History

The growth of standardized testing mirrors the development of 19th century science. As education, psychology and social studies moved onto a scientific footing, different types of tests gained in popularity as data-collection tools (Zytowski, 2008). The early 20th century saw developments in assessment like Binet's intelligence quotient (IQ) test and its variants. Tests were developed to assess career fitness and interest, military officer potential, and academic achievement. In the latter case, some notable events included the development of the Scholastic Aptitude Test and the Stanford Achievement Test in the 1920s, the Iowa Test of Basic Skills in 1935, and the California Achievement Test, today the Terra-Nova, which appeared in 1950 (FundingUniverse, 2010; Young, 2005; Zytowski, 2008).

The increase in numbers of tests administered created the need to find faster and accurate ways of scoring tests. During World War I, the United States Army administered some two million IQ tests as part of its Alpha program, all scored by hand (Zytowkski, 2008). While the development of the multiple-choice test in 1915 (Young, 2005) provided the structure for simplifying test scoring, a cheap, practical, automated scoring process did not emerge until the development of computers. A number of mechanical and chemical processes were developed during the 1920s and 1930s to speed up (and reduce the cost of) the scoring process, but these processes were still essentially manual, labor-intensive and relatively expensive. Early attempts at automation, including one that used the conductivity of graphite showed promise, but as Zytowski (2008) points out, "the cost, not to mention size and weight, of the [the scoring machine] limited its application to high-volume users, typically universities or test publishers." Even with the scoring problem on its way to resolution, effective ways of storing, processing and reporting data remained problems. The post-World War II digital revolution provided the various components for automating test scoring. The use of optical mark recognition (OMR) scanners, connected to computers, first used in the early 1950s, led to a standardized testing revolution, iconized by the #2 pencil and the bubble sheet.

As the capabilities of computers increased at the same time as the cost of computing dropped, standardized testing began shifting to a personal computer/Internet-based platform. When the Educational Testing Service introduced a computer-based Graduate Record Examination in 1993, a *New York Times* article summarized key benefits of testing on a computer: "Instead of sitting in a room with hundreds of other people on one of five annual test dates, students will be able to go to a computer center and take the G.R.E. on any of several days during the week, for a total of more than 150 days a year. Instead of waiting four to six weeks for results to arrive in the mail, students will be able to press a key on their computer at the end of the exam and get their scores immediately" (Winerip, 1993).

A re-telling of the rise of testing and the educational philosophy behind it, is beyond the scope of this paper. Suffice it to say that, like most school districts, Chicago Public Schools (CPS) has employed various standardized tests as part of its educational strategy for many years. In 1972, CPS began using the Iowa Test of Basic Skills (ITBS) as its primary assessment tool for determining student academic achievement. In 1988, the Illinois Goals Assessment Program

(IGAP) became the state wide assessment program. For a number of years, Chicago public school students took both the IGAP test and the ITBS. The IGAP was given in early March and the ITBS in late April. During the same time, Chicago Public Schools developed their own set of learning goals, separate from those of the state of Illinois. Gradually, Chicago Public Schools standards became more aligned with the Illinois State Board of Education (ISBE) standards and CPS accepted the Illinois Standards Achievement Test (ISAT) as the tool to measure student achievement, which also served as the Illinois assessment tool to satisfy the requirements of the No Child Left Behind Act. In 2005, CPS replaced the ITBS with Stanford Learning First reading tests, delivered in the fall, winter and early spring (Dell'Angela. 2005). These tests were intended as diagnostic tools for teachers and were not factors in student promotion. Math tests were added to the Learning First battery in 2006. Extended response questions for both reading and math also became a part of the tests. In the 2009-10 school year, these tests were referred to as the Chicago Reading Benchmark Assessment and the Chicago Math Benchmark Assessment.

In the fall of 2009, Chicago Public Schools began the process of implementing the Scantron Performance Series assessment which will supplant the Benchmark Assessments. The new Chief Executive Officer of CPS (as of March, 2009), Ron Huberman, has introduced an aggressive program of "performance management" in the evaluation of school and teacher performance. This includes the use of "value-added" measures, where standardized test scores are assumed to measure student growth, and are in turn used as indicators of teacher performance (Meyers, 2009; Harris, 2010). Because the Scantron Performance Series provides an immediate, grade-independent scaled score, and is administered three times a year, it provides a regular and timely measure for the value-added metric of teachers and schools. Because

Scantron will provide performance metrics unavailable with the local Chicago Benchmark Assessment, the computer adaptive tests will eventually replace the Benchmark Assessment program. By the end of the 2009-10 school year, all CPS schools should have administered the Scantron assessment at least once.

In the 2010-11 school year, CPS schools began yet another test-taking regime. The specific assessment tool has been left up to each CPS administrative area. These assessments were initially supposed to be administered seven times a year, or approximately mid-quarter and end-of-quarter, to provide data on student progress approximately every five weeks. Dvorak's area (Area 10) is supposed to implement a computer-based test from Riverside, the publishers of the Iowa Test of Basic Skills. The Benchmark Assessment tests will be optional for Chicago schools in the 2010-11 school year (CPS Office of Performance, 2010a).

Areas of concern

The introduction of a new test series raises the question of test validity (does it measure what it is intended to measure?) and reliability (does it consistently measure it?). Different testing regimes yield different packages of data. What is the best way to use that data? What the limits of what the data can and should be responsibly used for? These questions apply to any test regime. However, the shift from paper-and-pencil assessments to a computer-adaptive assessment presents another bundle of challenges.. There are major technology challenges, especially ensuring that there enough computers in the school to handle electronic testing. Training test administrators and teachers in test administration, scheduling testing time throughout a school, and preparing students for the assessment are also challenges. And with the change in test mode, might there also be unexpected or unintended differences in test results?

In most studies that compare student performance on different test modes, different groups of subjects take one test mode or another. The groups are identified as statistically similar. In general, previous research has found similar results in test results between paper-and-pencil versus computerized tests. The introduction of a new, computerized testing regime alongside of existing paper-and-pencil performance tests provides a unique opportunity to confirm if these findings hold up with the same students tested during the same time frame. If the results are different, why? Test publishers argue that their computerized tests yield similar results to other test tools (see, e.g., Scantron, 2010). How do these findings hold up with a very specific demographic, namely urban African-American students from low-income communities? If not, why not? If modal effects exist, do they change after several test administrations? Are there strategies which school administrations can use to minimize the modal effects?

Literature review

The possible difference in test results when the same test is administered via computer versus paper-and-pencil are called "mode effects." A computer-based test may simply be the same paper-and-pencil test done on a computer (referred to below as a computer-based test, or CBT), or the paper-and-pencil test (PPT) may be converted into a computer-adaptive test (CAT), where testing software determines what questions a tester will see, based on prior responses.

There are two areas where mode effects might arise. The first one is the change in media, from paper-and-pencil to computer (or PPT to CBT). The second area is the shift from a traditional test to an adaptive test (or PPT or CBT to CAT), where the questions get more or less challenging based on previous answers. The shift from a traditional PPT to a CAT combines two mode effects.

In looking at a media-effect only, Olsen (1986) reported on the results of the same test being administered in different formats. He found that paper-administered and computer-administered tests had comparable results, although interestingly, scores tended to be lower on the second test administered.

Mead and Drasgow (1993) conducted a meta-analysis of paper-and-pencil versus computerized tests to determine if the medium of test administration affected test results. Mead and Drasgow's review is noteworthy because it was chosen as the representative study of a body of research by the Scantron Corporation to support its computer-based adaptive Performance Series (Scantron, 2010). Mead and Drasgow found a high correlation (.97) for cross-mode (paper/pencil vs. computerized) timed power tests. Mead and Drasgow also found that computer-adaptivity, as opposed to simply a computer version of a PPT, was not a significant factor in test results.

