

**Symposium of
Administration Mode Effects in Computer-Based
Large-Scale Assessments**

**A Comparison Study of Testing Mode Using
Multiple-Choice and Constructed-Response Items
-- Lessons Learned from a Pilot Study**

Liru Zhang
Delaware Department of Education

C. Allen Lau
Harcourt Assessment, Inc.

Paper presented at the AERA Annual Conference
April 7-11, 2006, San Francisco, CA

A Comparison Study of Testing Mode Using Multiple-Choice and Constructed-Response Items -- Lessons Learned from a Pilot Study

Background

For over a decade, the information technology and widespread availability of computers have significant impacts on curriculum, instruction and student learning in education. The advancements of new technology provide the measurement community with considerable potentials in test development and an alternate test delivery, named online or computer-based test (CBT). The advantages of CBT over traditional paper-and-pencil test (PPT), such as immediate scoring and reporting, test security, flexible test administration schedules, using multimedia item types, and new generation of accommodations for students with special needs, have been recognized for large-scale assessments (Bennett, 2001, 2002; Boo & Vispoel, 1998; Folk & Smith, 1998; Parshall, Spray, Kalohn, & Devey, 2002; Schmit & Ryan, 1993; Klein & Hamilton, 1999). A 2003 survey from 46 states and 6 U.S. jurisdictions by the Council of Chief State School Officers (CCSSO) indicated that 80% of the states/jurisdictions were either developing or piloting online testing in one or more applications or interested in exploring the possibilities.

Under the requirements of *No Child Left Behind Act* (NCLB), the high-stakes, statewide assessment program should provide valid and reliable measures of students progress toward to the state content standards. If different versions of the test within a content area of a grade level is employed, (e.g., computer-based and paper/pencil versions of the same test), the state must demonstrate the comparability of test scores (Guidelines of Peer Review, 2005). According to the American Psychological Association Guidelines (1986), scores from the CBT and PPT versions “may be considered equivalent when (1) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (2) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (p.18). The Guidelines also indicate the importance of eliminating irrelevant influences on test scores such as computer anxiety and computer experience. The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) suggest that direct evidence of score equivalence should be provided.

Published and unpublished research studies have been conducted in recent years to explore the comparability of test scores between administration modes of computer-based and paper/pencil versions of a test in various content areas, such as Reading or Language, Mathematics, and Writing. Results from reported studies, however, vary from study to study depending on the research design and methodology, content area, grade level, sampling procedures, item format, scoring, and technology device and computer system involved.

Russell, et al (2000) reported that student performance was significantly better on the computer for both Language Arts Composition and open-ended items of the

Massachusetts composition test across grades 4, 8, and 10. Students receiving Special Education services were benefited even more by taking the written test on the computer. An unpublished study using a volunteer sample of students and a computer automated scoring system to investigate the effects of administration mode for a state online Writing test (2006) found statistically significant higher scores on the paper/pencil version than on the computer-based version for both grades 5 and 9. In Oregon, a study (Choi and Tinkler, 2002) focused on score comparability between PPT and CBT in Mathematics and Reading for grades 3 and 10. They created two roughly equivalent parts of the existing test form, one for paper/pencil and one for computer delivery; and used a separate test block as anchor that was presented on computer only for calibration. Classrooms within each school were assigned into groups randomly with a counter-balanced design for test taking to eliminate order effect. The findings show that, in sum, the estimated mean item difficulty was larger for grade 3 than for grade 10 and larger for Reading than for Mathematics. Differential patterns of item difficulty were generally harder on the CBT, particularly in Reading for grade 3. Additional analysis was conducted for Reading by textual categories. In 2004, Harcourt Assessment, Inc. conducted a comparability study (Wang, et al) for Stanford Diagnostic Reading and Mathematics tests. Over 3,500 students from grade 2 to grade 12 participated in six levels. The study employed the split plot repeated measure design with unequal group size to examine the linear relationship, rank order, frequency distribution, and mean of test scores between administration modes. This study provided empirical evidence to support the comparability of test scores across administration mode in both content areas. Differences in scores between modes did not exceed random error for most sub-tests. "For the few exceptions where there were significant differences in mean scores, the test results obtained from the different administration modes would be equated to allow for equivalent score interpretations...." (p.12-13) This report also pointed out that the larger different mean scores for grade 2 may be due to unfamiliarity with computer use and relatively smaller sample. Court (2005) reported consistently higher average scores on PPT than on CBT for grades 4 and 7 Mathematics and grades 5 and 8 Reading of a statewide assessment program. The author suggested that test scores derived from PPT and CBT versions were not interchangeable without an equating procedure, and, therefore, the consequences could be substantial on individual students, schools, and school districts.

