An Empirical Comparison of Achievement Level Estimates from Adaptive Tests and Paper-and-Pencil Tests

G. Gage Kingsbury
Northwest Evaluation Association

April, 2002

Presented to the American Educational Research Association annual meeting
New Orleans, LA

Since the early development of computerized adaptive testing, researchers have strongly suggested studies comparing performance of computerized adaptive tests (CAT) to their paper-and-pencil counterparts (Green, Bock, Humphreys, Linn, and Reckase, 1984). This is imperative in any situation in which an adaptive test is replacing an existing paper-and-pencil test, or in any situation in which paper tests and adaptive tests are being used concurrently. While many initial studies of adaptive testing (Weiss, 1978, 1983) and some recent large-scale adaptive testing applications (Sands, Waters, and McBride, 1997) have reported comparative results, additional empirical information is quite useful.

One of the first questions that must be asked of two tests that are going to be used interchangeably is whether or not they provide similar scores for similar test takers. Lord (1980) has suggested that if two tests are going to be used as parallel tests, it should be a matter of indifference to individuals which of the two tests they take. This is an unlikely goal when we are comparing an adaptive test to a paper test, since elements of the test such as availability, item difficulty, induced motivation, and need for computer familiarity may differ dramatically. However, before we use two tests interchangeably, we should be able to identify that scores from one test to the other will differ no more than expected due to the measurement characteristics of the tests. We should also be able to identify those characteristics of the tests and test takers which contribute to any differences that go beyond expected variability.

In the past decade, a variety of licensure tests, certification tests, and other tests designed for the adult population have started to be administered in an adaptive format. A substantial amount of information about the performance of these tests with adults is available. Over the past few years, various school districts and state departments of education have adopted adaptive testing as a primary measure of student achievement and growth. Very little information is available about the performance of young students with adaptive tests. This study compares scores of elementary school students taking paper and adaptive tests as part of their normal districtwide assessment within a public school system. It then identifies characteristics of the student which might make adaptive testing more (or less) appropriate.

Study Design

Students. Scores from 8560 tests taken by students enrolled in the fourth and fifth grade in the Meridian, Idaho public school system were used in the study. Meridian is a rapidly growing suburban community near Boise, and one of the largest school districts in Idaho. Each student took a paper-and-pencil in the spring of 2000, as a portion of the normal districtwide assessment. In the fall of 2000, these students were

administered either a second paper-and-pencil testing, or a CAT testing (depending on the school in which they were enrolled). 4883 scores were obtained from a second paper-and-pencil testing; while 3677 scores were obtained from CAT testing. Again, the assessment was a portion of the normal districtwide assessment program.

Tests. Paper-and-pencil scores came from tests that were 40 or 50 items in length. The tests used were Achievement Level Tests (ALT). In ALT testing, a student takes one level of a series of tests that vary systematically in difficulty. The specific level that a student took was determined either by past test performance or by the student's score on a short locator test. ALT functions very similarly two the second stage of a two-stage adaptive test.

The adaptive tests used were the Measures of Academic Progress (MAP). These tests were 35 to 50 items in length. MAP and ALT tests report scores on the same score scale. The MAP tests were designed to include the same goals in the same proportions as the paper-and-pencil ALT tests. The MAP tests drew from item pools of over 1200 items which had previously been calibrated to a common measurement scale using the Rasch model. In both tests, maximum-likelihood scoring was used.

Subject areas. Students took tests in Reading, Mathematics, and Language Usage. The number of students in each testing condition within each subject area is shown in Table 1. Only those students who received valid test scores in both testing seasons are include is Table 1 and in the analysis.

Table 1
Number of Students in each Subject Area and Testing Condition

	Testing Condition	
Subject Area	ALT/ALT	ALT/MAP
Language Usage	1627	1189
Mathematics	1578	1157
Reading	1678	1331

Analysis

The first analysis performed was an analysis of covariance. This analysis was designed to determine whether the scores observed in fall testing differed more than expected as a function of the fall testing mode. Spring test scores were used as a covariate to control for initial group differences.