Many other studies support Mead and Drasgow's findings, both between PPT and CBT, and those tests and CAT. Gorham and Bontempo (1996) looked at re-test rates for nurse licensing exam, and found little difference in re-test rates for PPT versus CAT. Schaeffer, Bridgeman, Golub-Smith, Lewis, Potenza and Steffen (1998), in a study funded by the Graduate Record Examinations Board, also found no significant difference in GRE scores between PPT and CAT. Bodmann and Robinson (2004) investigated the effect of several different modes of test administration on scores and completion times using undergraduates. They conducted experiments comparing PPT versus CBT, as well as the same test using different computer interfaces. They found little difference in scores, although completion times varied. Wang (2004), reporting on a Pearson company study, found little difference between the mean scores for PPT or CBT, with one exception, noted below. Wang also authored two meta-analyses of reading and math assessments, and found little mode effect (Wang, Jiao, Young, Brooks, and Olson, 2007, 2008[1]).

Still, despite the apparent uniformity of findings comparing test modes, there are also important issues raised in the studies that have implications for elementary school testing. For example, most of the tests included in the Mead and Drasgow (1993) meta-analysis were employment-related tests (like the Armed Services Vocational Aptitude Battery and the Western Personnel Test) or college admission exams (like the Graduate Record Examination). These are high stakes tests for the test-takers, and so test-takers have a high level of motivation to do well on the tests, no matter what the medium or mode. Likewise, Gorham and Bontempo looked at nurse licensing exams; and Schaeffer, Bridgeman, Golub-Smith, Lewis, Potenza and Steffen (1998) looked at GRE results -- again, high-stakes tests for the test-takers. The presumably high level of motivation by test-takers for the exams studied may not extend down to younger test-takers taking assessments with no consequences for them. This suggests a possible area of research for comparing tests (and not necessarily test media), in particular test results between high-stakes tests (where grade promotion is at stake) and low-stakes tests taken by the same student.

Mead and Drasgow (1993) also acknowledged that test content between paper/pencil and computerized tests as a moderating factor was not examined in enough of the studies included in

[1] It is perhaps worth noting these studies are by researchers associated with Houghton Mifflin Harcourt, a major educational publisher, and publisher of standardized tests, including computerized tests.

their meta-analysis. They made a particular note about the possible difficulty of reading long passages on a computer display. This limits the usefulness of their study, because reading long passages on a computer is an important factor in reading tests, and could affect test scores.

Some studies, while supporting little mode effect for older test-takers, do indicate issues for young test-takers. Choi and Tinkler (2002) reported on a study of K-12 students who took a PPT version and a CBT version of the same test. The scores on the two test modes showed little difference except for the youngest test-takers. Younger students had a more difficult time with the CBT reading test because of the unfamiliarity of the computer itself. The longer reading passages required the students to use a scroll down bar. The study found that some of the younger students had a hard time with comprehension due to having to stop reading and scroll down and pick up where they left off. Wang (2004) also found differences in PPT versus CBT for grade 2 test-takers, surmising that the differences were likely due to computer unfamiliarity for the younger children.

Although computer use proved to be a hurdle for some of the younger students, it did not seem to make a difference with older students who had the requisite technological skills. Another possible mode effect may arise from differences in what can be done with paper versus computers. Students are often taught reading strategies that include writing on the text (highlighting, underlining, margin notes). If the computer test-taking software supports it, the skills to highlight, strike-through or annotate can be complex and too much for many test-takers, especially younger test-takers.

Kolen (2000) outlined a number of issues that may cause a difference in scores when equivalent tests are administered in different formats or media. He presented a conceptual framework for organizing what he calls "threats to comparability" among alternate test formats. Kolen identified five main threats: differences in test questions (see e.g., Legg and Buhr, 1990); differences in scoring (of special relevance when weighting questions used in Item Response Theory, a key component of computer-adaptive testing); differences in testing conditions, which includes testing on a computer versus paper-and-pencil; differences in examinee groups; and violations of statistical assumptions used to establish test comparability. He then applied these categories to three applications, including paper-and-pencil versus computer-adaptive tests. Although Kolen described differing examinee groups as different test subjects, this category could be extended to include the same examinees, but tested at different times, as in the case of CPS students who take different tests within weeks of each other, but with different motivation, interest, physical state, and so on. Because paper-administered and computer-adaptive tests are very different in the nature of the questions, care must be taken in equating scores from the two types of tests. Besides differences in examinee groups, Kolen identified several test-condition issues that might affect the comparability of results between paper-administered and computer-adaptive tests. In particular, he noted "the ease of reading lengthy passages" (p. 85). Kolen noted that "mode effects for paper-and-pencil and computer-administered tests appear to be very complex" (p. 86) and will vary according to the testing program, so making generalized statements discounting mode effects and test comparability should be approached with healthy skepticism.

Although most of the studies of comparability, both of paper-and-pencil tests vs. computer-based tests and paper-and-pencil tests versus computer-adaptive tests suggest comparable scores between the two modes, there are a number of variables or conditions that might affect test results and have not been adequately studied in the literature reviewed above. As noted above, Kolen (2000) has outlined five categories that threaten score comparability. In addition, the studies have generally not taken into account external factors that might threaten comparability, like motivation. What happens when one compares a paper-and-pencil high-stakes test with a computer-based low stakes test, where both tests are intended to assess the same skills? Or in general, differing motivation might threaten score comparability. Factors like "test fatigue" -- student lack of interest or exhaustion as the result of repeated testing might affect comparability of results. These are not elements of the tests themselves, but are important elements of the overall testing process. The introduction of a new computer-adaptive testing regime alongside an established paper-and-pencil regime provides a unique opportunity to see how score comparability holds up in practice.

Although not directly related to mode effects, Corcoran (2010) describes a phenomenon where two different standardized tests yield different value-added measures for teachers. "We see that teachers who had high value-added on one test tended to have high value-added on the other, but there were many inconsistencies." These "inconsistencies" tend to be obscured by averages and correlations, but indicate potential problems when using tests for high-stakes purposes.

Observations

Dvorak grade 3 through 8 students have now taken the computerized Scantron Performance Series assessment three times: February, 2010; May, 2010; and September, 2010 (at the start of a new school year). In addition, students also took the paper-and-pencil Illinois

Standards Achievement Test (ISAT) in March, 2010. The close proximity of the test events provides a useful entry point for investigating possible mode effects on the Dvorak student body.

The general methodology of this analysis is to use student ISAT scores as a reference point, and compare Scantron scores with them. Both scores are scaled scores. Since ISAT is scaled for each grade level, comparisons are done for each grade. When grades are referenced in the tables, they refer to the students' current grade. A 4th grader took the 3rd grade ISAT in March, 2010. ISAT is familiar to older students, Scantron was a new test format in the winter of 2010. However, ISAT was also new to 3rd graders in March (although one 3rd grade class had also taken the NWEA online assessment, another computer-adaptive test). This section includes data observations, and is followed by a Discussion section, where some tentative conclusions are offered. The tables and figure appear in the Appendix.

When Dvorak grade-level ISAT reading scores and Scantron reading scores for any of the examined Scantron events are compared by grade level, they show a relatively high level of positive correlation, ranging from .73 to .83, with a mean of .79 (see Table 1). The ISAT-Scantron correlation for the math test is more varied, ranging from .58 to .84, but still demonstrates a relatively high positive correlation, with a mean of .76 (see Table 2).