To investigate the equivalence of PPT and CBT tests, Ewing, et al (2003) used structural equation modeling to examine the degree of structural invariance across mode for administering College Algebra and English Composition without essay. For the algebra test, support for partial structure invariance was found across mode for African American, Asian, and Hispanic students. For the composition test, however, the "goodness-of-fit indices did not support the plausibility of the three-group congeneric model for the total or subgroups" (p.15). Using DIF approach, Schwars, et al (2003) explored the cross-mode comparability of test scores on InView, a norm-referenced aptitude test, at the item level. Without providing a general conclusion, the authors suggested "one expectation is that many items should have been flagged on an inspection of the mean differences between modes of administration These differences in ability distributions do not necessarily give rise to DIF since statistics are conditional on ability" (p.8).

Purposes of Study

The Delaware Student Testing Program (DSTP) is a standards-based, mandated assessment system. The DSTP scores in Reading and Mathematics have been used as the primary indicators for high-stakes accountability decisions. Summer school, Individual Improvement Plans (IIP), and/or retests are required for students of grades 3, 5, and 8 whose performance falling into the categories of Below the Standard and Well Below the Standards in Reading; and students of grade 8 whose performance falling into the same two categories in Mathematics.

By the end of summer, decisions on student promotion and/or the IIP for instructional needs for the next school year largely depend on the results of retesting. Due to the immediate demand of retest scores from schools, the Department of Education was considered providing students with an option of online delivery administration. The feasibility of a CBT version for retest and the comparability of test scores across administration mode were the primary objectives of the pilot study. Since the DSTP employs short-answer (SA) and open-ended (OE) formats in addition to multiple-choice (MC) items, the quick turnaround of test results heavily depending on the scoring process. In winter 2002, a pilot study was conducted to collect empirical evidence to explore the feasibility of using online administration for summer retest in order to provide schools and school districts with an immediate turnaround of test results for the pressing decisions. The study focused on:

- (1) The comparability of test scores between the paper/pencil and computer-based test administration modes; and
- (2) The consistency between human-scoring and computer-automated scoring on short-answer and open-ended items.

Design and Methodology

Assessment Instruments: The test used in the pilot study was one of the operational test forms, including the field test items, in Reading for grades 5 and 8 and in Mathematics for grade 8. Each test form consists of two components: the abbreviated version of Stanford Achievement Test, 9th edition (SAT9) Reading Comprehension or Mathematics Problem Solving and Delaware-developed items. Table 1 shows number of items and maximum score points by test and item format. Even though students took all the 30 SAT9 items during test administration for reporting the norm-referenced test scores, only selected items that measure the corresponding Delaware Content Standards were used for equating each year and counted for reporting the standards-based scores (scale scores). Since SAT9 was taken in the same paper/pencil mode as the common items, Delaware-developed items were ‘converted’ to the computer-based version with minor modifications for a few Mathematics items (e.g., layout, context) due to the software limitations.

Table 1. Number of Items and Maximum Points by Test and Grade

Category	Reading				Mathematics	
	Grade 5		Grade 8		Grade 8	
	<i>N. of Items</i>	<i>Maximum Points</i>	<i>N. of Items</i>	<i>Maximum Points</i>	<i>N. of Items</i>	<i>Maximum Points</i>
Item Format						
Multiple-Choice	55	55	55	55	52	52
Short Answer	8	16	8	16	8	16
Open-Ended	4	16	4	16	3	12
Total	67	87	67	87	63	80
DSTP Component						
SAT9	30	30	30	30	30	30
Selected SAT9 Items	24	24	27	27	28	28
DE-Developed	37	57	37	57	33	50
Total	61	81	64	84	61	78

Sample of Subjects: The subjects used in the study were volunteer participants. Eligible subjects were students who were enrolled in grades 5 (retention) or 6; or grade 8 (retention) or 9 of the fall 2002. Those students either had taken or should have taken the grade 5 Reading or grade 8 Reading and/or grade 8 Mathematics test(s) in spring or summer of 2002 (using different test forms), regardless of their performance levels achieved. School districts or schools assigned their participants into the PPT or the CBT group with some consultation with students. Each school assigned about one half of their participants into the PPT and one half into the CBT group with some consultation with students. Over 1,600 students participated in the pilot study, 570 for grade 5 Reading, 330 for grade 8 Reading, and 801 for grade 8 Mathematics. Approximately one half of the students were assigned to the PPT Group and one half to the CBT Group (Table 2).

Students were notified prior to the test that there would be no negative impact on any participants. The state would recognize a higher individual score on the pilot study than his/her score on the 2002 spring test administration or the summer retest and use the higher score for student accountability.

Table 2. Sample of Subjects by Test, Grade and Student Background

Student Background	Grade 5 Reading				Grade 8 Reading				Grade 8 Math			
	PPT		CBT		PPT		CBT		PPT		CBT	
	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>
Gender												
Female	117	42	119	43	83	48	66	42	214	52	218	57
Male	161	58	157	57	91	52	90	58	200	48	168	44
Racial/Ethnic Group												
American Indian	1	0	0	0	0	0	1	1	0	0	1	0
Afri. American	122	44	124	45	87	50	66	42	167	40	149	39
Asian	3	1	3	1	0	0	2	1	0	0	2	1
Hispanic	36	13	26	9	16	9	10	6	28	7	22	6
Caucasian	116	42	123	45	71	41	77	49	219	53	212	55
Student with Disability												
No	192	69	216	78	129	74	121	78	341	82	330	86
Yes	86	31	60	22	45	26	35	22	73	18	56	15
Eligible for Free Lunch												
No	108	39	124	45	79	45	79	51	243	59	227	59
Yes	170	61	152	55	95	55	77	49	171	41	159	41
Total*	293	51	277	49	174	53	156	47	414	52	387	48

* The discrepancy of number of subjects is due to missing data.