While the ANCOVA is a necessary first step, it is at least as important to study the bivariate distributions of student scores as they move from spring to fall, to identify whether the mode of testing used in the fall has a differential impact on student scores. Correlations of spring and fall scores were calculated for each fall

test modality. This is essentially a test-retest correlation for the students taking two paper-and-pencil tests. For the students taking a paper-and-pencil test in the spring and a CAT test in the fall, the correlation should approach the paper-and-pencil test-retest correlation, if the paper test and the adaptive tests are acting as essentially parallel instruments.

Finally, a more detailed analysis of the performance of students with the most discrepant score changes from spring to fall was conducted. In this analysis, all students with fall scores that varied from spring scores by more than one standard deviation (based on the spring standard deviation) were identified and their test results were subjected to inspection. This group of students included those who had surprising change moving from one paper test to another, as well as those with surprising changes moving from a paper test to the adaptive test. This investigation included identifying demographic characteristics of the students, examining the spring response patterns of the students, and comparing performance on specific items which appeared in the paper-and-pencil tests which also appeared with some frequency in the adaptive tests. The focus of the analysis was to identify particular characteristics that might cause student scores to be more variable when moving to an adaptive test from a paper test.

Results

Table 2 shows descriptive statistics for the scores obtained by students in each subject and testing season, for each test modality.

Table 2
Descriptive Statistics for Test Scores for Each Test Modality and Testing Season

Subject Area and Test Modality	Spr Mean	Spring SD	Fall Mean	Fall SD	Mean Growth
Language Usage					
ALT/ALT	208.34	12.57	208.57	11.49	0.23
ALT/MAP	209.12	12.67	210.14	11.02	1.01
Mathematics					
ALT/ALT	205.55	11.62	203.95	11.96	-1.59
ALT/MAP	206.86	12.13	206.87	12.24	0.02
Reading					
ALT/ALT	206.39	11.89	207.20	12.35	0.80
ALT/MAP	206.47	12.58	206.98	12.77	0.51

The differences in the scores observed for the two test modalities were consistently small. The largest observed mean difference was 1.5 scale score points (for Mathematics) and the other differences were less than a scale score point. This difference is less than the difference observed between the two groups of students during their spring testing, which was all done with paper-and-pencil.

The results of the analysis of covariance using spring scores as the covariate and test modality as the dependent variable are shown in Table 3, for each subject area. The analysis indicates a difference that was significant at the .01 level for Language Usage and Mathematics. This is not particularly surprising, since the large sample sizes result in a very powerful test. The more important information is that the impact of test modality on student performance, while significant in a statistical sense, is quite modest with respect to the standard deviation of test scores.

Table 3
ANCOVA Results for Impact of Test Type on Fall Test Scores
with Spring Test Scores as a Covariate

Subject Area	F-Value	P-Value
Language Usage	20.651	.000
Mathematics	59.697	.000
Reading	1.451	.229

Table 4 shows the correlations between spring and fall scores, for each test modality and subject area. Correlations between spring paper-and-pencil scores and fall paper-and-pencil scores range from .88 to .90. Correlations between spring paper-and-pencil scores and Fall CAT scores range from .83 to .85. The test-retest correlations observed in both cases are high. Those observed in changing test modes from paper-and-pencil to CAT are somewhat lower than those observed when the test mode remains constantly paper-and-pencil. Some of the variation in correlations may be due to the differing error structures of the two test forms. The MAP test tends to result in scores with slightly lower SEM values, particularly for students at the extremes of the achievement distribution. While this should result in high reliability for the MAP scores, the difference in error structures may actually reduce the correlation between the ALT and MAP scores.

Table 4
Correlations between Fall and Spring Test Scores for each Subject Area
and Testing Condition

	Testing Condition	
Subject Area	ALT/ALT	ALT/MAP
Language Usage	.90	.83
Mathematics	.89	.85
Reading	.88	.83

Given the ANCOVA and correlational findings, it seems prudent to examine the characteristics of the fall test scores for the two types of tests as a function of the spring test scores. The mean fall test scores for groupings of students based on spring test scores for each subject area and each test mode are shown in the Figures 1, 2, and 3. Grouping was done by combining all students with spring scores within a ten-point score band (175-185, 185-195, etc.). From these figures it can be seen that the two test modes in the fall have very similar relationships with spring scores. However, some students do seem to achieve

substantially higher scores on the Fall CAT test than we would anticipate from their spring score. This is seen clearly in Language Usage and Mathematics, and seems to be more prevalent for students with very low scores in the spring testing.