The CPS Office of Performance recognizes a strong enough correlation to have released to principals a table of predicted 2011 ISAT scores based on Fall, 2010 Scantron scores, using 2009 Fall Scantron and 2010 ISAT scores for their correlation (Chicago Public Schools Office of Performance, 2010b). The Scantron Corporation has also published large scale correlations of the Performance Series based on the Spring, 2008 ISAT administration, with math correlations ranging from .749 to .823, and reading score correlations ranging from .755 to .844 for grades 4 -

8. Scantron uses these correlations to argue the "concurrent validity" of its assessment series (Scantron Corporation, 2010a).

For reading, the correlations across test sessions decrease or are stagnant. Math correlations show a variety of increase, stagnation and decrease.

The ISAT / Scantron scores per classroom show a wider range of correlations (see Tables 3 and 4).

For the reading test, the mean Fall 2010 correlation was .55; for math it was .53. One classroom in particular (7B) had a low correlation, close to zero, of both reading and math.

The gains for students between the Winter 2010 assessment and the Spring 2010 assessment, about 12 weeks later, show a wide range of values. Tables 5 and 6 show the gains by classroom. The standard deviation of the reading gains, the measure of how spread out the values are, ranges from 78 to 290. The range of the gain (or difference between the two test scores), ranges from 378 points to 1139. In the latter case, a student dropped by almost 570 points, and another rose by an almost equal amount. With a $\partial = 290$, assuming a normal curve, about two-thirds of the students scores changed within a rather large range of almost 600 points over a 12-week period. The math gains are not as widely dispersed.

When the sets of scores are plotted, there are distinct outliers, which tend to be at the lower end of the score scales (see Figure 1).

Both the Scantron Performance Series and ISAT report a grade-level national percentile ranking (NPR) of student scaled scores. About 15% of students that took the 2010 ISAT and the Spring 2010 Scantron (38 out of 250; Spring Scantron was taken about 8 weeks after ISAT) appear in the top 60th percentile and above on the ISAT reading test, but appear in the 40th percentile or lower on the Spring, 2010 ISAT. The numbers are similar for the math test, and when compared with the Fall 2010 Scantron assessment.

The Scantron Performance Assessment includes a safeguard against students simply clicking answers to finish the test quickly. If a student answers five questions "in a rapid fire manner at a rate faster than possible to even read the questions, and if those rapid answers are no better than guessing, the test will be spoiled" (Scantron Corporation, 2010b). In the report it makes available to test administrators, Scantron describes this reason for test spoilage as "testing irregularities." Table 4 shows the number of tests spoiled in Dvorak's Fall, 2010 test administration.

Fall 2010 test spoilage by class varied from no tests spoiled, to over 30% of the students in the class spoiling the math test at least once. There does not appear to be any relationship between the number of students spoiling a test, the correlation of test scores, or the phenomenon of students scoring in a high percentile on ISAT and a low one on Scantron. There is virtually no correlation between the percent of students spoiling the Fall reading test in a class and the class correlation between ISAT and Scantron (r = -0.05). There is a small negative correlation between the percent of students spoiling the Fall math test in a class and the class correlation between ISAT and Scantron (r = -0.29).

The Scantron Performance Series reports a Standard Error Measure (SEM) for its scores. The SEM expresses a range of scores that a student would be likely to get if the student re-took the test. For example, a score of 2000 with a SEM of 50 would mean that if a student immediately re-took the Scantron assessment the student would likely score between 1950 and 2050. The overall SEM numbers are relatively low (see Tables 9 - 11). The Reading SEM is higher than the Math SEM for all grades for all Scantron assessments examined, ranging from 23% to 79%.

One additional item related to the SEM is the SEM of the score difference between test administrations. The difference in scores between, say, the Spring 2010 and Winter 2010 administration (done about 12 weeks prior) is referred to as the gain, which can be positive (student scored higher on the later assessment) or negative (student scored lower). The SEM of the score difference (the gain) indicates the significance of the gain. If the absolute value of the gain is less than the SEM, then the score should be interpreted as effectively unchanged (or rather, no conclusions can be reliably drawn from the change). Table 14 shows the number and percentage of students with a drop (a negative gain) greater than the SEM of the scaled score (SS) difference between the Winter 2010 and Spring 2010 Scantron assessments.

Table 15 shows the numbers of students with significant drops in their Spring 2010 scaled scores who also had high ISAT / low Scantron NPRs. The table indicates a small overlap between the two groups.

Table 16 shows the SEM for individual strands assessed by Scantron. The strand SEM is in most cases more than twice as large as the overall SEM for the assessment. The table also includes the SEM for students who spoiled at least one test.

Although both ISAT and Scantron provide national percentile rankings (NPR), which suggests a common point of comparison to establish concurrent validity, the NPRs vary substantially for students. Table 13 shows the mean difference in the NPR scores between the two tests. The correlation between the NPRs is fairly high, which is to be expected -- overall, high or low performance in one test suggests high or low performance in the other. However as overall measures against a national sample, one or both of the tests are failing to give a clear picture. Over 68% of the students who took both the Winter 2010 Performance Series reading assessment and the 2010 ISAT reading test showed a greater than 25% drop in their Scantron NPR versus the ISAT NPR. In percentile terms, 45% of students showed a Scantron reading NPR more than 20 points lower that the ISAT reading NPR. The differences are not so great for the math assessments, although still high. Some 57% of students showed more that a 25% drop in the Scantron math NPR versus the ISAT math NPR; and 39% had a Scantron math NPR over 20 points lower that the ISAT math NPR.

Discussion

This analysis set out to investigate first, the validity and reliability of Scantron for Dvorak, and second, the possible presence of mode effects in the a new computer-adaptive test regime. The parallel use of the paper-and-pencil ISAT test provided an useful control for the investigation. If mode effects did exist, then the Scantron tests would yield different results than ISAT. The analysis above indicates that, in general, at the school level and above, ISAT and Scantron scores correlate positively and strongly. The Dvorak experience affirms the assertions by CPS and Scantron about the concurrent validity of the Scantron Performance Series. This indicates, like the meta-analysis research, that there is little mode effect between the paper-and-pencil ISAT and the computer-adaptive Scantron Performance Series.

On the other hand, the large differences between the NPRs for the two tests presents a puzzling discrepancy. What might account for the discrepancies? The Scantron NPR is based on a normative sample of Fall 2005 - Spring 2006 test-takers (Scantron, 2010a). ISAT derives an

NPR from answers on 30 norm-referenced questions from the Stanford Achievement Test Series 10 (ISBE, 2009). The ISAT NPR is based on a norm referenced group of the students taking the same test (so in this case, other students taking the 2010 ISAT test). The difference in norm groups might be one possible cause of the discrepancy. Perhaps a difference in methodology in calibrating scores resulted in a difference. The correlation between NPRs suggests that one of these two reasons may be the case. Still, a national percentile ranking by definition should yield similar numbers if the two tests are concurrently valid. This would suggest that one or both of the tests are not valid. Nevertheless, this analysis assumes that ISAT is valid, and for the following discussion and recommendations, assume that the correlation evidence establishes some correspondence of results. Research into the differences in NPRs requires further investigation.

The analysis at the grade level and above suggests the absence of a mode effect, and even more, that the introduction of a new standardized testing regime yields consistent results with established assessments. However, within the overall Dvorak data, and especially when examined at the classroom level and below, there are features and anomalies in the data that deserve a further look. Some of these features and anomalies may be related to mode effects, or due to other factors. The discrepancies noted above include:

(1) varying correlations of Scantron and ISAT by classroom

(2) large SEMs for tested strands

(3) the wide range of gains (from negative to positive)

(4) the numbers of students scoring in a high percentile on ISAT and a low percentile on Scantron

(5) issues around test spoilage.

The smaller correlations between ISAT and Scantron at the classroom level suggest that grade-level aggregation and above hides possible issues at the classroom and individual student level. When the sets of scores are plotted, there are distinct outliers, which tend to be at the lower end of the score scales (see Figure 1 above). This indicates that some students are not generating similar results across test regimes.