Test Administration and Training: The test was given under the same conditions as regular DSTP administration. The majority of accommodations were available except those accommodations that required additional materials, such as Braille, large-print, audio tape, and the Spanish version of the Mathematics test. A questionnaire was given to all students about their access to a home computer and experiences in using computers (Appendix). To collect additional information regarding the feasibility of CBT, observations and incident reports for the CBT administration were also planned.

A training session was provided for test administrators, Technology Coordinators, and Test Coordinators of schools and school districts. The Online Tutorial became available a week before the test to provide for students of the CBT group to be comfortable and familiar with the online testing environment. Sample items presented in the Online Tutorial also gave school technology coordinators the opportunity to anticipate and resolve potential issues and problems that they might encounter during testing.

The CBT version was delivered to each school via a secured web browser developed by a subcontractor. The secure browser prevented students from using features, such as e-mail, instant messaging, spell-check, and viewing other web sites while taking the test. Each student could reach the assigned test using a password and special ID.

Research Design and Data Analyses: This study employed the common-item, common-testing-mode, non-equivalent groups design. To equate the two administration modes, the same SAT9 items were administered in the same paper/pencil format for both groups while the Delaware-developed items were delivered in the paper/pencil format for the PPT Group and by web browser for the CBT Group.

- (1) **Equating Procedure:** Separate calibration for CBT and PPT was conducted, each with SAT9 items, using the Rasch Partial Credit Model (PCM) by BIGSTEPS. Each CBT version was equated to the corresponding PPT version with the SAT9 items as common part. The linking constant, the mean difference of the average item difficulty of the CBT Group from the average item difficulty of the PPT Group on the common SAT9 part, would place the calibrations from the CBT version on the same scale as the calibrations from the PPT version. Since the two modes of the test were equated, the average administration mode effect could be computed. The mean difference of the average item difficulty between equated CBT calibrations and the PPT calibrations on the Delaware-developed items represented the mode effect.
- (2) **Raw Score to Scaled Score Conversion:** A conversion table was developed for the CBT test administration mode in this pilot study. Based on the design, the Delaware-developed items were administered in the CBT format while the SAT9 items administered in the paper/pencil format. The magnitude of the mode effect was not added to the SAT9 item calibrations by running the BIGCR program to develop the raw score to scale score conversion tables.
- (3) **Assumptions of the Design:** The primary purpose of the pilot study was to examine the comparability of test scores across test administration mode. If the difference between the average paper/pencil item calibrations and the average computer-based item calibrations was zero, there is no mode effect and the test scores were comparable. If the difference did not equate to zero, the average administration mode could be determined either positive or negative. In this case, a new conversion table would be needed.
- (4) **Consistency of Scoring:** The consistency of human-scoring and computer-automated scoring was examined in this study for short-answer items (0-2 points) and open-ended items (0-4 points). Data analyses included summary descriptive statistics, cross-tabulations by item, percent agreement, Cohen's Kappa coefficient, and Pearson product-moment correlation. Kappa was computed by collapsing the matrix to be symmetrical.

Results and Discussion

Descriptive Statistics: The characteristics of the student sample is presented in Table 2, including number and percent of students by grade, test, and background information (e.g., gender, students with disability). Data suggest that more minority students, students with disabilities, and students from low-income families might be more comfortable with the traditional paper/pencil format than their counterparts, particularly in Reading. For instance, 9% more students with disabilities were in the PPT Group than in the CBT Group for grade 5 Reading; 8% more African American students were found in the PPT Group than in the CBT Group for grade 8 Reading.

Means and standard deviations of raw scores and scale scores for SAT9, Delaware-developed items, and the total test are presented in Tables 3a to 3c for each test by the administration mode. The reliability of test scores for each components of the test ranges from .81 to .90 for grade 5 Reading, from .76 to .89 for grade 8 Reading, and from .66 to .82 for grade 8 Mathematics. The consistency of the average scale scores between the two modes with the means of the common items from SAT9 provided a reasonable and essential condition for the equating procedure using the common-item, common-testing-mode, and non-equivalent group design for all three grades. In Reading, the mean scale score is 444.4 for CBT and 437.2 for PPT; while the SAT9 scaled scores are 643.0 and 635.5, respectively, for grade 5; the mean scale score is 480.1 for CBT and 436.3 for PPT; while the SAT9 scaled scores are 668.5 and 665.0, respectively, for grade 8. In Mathematics, the mean scale score is 472.1 for CBT and 470.8 for PPT; while the SAT9 scaled scores are 659.9 and 658.3, respectively.