The analysis of students who had a score change of more than a standard deviation (approximately 15 scale points) from spring to fall indicated no relationship with student gender or ethnic identifier. Two sets of students were over represented in the group making surprising changes when taking a paper test in the spring and an adaptive test in the fall. The first group included students who had to take two paper tests in the spring because the first score was invalid. This group comprised 10.7% of those students making surprising change from ALT to MAP, but only 4.0% of the entire ALT to MAP sample. The second set included students who omitted more than 10 percent of the questions on the paper test in the spring. This group comprised 7.7% of the students with surprising change from ALT to MAP, while it comprised less than 1.0% of the entire sample. In both of these groups, the change for from spring to fall was positive without exception. Students with surprising gains who took a paper test in the spring and another paper test in the fall did not show overrepresentation of these groups.

Discussion

In general, this study provided some support for the use of the adaptive test alongside the paper-and-pencil test. As students went from one test in the spring to another in the fall, the difference in the second score that was attributable to the test modality was no greater than one-tenth of a standard deviation in terms of average scale scores. While this result was statistically significant for two subject areas, the actual impact on any student's score can be expected to be quite small.

While the two different test modes have very little effect on scale scores, it is interesting to look more closely at the students who did show more change than expected when moving from paper-and-pencil to CAT. These students include a very large percentage of students who had to take a retest in the spring testing. These students also include a high percentage of students who omitted many items on the spring test (but not enough to invalidate the score). It is possible that these students were helped by the characteristics of the CAT system. A student who was misplaced in the paper-and-pencil system might encounter frustrating items and lose focus. The adaptive nature of the CAT system eliminates this possibility. In addition, omissions are not allowed in this adaptive testing system. As a result, a conservative student who might omit rather than guess in a paper-and-pencil test must use partial information to give an answer in CAT. In both conditions, a more accurate score from the CAT system may result.

Young students may benefit from an adaptive test for any number of reasons. They may be more motivated by the adaptive format. They may be more focused because only one item is presented at a time. They may have less difficulty responding because they do not have to deal with an answer sheet. Any of these characteristics may make it more difficult to create adaptive and paper-and-pencil tests that result in exactly comparable scores. While the analysis in this study indicated substantial correspondence between these tests, it also indicated elements of an adaptive testing program that might cause scores to diverge from those obtained for a paper-and-pencil test. Consistent research related to these issues should allow us to clarify and control those elements.

References

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests, Journal of Educational Measurement, 21, 347-360.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). Computerized Adaptive Testing : From Inquiry to Operation. Washington, DC: American Psychological Association.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Weiss, D. J. (Ed.) (1978). Proceedings of the 1977 computerized adaptive testing conference. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (Ed.) (1983).New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Zara, A. R. (June, 1989). A research proposal for field testing CAT for nursing licensure examinations. In Delegate Assembly Book of Reports 1989. Chicago: National Council of State Boards of Nursing, Inc.