The Standard Error Measure (SEM) provides another possible indicator for reliability. A small SEM suggests that the Performance Series assessment has zeroed in on a student's performance, and the student is performing consistently, yielding similar scores on test re-takes. A large SEM, on the other hand, suggests that a student performance on the test is varying widely. The student is performing inconsistently. The size of the SEM, then, may serve as a reliability indicator for given students. A small SEM means a reliable score, a large SEM means a large possible range of scores, hence an unreliable, or at least highly variable performance such that the test may not accurately reflect student performance. It should be noted that the grade level range of SEM for reading has not changed with successive test administrations (see Tables 11 through 13). When comparing Winter and Spring 2010 grades with the next higher Fall 2010 Standard Error of Mean Scaled Scores, the numbers showed little difference. The lowest standard errors for Fall 2010 are generated by 3rd graders, who took Scantron for the first time.

As with other statistical processes seen above, the larger the pool of results, the more a general picture of the whole emerges, but interesting details are lost. For example, the overall SEM for the math and reading tests are much lower than the SEMs for individual learning strands (see Table 15). The suggests that the reliability of the assessments for specific strands is much lower than the test as an overall, albeit hazy, picture of student achievement.

As Tables 7 and 15 show, many students scored in a high ISAT NPR, but a low Scantron NPR. For the Spring 2010 assessment, the numbers for both reading and math are about 20% or one-in-five students. Even where students showed large discrepancies in the NPRs, the Scantron SEM was close to average, indicating that Scantron's algorithms determined that the results were consistent. In only one case, where a student appeared in the 60th percentile on ISAT, and the 1st percentile on Scantron (the student apparently had not made much effort on Scantron), was the SEM much higher than the average.

The wide range of gains within classrooms further supports the idea that the Scantron test results are not valid or reliable for some students. While not to belittle teacher efforts, that a student acquired enough new learning in 12 weeks to achieve a 500-point score gain seems absurd, and either the first or second test (or both) had erroneous results. Equally absurd is the notion that a student lost 500 points worth of skills in 12 weeks. Even the idea of losing more than the SEM of the score difference is counter-intuitive. Were the skills so shallowly learned that the student forgot the skills in three months? Or was there some other kind of cognitive failure? While possible, this does not seem reasonable. The 500-point drop is the extreme example, but that 1-out-of-6 students who took the Spring 2010 reading assessment showed a drop more than the SEM, and over 10% had a drop of more than 100 points more than the SEM (see Table 14) also challenges the validity and reliability of the assessment for many students. It should be noted that Scantron allows students to retake the test within the same assessment window, and administrators may spoil student tests. This in fact was done at Dvorak for some students whom teachers identified as having scores they felt were unreasonable. The retake almost always resulted in a higher score. Table 15 indicates that there was little overlap between the students who dropped significant points, and the students who showed large discrepancies in the ISAT and Scantron percentiles. Different sets of students are generating different kinds of inconsistent results on the Scantron test.

Test spoilage provides an interesting glimpse of student test-taking behavior. The test software has specific criteria for spoiling a test (see above), but the reason for the student tripping the test spoilage criteria may be for many reasons: student disinterest, difficulty with the testing medium, a desire to stay away from class, lack of focus, or even active or passive resistance to testing. As Table 8 shows, about 10 percent of students spoiled a math or reading assessment (slightly more math than reading -- see below), and about half of them spoil a test more than once (four students spoiled a test 10 or more times). When the spoilers successfully do complete a test, their scores correlate with ISAT, and their SEMs are close to the overall SEM, and in some cases less (see Table 16). This suggests that once the test is completed by a spoiler, the results tend to be more reliable overall. This might suggest that the encouragement required to get the student to finish the test also encourages them to work at their skill level. Spoilers are not necessarily the lowest performers. Furthermore, some classes show a high rate of spoilers (see Tables 9 and 10). This phenomenon may be controlled by teacher presence, clearer instructions and expectations, and making the test more meaningful to the student. Spoilers need to be coached (or cajoled) to complete a test, which may be a factor in a better effort. In any case, the phenomena of test spoilers reinforces the fact that test-taking is not the same as temperature-taking -- the test-taker must perform, and for valid test results, the test-taker must be a willing participant in the process.

The literature review above raised the possibility that an online reading test might show a greater mode effect than a math test. The results from the Dvorak experience are inconclusive. The higher SEMs for the reading tests than the math tests (in some cases almost 80%; see Tables 11 - 13) indicate that the reading test yields a wider range of scores, which indicates that something is different about taking the reading test. The larger drops in reading scores (negative gains) supports this idea as well. On the other hand, a higher percentage of students spoiled the math test than the reading test, which suggests that the nature of the online reading test did not pose a problem to students staying focused on answering the questions.

Unfortunately, this analysis has not turned up any identifiable patterns that might allow teachers and test administrators to identify students who might yield invalid results, that might possibly be addressed by intervention. After a test is finished, a significant drop in score may indicate low student effort. Teachers, from personal observation, can identify students whose scores do not reflect ability.[2] The re-take feature of Scantron does provide a means of easing of obtaining more accurate test results.

There is no conclusive evidence from the Dvorak data that indicates a mode effect. As outlined above, there are a number of issues with the Scantron test, but they could be attributable to many factors: student interest and motivation, test fatigue, teacher presence, self-confidence, self-motivation, as well as mode effects stemming from test design, computer familiarity, lab environment and so on. Mode effects could be transitory, the results of a brand new test medium that disappears as the medium becomes more familiar. This does not appear to be the case, at

[2] That teachers already have a good qualitative sense of student ability reinforces the idea that testing is not for the students or the teachers, but for school, district, state and national administrators.

least for most students. Mode effects could also be an ongoing phenomena, but this also does not appear to be the case at Dvorak with the Scantron Performance Series.

The Scantron issues do raise important concerns for teachers, parents and administrators. If the goal of testing is to accurately assess student learning, then something is amiss with the testing of these students. Most student results do not suggest problems (that is, their scores correlate with ISAT, the SEMs are low, their gains are within the SEM or show growth, tests aren't spoiled), but a sizable percentage do. This suggests that the Scantron test is useful, but within important limits. At the school-level, Scantron appears to provide a consistent overall grade-level performance picture when compared to other accepted reference points like the ISAT, in the narrow sense of specific skills. As a result, Scantron can help to identify student skill needs. However, the discrepancies in the data should precludes its use for high-stakes purposes, in particular teacher evaluation. This conclusion is consistent with other research on using test scores for teacher evaluation (see e.g., Schochet and Chiang, 2010 and Corcoran, 2010).

Recommendations

The experience of introducing a new testing medium to Dvorak Technology Academy, described above, suggests several actions for teachers, school administration, and district leadership. The recommendations for action fall into two categories. First, the lessons learned in introducing a new testing medium lead to several recommendations for introducing new test media. Second, the validity of new media, in particular the Scantron Performance Assessment, is ambiguous, and so leads to recommendations on what can be responsibly done with the data from the assessment.

Recommendation #1: Prepare students for new testing modes. The school administration should ave a training session to allow students to get familiar with test operation, as well as train students in the computer operations necessary to successfully complete the assessment. If a training mode is not available, then treat the first assessments as a training assessment, and provide students with necessary supports during the assessment period. This will also allow teachers to discuss test-taking strategies, and frame testing as a genre (New Teacher Center, 2009). Administrators should recognize that the initial assessments are for training purposes, and treat the data accordingly.