Table 3a. Descriptive Statistics for Grade 5 Reading by Administration Mode

Score	PPT			CBT		
	Mean	S.D.	Alpha (SEM)	Mean	S.D.	Alpha (SEM)
SAT9						
<i>Raw Score</i>	15.9	5.4	0.81 (2.4)	16.9	5.7	0.83 (2.4)
<i>Scaled Score</i>	635.5	35.6		643.0	40.2	
Delaware						
<i>Raw Score</i>	26.7	10.7	0.88 (3.7)	25.6	9.6	0.86 (2.7)
Total						
<i>Raw Score</i>	39.8	14.0	0.90 (4.4)	39.5	14.1	0.90 (4.1)
<i>Scale Score</i>	437.2	38.1		444.4	34.2	

Table 3b. Descriptive Statistics for Grade 8 Reading by Administration Mode

Score	PPT			CBT		
	Mean	S.D.	Alpha (SEM)	Mean	S.D.	Alpha (SEM)
SAT9						
<i>Raw Score</i>	16.7	5.1	0.78 (2.4)	17.4	4.9	0.76 (2.4)
<i>Scaled Score</i>	665.0	30.1		668.5	28.4	
Delaware						
<i>Raw Score</i>	22.2	9.4	0.86 (3.5)	22.1	9.0	0.84 (3.6)
Total						
<i>Raw Score</i>	37.9	12.5	0.89 (4.1)	38.5	12.0	0.87 (4.3)
<i>Scale Score</i>	476.3	34		480.1	31.8	

Table 3c. Descriptive Statistics for Grade 8 Mathematics by Administration Mode

Score	PPT			CBT		
	Mean	S.D.	Alpha (SEM)	Mean	S.D.	Alpha (SEM)
SAT9						
<i>Raw Score</i>	13	4.3	0.67 (2.5)	13.3	4.2	0.66 (2.4)
<i>Scaled Score</i>	658.3	45.1		659.9	23.7	
Delaware						
<i>Raw Score</i>	17.2	7.0	0.76 (3.4)	14.5	6.0	0.71 (3.2)
Total						
<i>Raw Score</i>	30.3	10.2	0.82 (4.3)	27.9	8.8	0.79 (4.0)
<i>Scale Score</i>	470.8	26.5		472.1	24.3	

Table 4 shows the percentage of students in the performance levels for each grade by administration mode. It is obvious that the majority of participants (59-78%) were low-achieving students (Below or Well below the Standard) regardless of the administration mode. Although the frequency distributions of performance levels are similar across modes, the percent of students Below the Standard slightly higher by 3-5% for the PPT Group; while the percent of students Meets the Standard slightly higher by 2-5% for the CBT Group. To compare the changes of performance levels from the spring administration to the pilot study between the PPT and the CBT Groups, a matched sample was created. About 94% to 100% of the participants with a valid score on both administrations were included in the matched sample for each grade. The results of the cross-tab comparison (Tables 5a to 5c) suggest a similar performance pattern between the PPT and CBT Groups across administrations with slight discrepancies. In grade 5

Reading, for example, 27-28% of the students who had performed Below the Standard in spring received a lower performance level of Well below the Standard in the pilot study across modes, 38-39% of them remained the same performance level of Below the Standard, and 32-34% of them received a higher performance level of Meets the Standard. In grade 8 Reading, the majority of the students in Well below the Standard in spring remained the same performance level in the pilot, 65% in PPT and 57% in CBT. Eight percent more students in CBT than in PPT performed better in the pilot study.

Table 4. Percent of Students in Performance Level by Test and Administration Mode

Performance Level	Grade 5 Reading		Grade 8 Reading		Grade 8 Mathematics	
	PPT	CBT	PPT	CBT	PPT	CBT
<i>Well below the Standard</i>	36	31	44	41	44	39
<i>Below the Standard</i>	31	28	30	31	34	38
<i>Meets the Standard</i>	30	35	26	28	21	23
<i>Exceeds the Standard</i>	1	5	0	0	1	0
<i>Distinguished</i>	3	1	0	0	0	0

Table 5a. A Cross-Tab Comparison of Performance Level for Grade 5 Reading

Mode	Pilot Level	2002 Spring Administration									
		<i>Well Below</i>		<i>Below</i>		<i>Meets</i>		<i>Exceeds</i>		<i>Distinguished</i>	
		N.	%	N.	%	N.	%	N.	%	N.	%
PPT	<i>Well below</i>	58	60	38	28	2	7	0	0	0	0
	<i>Below</i>	26	27	53	39	4	14	1	13	0	0
	<i>Meets</i>	12	13	43	32	21	75	4	50	2	29
	<i>Exceeds</i>	0	0	0	0	1	4	2	25	0	0
	<i>Distinguished</i>	0	0	1	1	0	0	1	13	5	71
	Total		96	100	135	100	28	100	8	100	7
CBT	<i>Well below</i>	44	65	41	27	1	3	0	0	0	0
	<i>Below</i>	14	21	58	38	4	14	0	0	0	0
	<i>Meets</i>	10	15	51	34	21	72	8	62	3	38
	<i>Exceeds</i>	0	0	1	1	3	10	5	38	3	38
	<i>Distinguished</i>	0	0	0	0	0	0	0	0	2	25
	Total		68	100	151	100	29	100	13	100	8