Figure 1
Relationship between Spring and Fall Language Usage Scores

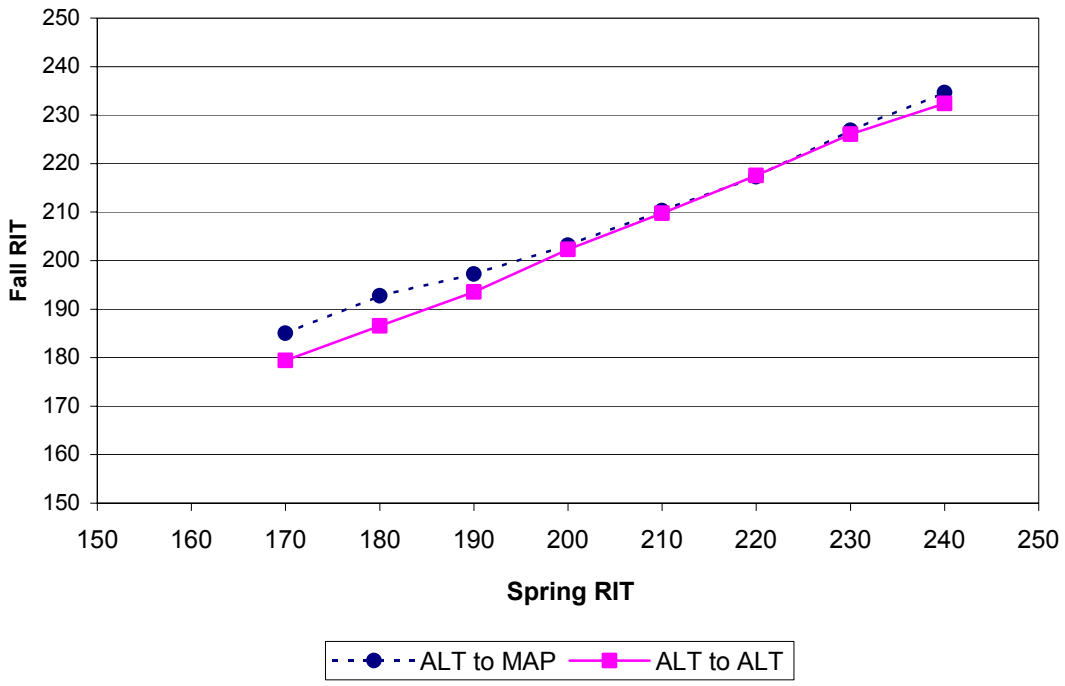

Figure 2
Relationship between Spring and Fall Mathematics Scores

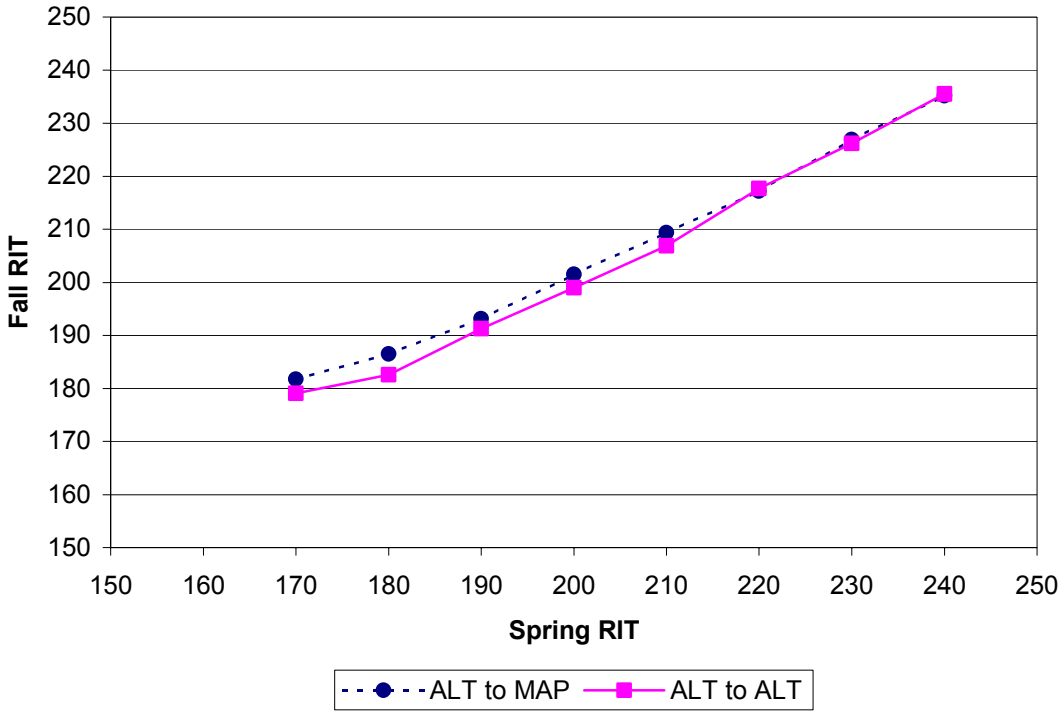

Figure 3
Relationship between Spring and Fall Reading Scores