Recommendation #2: Get student buy-in to ensure useful results. Assessments will only yield valid and reliable results if students actively work to do well on a test. There are many reasons why a student may not make a serious effort on a test: lack of interest, fear of failure, lack of self-confidence, learned helplessness. If a test is not recognized as "high stakes", student engagement may be less. Somehow, school leaders need to incentivize test-taking. Ideally, students should be self-motivated to do their best on a test. If self-motivation is non-existent, external motivations may help. These could include a grade or individual or classroom rewards. Teachers should discuss the test results with students, to demystify test-taking, and help them understand that testing helps the teacher help the student. Teachers should discuss test results with students, so they can (a) see how they did and (b) see that the data is in fact being used, and may actually help the student.

Recommendation #3: Help teachers learn the test, how to use it, how to read the data, and how to talk to students about test results. This can be done via professional development, but it should not just be done by the test vendor. Training should be honest about the test medium, and what the test means and doesn't mean.

Recommendation #4: The results from a new test medium should not be used in evaluating teacher performance or for student promotion or placement. Test data should be treated as another source of data for the teacher and administrators, and not a summary of student or teacher performance.

References

Bodmann, S.M., Robinson, D.H. (2004). Speed and performance differences among computer-based and paper-pencil test. Retrieved May 30, 2010 from http://www.utexas.edu.

Chicago Public Schools Office of Performance (2010a). CPS elementary assessment calendar for 2010-11. Retrieved June 12, 2010 from http://research.cps.k12.il.us/export/sites/default/accountweb/Assessment/Elementary_School_Assessment_Calendar_FINAL.pdf

Chicago Public Schools Office of Performance (2010b). 2010 ISAT predictions based on Fall Scantron scores. Unpublished document.

Chicago Public Schools Office of Performance (2010c). No Child Left Behind (NCLB) accountability summary 2010: Dvorak Elem Specialty Academy (Unit 6760). Retrieved September 26, 2010 from https://research.cps.k12.il.us

Choi, S.W., Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. Retrieved May 30, 2010 from http://www.ncme.org

Corcoran, S. (2010). Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice.

Providence, RI: Annenberg Institute for School Reform at Brown University. Retrieved November 26, 2010 from http://www.annenberginstitute.org/Products/Corcoran.php

Dell'Angela, T. (2005). Chicago schools to ditch Iowa test. Retrieved on June 2, 2010 from http://articles.chicagotribune.com/2005-08-21/news/0508210001_1_standardized-tests-iowa-test-test-change

FundingUniverse (2010). Educational Testing Service. Retrieved September 25, 2010 from http://www.fundinguniverse.com/company-histories/Educational-Testing-Service-Company-History.html

Gorham, J. L. and Bontempo, B. D. (1996). Repeater patterns on NCLEX using CAT versus NCLEX using paper-and-pencil testing. Retrieved on May 28, 2010 from http://www.mountainmeasurement.com/pdfs/CATvsPP.pdf

Harris, R. (2010). "Helping principals, teachers the next wave of performance management." Catalyst notebook. Retrieved May 11, 2010 from http://www.catalyst-chicago.org/notebook/index.php/entry/533

Illinois Government News Network. (January 15, 2010). Governor Quinn signs legislation to improve teacher, principal evaluations and training. Press release. Retrieved November 25, 2010 from http://www.illinois.gov/PressReleases/ShowPressRelease.cfm?SubjectID=1&RecNum=8176

Illinois State Board of Education. (2009). Interpretive guide 2009 Illinois Standards Achievement Test. Retrieved November 7, 2010 from www.isbe.state.il.us/assessment/pdfs/ISAT_Interpr_Guide_2009.pdf

Jachino, R. (2010). 2010 Mathematics ISAT. Retrieved June 12, 2010 from http://www.isbe.state.il.us/assessment/pdfs/2010/math_isat.pdf

Kolen, M. (2000). Threats to score compatibility. *Educational assessment.* 6(2), 73-96.

Legg, S.M.; Buhr, D.C. (1990). Investigating differences in mean scores on adaptive and paper and pencil versions of the college level academic skills reading test. Retrieved on June 1, 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/ 0000019b/80/20/67/6f.pdf

Love, N. (Ed.). (2009). *Using data to improve learning for all*. Thousand Oaks, CA: Corwin Press.

Mead, A. and Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis." *Psychological bulletin*. 114(3), 449-458.

Meyers, J. (2009). "Huberman's performance SWAT team." Catalyst notebook. Retrieved May 11, 2010 from http://www.catalyst-chicago.org/notebook/index.php/entry/290/

New Teacher Center. (2009). Meaningful test prep: Teaching testing as a genre. Workshop handout. Chicago: Chicago New Teacher Center.

Northern Illinois University (2010). Interactive Illinois Report Card. Retrieved June 14, 2010 from http://iirc.niu.edu/Default.aspx

Olsen, J. (1986). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986). Abstract retrieved June 6, 2010 from http://eric.ed.gov/ERICWebPortal/custom/ portlets/recordDetails/detailmini.jsp?_nfpb=true&_& ERICExtSearch_SearchValue_0=ED274714&ERICExtSearch_SearchType_0=no&accno

=ED274714

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. Retrieved on April 25, 2010 from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1047&context=jtla

Scantron Corporation (2010a). *Scantron Performance Series technical report*. Retrieved May 11, 2010 from http://docs.achievementseries.com/docsPerformance/ PerformanceSeriesTechnicalReport.pdf

Scantron Corporation (2010b). *Scantron Performance Series*. Retrieved November 13, 2010 from https://admin.edperformance.com/home/home.ssp

Schaefer, G.A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M.T., Steffen, M. (1998). Comparability of paper-and-pencil and computer adaptive test scores on the GRE general test. Retrieved on June 5, 2010 from http://www.ets.org/Media/Research/pdf/RR-98-38-Schaeffer.pdf

Schochet, P. Z. and Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. United States Department of Education. Retrieved October 10, 2010 from http://ies.ed.gov/ncee/pubs/20104004/pdf/ 20104004.pdf

Wang, S. (2004). Online or paper: does delivery affect results? Retrieved June 1, 2010 from http://www.pearsonassessments.com/NR/rdonlyres/D381C2EA-18A6-4B52-A5DC-DD0CEC3B0D40/0/OnlineorPaper.pdf

Wang, S., Jiao, H., Young, M., Brooks, T. and Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and psychological measurement,* *67*, 219 - 238.

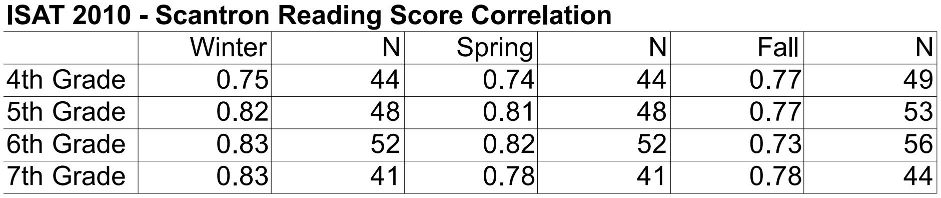
Wang, S., Jiao, H., Young, M., Brooks, T. and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and psychological measurement, 68*, 5 - 24.

Winerip, M. (1993). No. 2 pencil fades as graduate exam moves to computer. *New York Times*. November 15, 1993. Retrieved September 25, 2010 from http://www.nytimes.com/1993/11/15/us/no-2-pencil-fades-as-graduate-exam-moves-to-computer.html

Young, K. (2005). Standardized testing. Retrieved on September 25, 2010 from https://www.msu.edu/~youngka7/testing.html

Zytowski, D. (2008). From #2 pencils to the World Wide Web: A history of test scoring. *Journal of career assessment*. November 2008; vol. 16, 4: pp. 502-511.