Table 5b. A Cross-Tab Comparison of Performance Level for Grade 8 Reading

Mode	Pilot Level	2002 Spring Administration									
		<i>Well Below</i>		<i>Below</i>		<i>Meets</i>		<i>Exceeds</i>		<i>Distinguished</i>	
		N.	%	N.	%	N.	%	N.	%	N.	%
PPT	<i>Well below</i>	47	65	26	28	3	33	0	0	0	0
	<i>Below</i>	17	24	33	35	2	22	0	0	0	0
	<i>Meets</i>	8	11	34	37	4	44	0	0	0	0
	<i>Exceeds</i>	0	0	0	0	0	0	0	0	0	0
	<i>Distinguished</i>	0	0	0	0	0	0	0	0	0	0
	Total	72	100	93	100	9	100	0	0	0	0
CBT	<i>Well below</i>	32	57	25	30	5	50	0	0	0	0
	<i>Below</i>	17	30	27	32	3	30	0	0	0	0
	<i>Meets</i>	7	13	32	38	2	20	0	0	0	0
	<i>Exceeds</i>	0	0	0	0	0	0	0	0	0	0
	<i>Distinguished</i>	0	0	0	0	0	0	0	0	0	0
	Total	56	100	84	100	10	100	0	0	0	0

Table 5c. A Cross-Tab Comparison of Performance Level for Grade 8 Mathematics

Mode	Pilot Level	2002 Spring Administration									
		<i>Well Below</i>		<i>Below</i>		<i>Meets</i>		<i>Exceeds</i>		<i>Distinguished</i>	
		N.	%	N.	%	N.	%	N.	%	N.	%
PPT	<i>Well below</i>	141	62	39	22	0	0	0	0	0	0
	<i>Below</i>	69	31	72	40	0	0	0	0	0	0
	<i>Meets</i>	15	7	67	37	1	100	0	0	0	0
	<i>Exceeds</i>	1	0	1	1	0	0	0	0	0	0
	<i>Distinguished</i>	0	0	1	1	0	0	0	0	0	0
	Total	226	100	180	100	1	100	0	0	0	0
CBT	<i>Well below</i>	118	62	32	17	0	0	0	0	0	0
	<i>Below</i>	59	31	85	45	0	0	0	0	0	0
	<i>Meets</i>	14	7	69	37	4	100	0	0	0	0
	<i>Exceeds</i>	0	0	1	1	0	0	0	0	0	0
	<i>Distinguished</i>	0	0	0	0	0	0	0	0	0	0
	Total	191	100	187	100	4	100	0	0	0	0

CBT Effects: As discussed earlier, the average CBT effect could be computed through equating procedure for the two modes of the test. The mean difference of the average item difficulty between equated CBT calibrations and the PPT calibrations on the Delaware-developed items represented the mode effect. A positive difference indicates that the CBT items are harder than the PPT items; a negative difference indicates that the PPT items are harder than CBT items. The number of steps, means, and standard deviations of the Rasch step values for the equated CBT items and for the PPT items are summarized in Table 6 by item format for each test. The mean difference or the CBT effect was treated as a linking constant for all test items, including SAT9 items, for adjustment. The average changes of scale scores were calculated by using the mean difference times 40, the multiplicative constant for scaling.

For grade 5 Reading, the average change of scale scores is 11.6 with the CBT effect of .29. The difference of 11.6 scale score points approximately equals to the change of 4.5 raw score points in the middle of the raw score distributions. For grade 8 Reading, the average change of scale scores is 2.8 with the CBT effect of .07. The change of 2.8 scale score points is around to 1 raw score in the middle of the raw score distributions. For grade 8 Mathematics, the average change of scale scores is 12.8 with the CBT effect of .32. The scale score change represents a difference of almost 6 raw score points in the middle of the frequency distributions.

In this study, the CBT effects were also examined by item format across the three tests. First of all, a positive CBT effect was found for most comparisons with a few exceptions. A negative CBT effect was found for MC items in grade 5 Reading (-.16) and for SA items (-.24) in grade 8 Reading. Secondly, the results suggested an inconsistent pattern in terms of the effect size and directions by item format across tests. For grade 5 Reading, a large, positive CBT effect was shown for short answer (.59) and open-ended items (.76) as well as for all constructed-response items (.66), a negative effect was found for multiple-choice items (-.16). In grade 8, the CBT effect (.07) is the same for all Reading items, for MC items, and for all constructed-response items. A large CBT effect was found (.41) for open-ended items, however, a negative effect was for short answer items (-.24). The results of grade 8 Mathematics suggest a large effect for short answer (.74) and all constructed-response items (.50), but a relatively smaller effect for multiple-choice (.11) and open-ended items (.10).

Although the results did not support a consistent pattern by item formats, a larger CBT effect was observed for constructed-response items (including short-answer and open-ended items) than for multiple-choice items in most cases of this study. For instance, the average difference of scale scores based on all constructed-response items is 26 points for grade 5 Reading and 20 points for grade 8 Mathematics, which represent the changes of 9-10 raw score points, respectively, from the corresponding raw score distributions.

Conversion Tables for CBT: Due the CBT effects, a new conversion table from raw scores to scale scores was developed for each test. Figures 1 to 3 compare the raw scores to scale scores conversions for the three tests between the computer-based version and the paper/pencil version.

Table 6. Statistics of Rasch Step Values for Equated Items

GR 5 Reading	All Items			MC Items			All OE Items			SA Items			ECR Items		
	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff
<i>N of Steps</i>	56	57		25	25		31	32		16	16		15	16	
<i>Mean</i>	0.56	0.27	0.29	-0.17	-0.01	-0.16	1.16	0.5	0.66	0.76	0.17	0.59	1.58	0.82	0.76
<i>SD</i>	1.36	1		0.58	0.53		1.52	1.21		0.95	0.54		1.91	1.59	
GR 8 Math	All Items			MC Items			All OE Items			SA Items			ECR Items		
	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff
<i>N of Steps</i>	56	56		25	25		31	31		16	16		15	15	
<i>Mean</i>	0.63	0.56	0.07	0.11	0.04	0.07	1.05	0.98	0.07	0.74	0.98	-0.24	1.4	0.99	0.41
<i>SD</i>	1.17	1.2		0.76	0.61		1.28	1.39		1.17	1.25		1.35	1.57	
GR 8 Math	All Items			MC Items			All OE Items			SA Items			ECR Items		
	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff	CBT	PPT	Diff
<i>N of Steps</i>	46	48		20	20		26	28		16	16		10	12	
<i>Mean</i>	0.59	0.27	0.32	-0.07	-0.18	0.11	1.1	0.6	0.5	1.47	0.73	0.74	0.51	0.41	0.1
<i>SD</i>	1.32	1.16		0.79	0.68		1.43	1.32		1.6	1.52		0.89	1.02	

The data visually illustrate that test items were delivered via computer became harder than delivered by paper/pencil of the same test. Students who earned a same raw score would receive different scale scores, a higher scale score for the CBT and a lower scale score for the PPT, depending on where the raw score was located on the scale. As discussed earlier, a larger CBT effect was found for grade 5 Reading and grade 8 Mathematics, where the average difference of scale scores was 11.6 and 12.8, respectively.

The mode effect has great impact on high-stakes decisions for individual students as well as for school and educator accountability. In grade 5 Reading, students need to earn 3 more raw score points on the PPT version (46 points) than on the CBT version (43 points) to achieve the performance level of Meets the Standard. Similarly, students who take the PPT version must have 1 to 3 more raw score points to reach the performance levels of Distinguished, Exceeds the Standards, and even Below the Standard, respectively, for grade 5 Reading. In grade 8 Mathematics, students need to earn 3 more raw score points on the PPT version (39 points) than on the CBT version (36 points) to achieve the Mathematics performance level of Meets the Standard. Similarly, students who take the PPT version must have 2 to 3 more raw score points to reach the performance levels of Distinguished, Exceeds the Standards, and even Below the Standard, respectively, for grade 8 Mathematics.