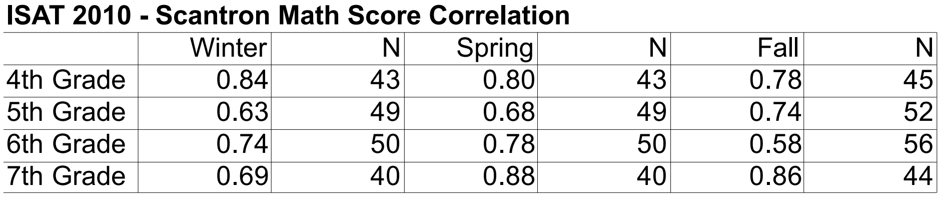
Appendix - Tables and Figures



ISAT 2010 - Scantron Reading Score Correlation

	Winter	N	Spring	N	Fall	N
4th Grade	0.75	44	0.74	44	0.77	49
5th Grade	0.82	48	0.81	48	0.77	53
6th Grade	0.83	52	0.82	52	0.73	56
7th Grade	0.83	41	0.78	41	0.78	44

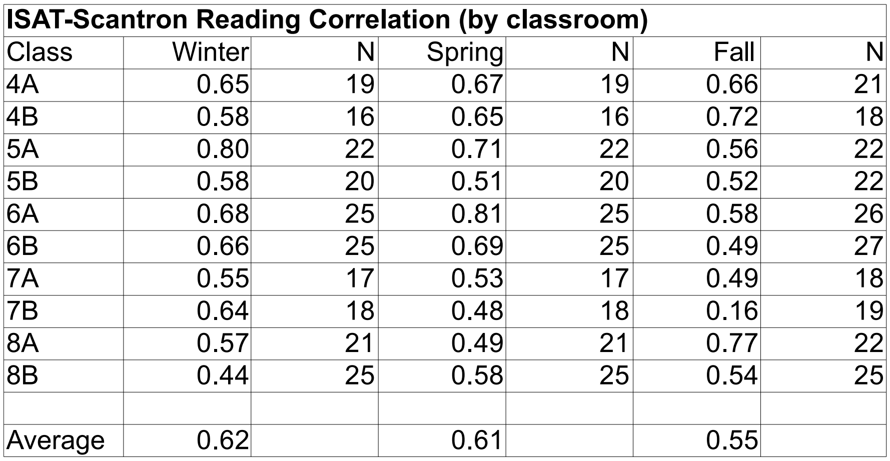
Table 1



ISAT 2010 - Scantron Math Score Correlation

	Winter	N	Spring	N	Fall	N
4th Grade	0.84	43	0.80	43	0.78	45
5th Grade	0.63	49	0.68	49	0.74	52
6th Grade	0.74	50	0.78	50	0.58	56
7th Grade	0.69	40	0.88	40	0.86	44

Table 2



ISAT-Scantron Reading Correlation (by classroom)

Class	Winter	N	Spring	N	Fall	N
4A	0.65	19	0.67	19	0.66	21
4B	0.58	16	0.65	16	0.72	18
5A	0.80	22	0.71	22	0.56	22
5B	0.58	20	0.51	20	0.52	22
6A	0.68	25	0.81	25	0.58	26
6B	0.66	25	0.69	25	0.49	27
7A	0.55	17	0.53	17	0.49	18
7B	0.64	18	0.48	18	0.16	19
8A	0.57	21	0.49	21	0.77	22
8B	0.44	25	0.58	25	0.54	25
Average	0.62		0.61		0.55	

Table 3

ISAT-Scantron Math Correlation (by classroom)						
	Winter	N	Spring	N	Fall	N
4A	0.69	20	0.81	20	0.60	21
4B	0.85	18	0.61	18	0.86	19
5A	0.30	21	0.42	21	0.67	22
5B	0.50	21	0.53	21	0.46	22
6A	0.69	25	0.80	25	0.32	26
6B	0.50	23	0.45	23	0.32	27
7A	0.61	18	0.81	18	0.84	18
7B	0.30	18	0.41	18	0.09	19
8A	0.48	20	0.27	20	0.48	22
8B	0.53	25	0.76	25	0.66	25
Average	0.55		0.59		0.53	

Table 4

Reading										
Grade	Class	N	Winter SS	Spring SS	Gain	∂ of Gain	Min Gain	Max Gain	Gain Range	Avg SEM Diff
03	3A	22	2223	2361	139	290	-566	573	1139	95
03	3B	28	1891	1971	81	147	-233	382	615	98
04	4A	24	2035	1989	-46	196	-423	399	822	93
04	4B	27	2477	2433	-44	203	-418	654	1072	91
05	5A	30	2339	2437	98	131	-357	366	723	90
05	5B	29	2591	2657	66	132	-257	434	691	92
06	6A	21	2738	2814	76	78	-68	310	378	93
06	6B	22	2465	2557	92	170	-311	335	646	89
07	7A	27	2887	2998	111	113	-97	365	462	95
07	7B	27	2556	2607	51	183	-426	427	853	89
08	8A	23	2669	2753	84	201	-407	358	765	89
08	8B	31	2911	2911	0	122	-252	312	564	90

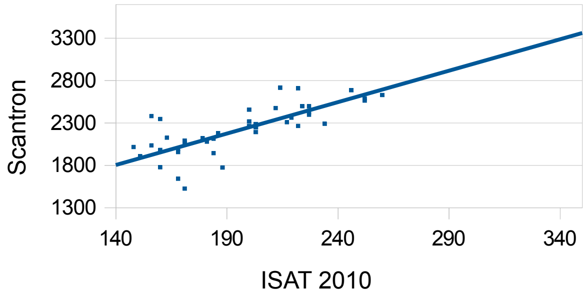
Table 5

Math

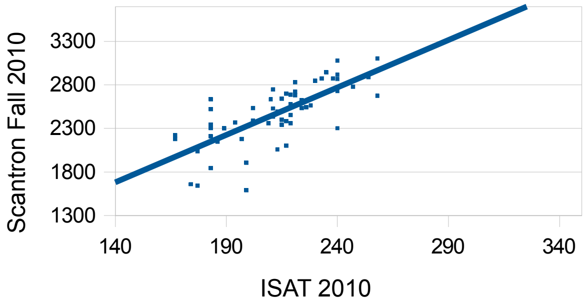
Grade	Class	N	Winter SS	Spring SS	Gain	StdDev of Gain	Min Gain	Max Gain	Gain Range	Avg SEM Diff
03	3A	22	2245	2347	102	97	-108	322	430	77
03	3B	29	2024	2143	120	85	-157	284	441	77
04	4A	23	2076	2139	63	130	-140	479	619	77
04	4B	28	2272	2341	69	85	-103	351	454	77
05	5A	28	2279	2309	29	104	-140	295	435	77
05	5B	29	2443	2513	70	111	-163	300	463	77
06	6A	21	2538	2687	149	109	-27	410	437	77
06	6B	23	2345	2393	48	99	-87	394	481	77
07	7A	26	2732	2773	41	100	-197	230	427	77
07	7B	26	2499	2528	30	128	-217	366	583	77
08	8A	25	2510	2512	2	160	-455	357	812	77
08	8B	31	2613	2774	161	133	-63	453	516	77

Table 6

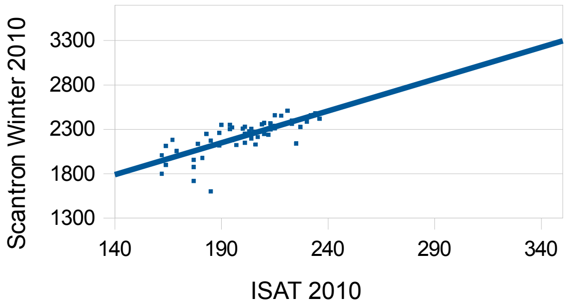
3rd - 4th ISAT/ Scantron Fall Reading



5th - 6th ISAT/Scantron Reading Fall



4th - 5th ISAT/Scantron Math Fall



5th - 6th ISAT/Scantron Math Fall

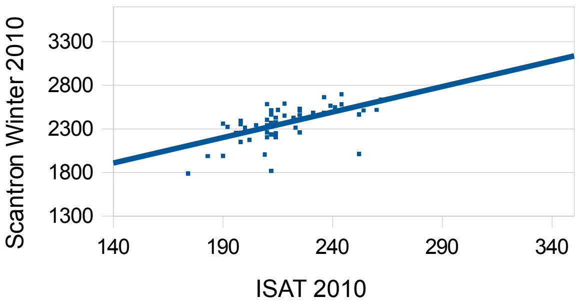


Figure 1

Fall, 2010	Top 40 ISAT NPR / Bottom 40 Scantron NPR	N	Percent	Also spoiled tests	Percent
Reading	39	255	15.3%	16	41.0%
Math	34	255	13.3%	12	35.3%

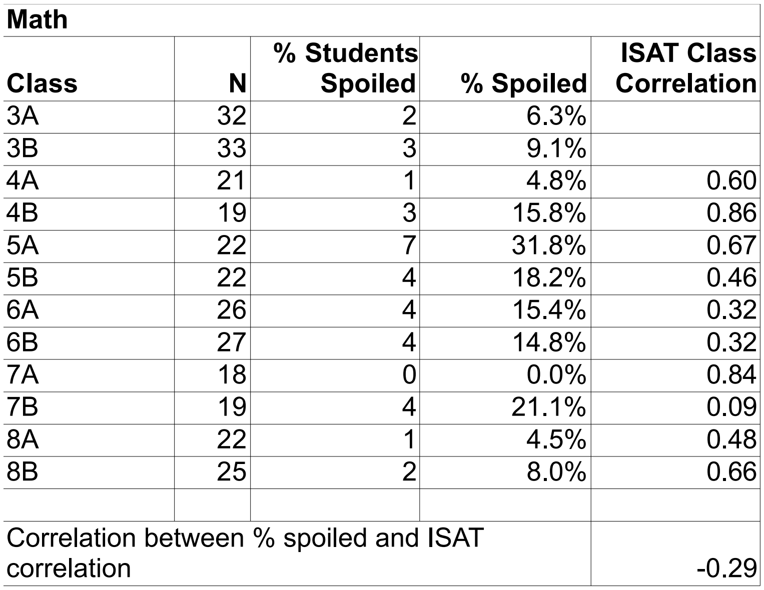
Table 7

Test spoilage due to testing irregularities - Fall, 2010		
	Reading	Math
Total students tested	343	342
Students with test irregularities	33	39
Percent of total students tested with testing irregularities	9.62%	11.40%
Spoiled more than once	17	21
Percent of total students tested with testing irregularities	4.96%	6.14%
Percent of students with testing irregularities that had more than one test spoiled due to testing irregularities	51.52%	53.85%

Table 8

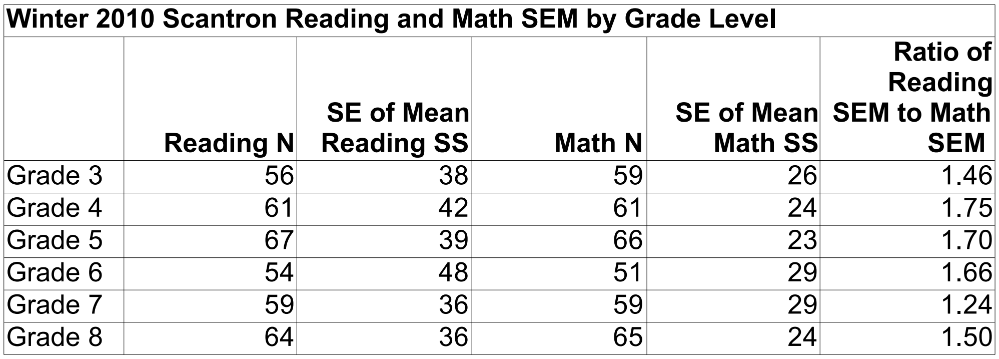
Reading				
Class	N	# Students Spoiled	% Spoiled	ISAT Class Correlation
3A	32	0	0.0%	
3B	33	7	21.2%	
4A	21	5	23.8%	0.66
4B	18	1	5.6%	0.72
5A	22	4	18.2%	0.56
5B	22	2	9.1%	0.52
6A	26	0	0.0%	0.58
6B	27	3	11.1%	0.49
7A	18	1	5.6%	0.49
7B	19	2	10.5%	0.16
8A	22	1	4.5%	0.77
8B	25	0	0.0%	0.54
Correlation between % spoiled and ISAT correlation				-0.05

Table 9



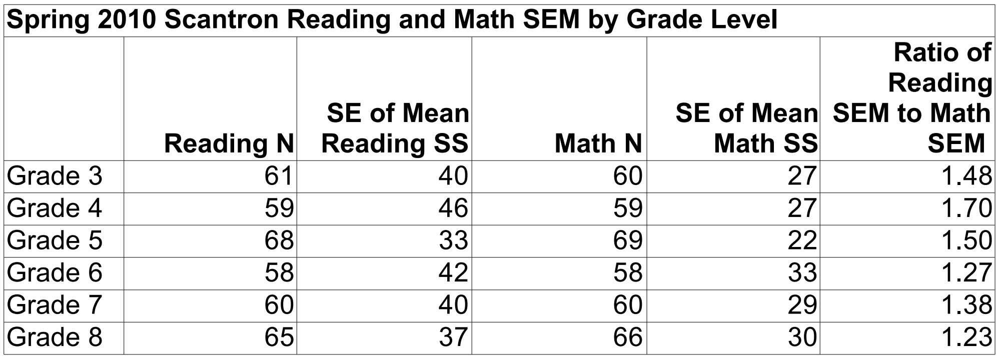
Math				
Class	N	% Students Spoiled	% Spoiled	ISAT Class Correlation
3A	32	2	6.3%	
3B	33	3	9.1%	
4A	21	1	4.8%	0.60
4B	19	3	15.8%	0.86
5A	22	7	31.8%	0.67
5B	22	4	18.2%	0.46
6A	26	4	15.4%	0.32
6B	27	4	14.8%	0.32
7A	18	0	0.0%	0.84
7B	19	4	21.1%	0.09
8A	22	1	4.5%	0.48
8B	25	2	8.0%	0.66
Correlation between % spoiled and ISAT correlation				-0.29

Table 10



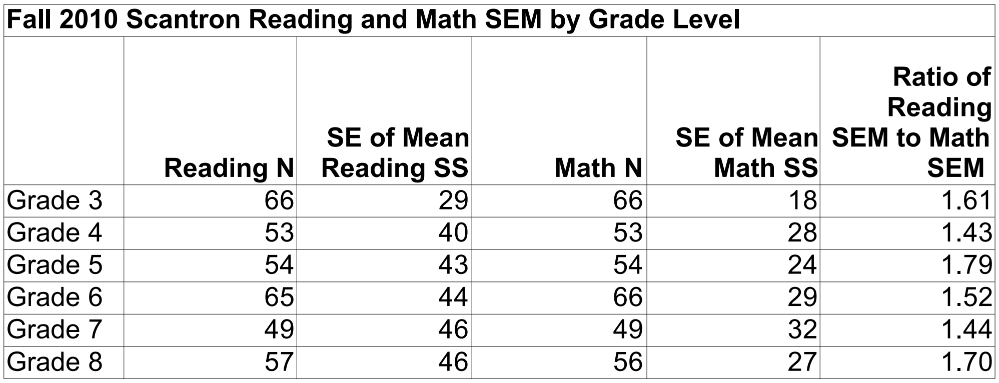
Winter 2010 Scantron Reading and Math SEM by Grade Level					
	Reading N	SE of Mean Reading SS	Math N	SE of Mean Math SS	Ratio of Reading SEM to Math SEM
Grade 3	56	38	59	26	1.46
Grade 4	61	42	61	24	1.75
Grade 5	67	39	66	23	1.70
Grade 6	54	48	51	29	1.66
Grade 7	59	36	59	29	1.24
Grade 8	64	36	65	24	1.50

Table 11



Spring 2010 Scantron Reading and Math SEM by Grade Level					
	Reading N	SE of Mean Reading SS	Math N	SE of Mean Math SS	Ratio of Reading SEM to Math SEM
Grade 3	61	40	60	27	1.48
Grade 4	59	46	59	27	1.70
Grade 5	68	33	69	22	1.50
Grade 6	58	42	58	33	1.27
Grade 7	60	40	60	29	1.38
Grade 8	65	37	66	30	1.23

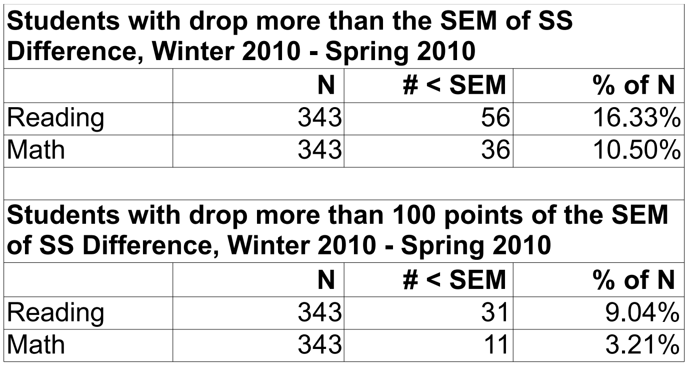
Table 12



Fall 2010 Scantron Reading and Math SEM by Grade Level

	Reading N	SE of Mean Reading SS	Math N	SE of Mean Math SS	Ratio of Reading SEM to Math SEM
Grade 3	66	29	66	18	1.61
Grade 4	53	40	53	28	1.43
Grade 5	54	43	54	24	1.79
Grade 6	65	44	66	29	1.52
Grade 7	49	46	49	32	1.44
Grade 8	57	46	56	27	1.70

Table 13



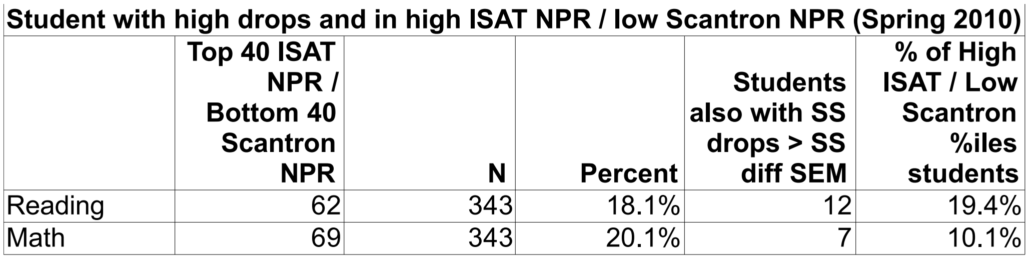
Students with drop more than the SEM of SS Difference, Winter 2010 - Spring 2010

	N	# < SEM	% of N
Reading	343	56	16.33%
Math	343	36	10.50%

Students with drop more than 100 points of the SEM of SS Difference, Winter 2010 - Spring 2010

	N	# < SEM	% of N
Reading	343	31	9.04%
Math	343	11	3.21%

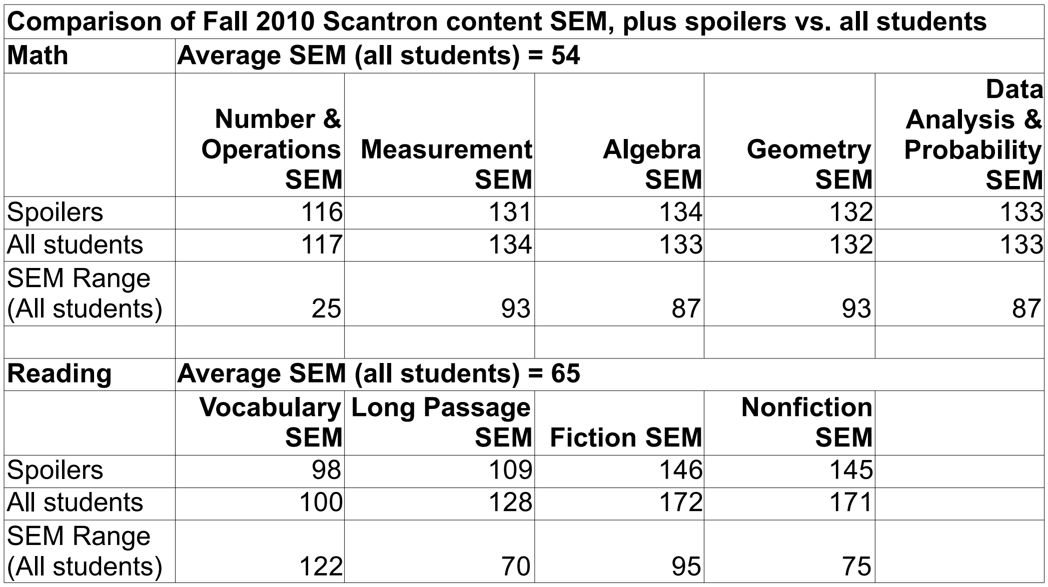
Table 14



Student with high drops and in high ISAT NPR / low Scantron NPR (Spring 2010)

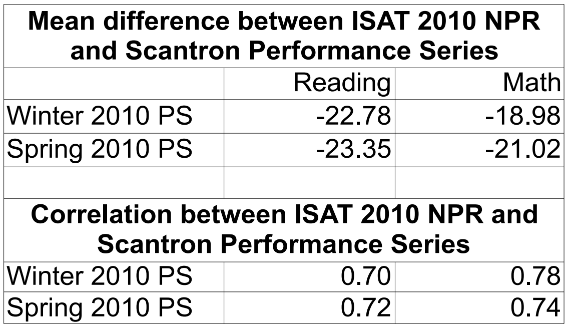
	Top 40 ISAT NPR / Bottom 40 Scantron NPR	N	Percent	Students also with SS drops > SS diff SEM	% of High ISAT / Low Scantron %iles students
Reading	62	343	18.1%	12	19.4%
Math	69	343	20.1%	7	10.1%

Table 15



Comparison of Fall 2010 Scantron content SEM, plus spoilers vs. all students					
Math	**Average SEM (all students) = 54**				
	Number & Operations SEM	**Measurement SEM**	**Algebra SEM**	**Geometry SEM**	**Data Analysis & Probability SEM**
Spoilers	116	131	134	132	133
All students	117	134	133	132	133
SEM Range (All students)	25	93	87	93	87
Reading	**Average SEM (all students) = 65**				
	Vocabulary SEM	**Long Passage SEM**	**Fiction SEM**	**Nonfiction SEM**	
Spoilers	98	109	146	145	
All students	100	128	172	171	
SEM Range (All students)	122	70	95	75	

Table 16



Mean difference between ISAT 2010 NPR and Scantron Performance Series		
	Reading	Math
Winter 2010 PS	-22.78	-18.98
Spring 2010 PS	-23.35	-21.02
Correlation between ISAT 2010 NPR and Scantron Performance Series		
Winter 2010 PS	0.70	0.78
Spring 2010 PS	0.72	0.74

Table 17