Figure 1. Comparison of Conversions for Grade 5 Reading

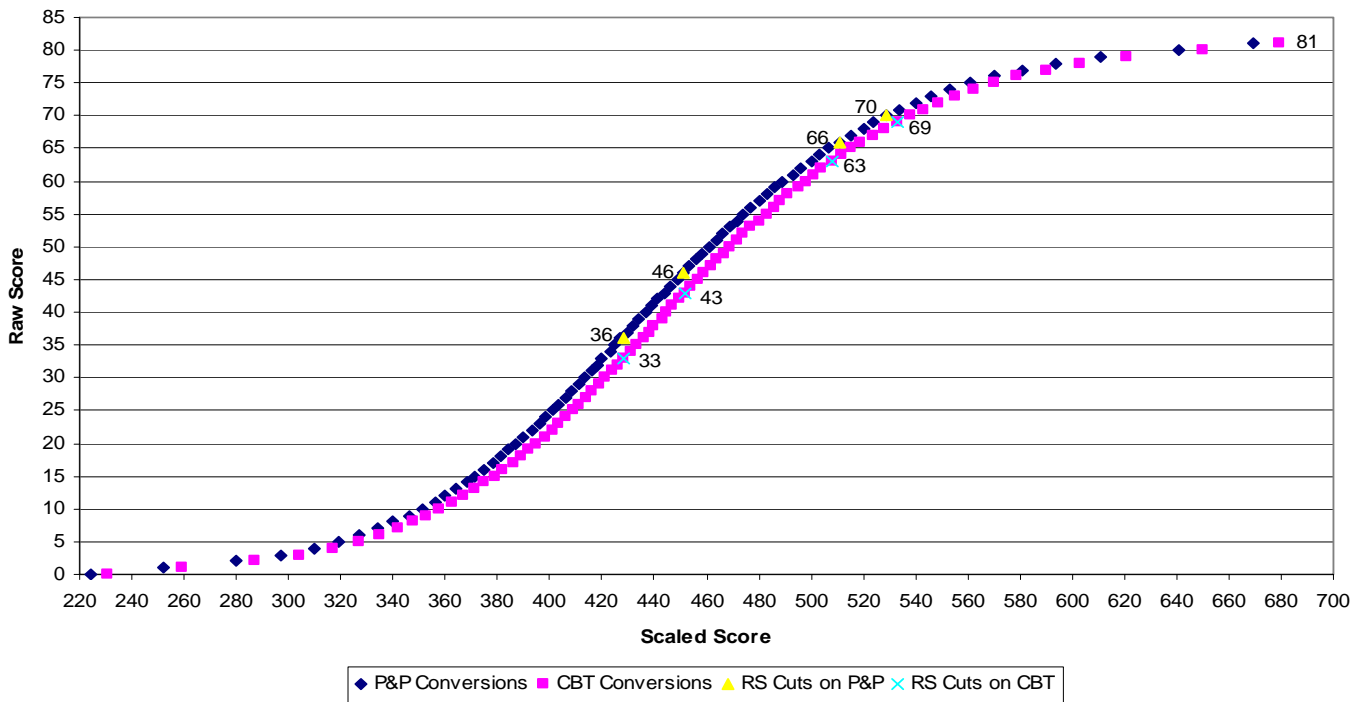


Figure 2. Comparison of Conversions for Grade 8 Reading

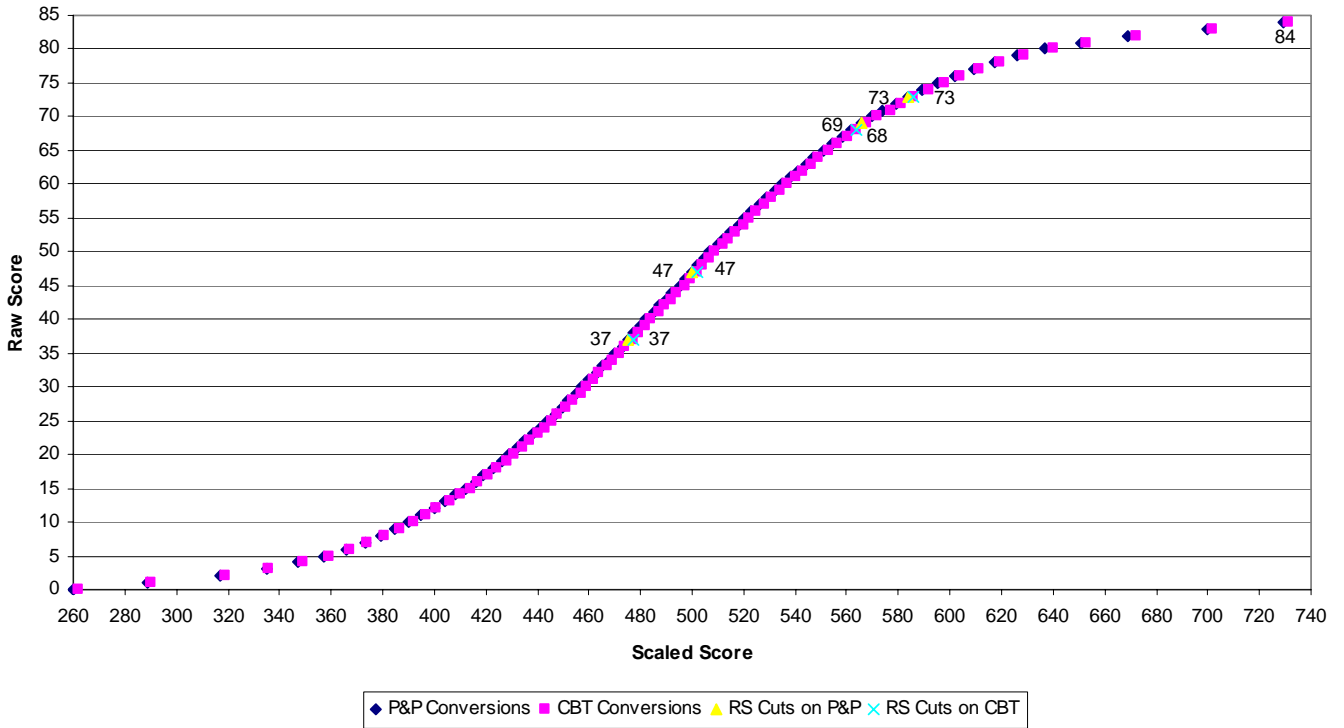
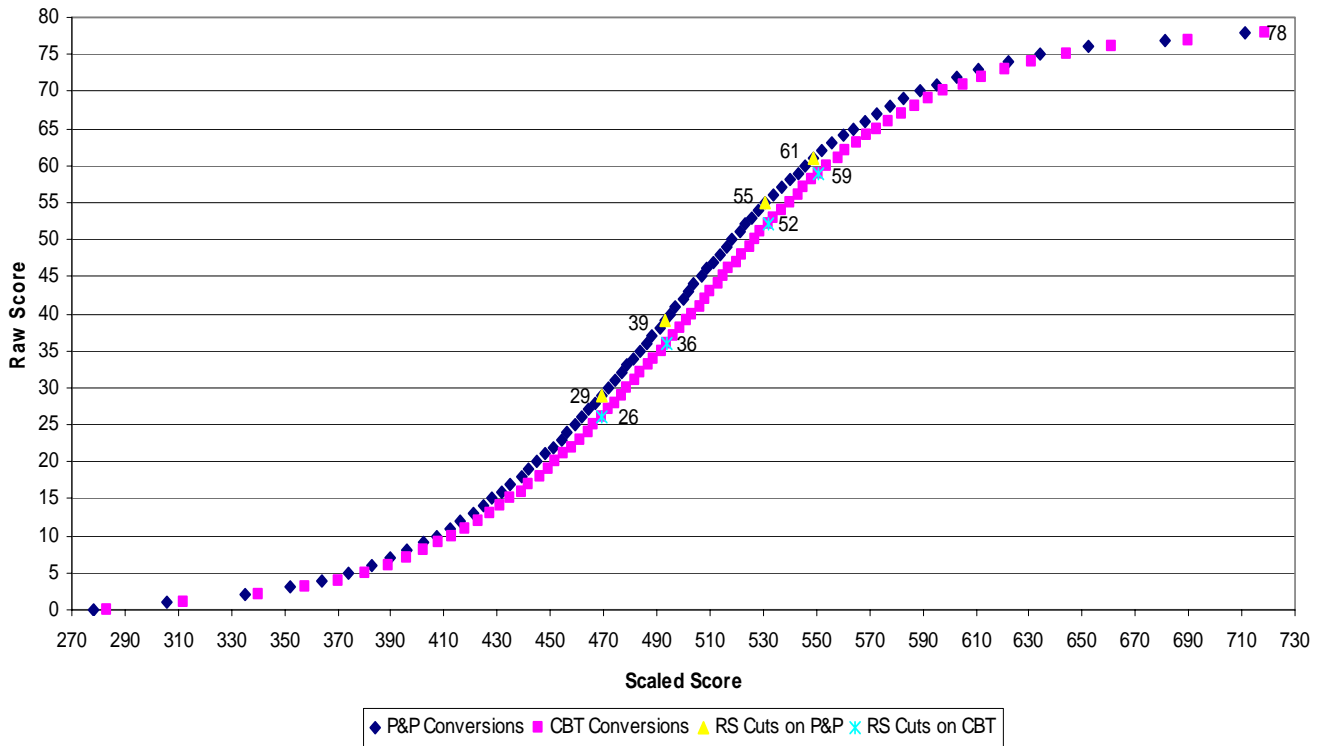


Figure 3. Comparison of Conversions for Grade 8 Mathematics



Consistency of Scoring: To examine the comparability of human-scoring and computer-automated scoring, cross-tab analysis was used for all constructed-response items (short-answer and open-ended formats). The results of comparisons are presented in Table 7 for each item by item format and test, including percent of perfect agreement, mean difference (human-scoring–computer-automated scoring), correlation coefficient, and Kappa Index for agreement. SA1 is short-answer with 0 and 2 points; SA2 has 0, 1, and 2 score points.

Table 7. Summary Results of Comparability of Scoring

Item Format	Agreement %	Kappa Index	r	Mean Diff	Item Format	Agreement %	Kappa Index	r	Mean Diff
Grade 5 Reading									
SA1*	74.3	53.3	0.57	0.35	OE	49.6	32.1	0.66	0.03
SA1	84.4	70.3	0.71	0.15	OE	56.2	42.9	0.79	-0.02
SA1	98.6	96.3	0.96	0.00	OE	59.1	41.6	0.74	-0.03
SA1	92.4	79.3	0.8	-0.07	OE	56.9	38.6	0.61	-0.12
SA1	96.4	96.9	0.97	0.01					
SA2*	70.3	54.2	0.76	-0.18					
SA2	62.0	44.6	0.72	0.17					
SA2	57.2	36.3	0.53	-0.19					
Grade 8 Reading									
SA1	98.7	92.5	0.93	0.04	OE	62.6	46.9	0.73	-0.19
SA1	79.4	42.2	0.42	0.00	OE	52.3	36.2	0.63	-0.47
SA1	79.4	25.2	0.26	0.12	OE	51.6	29.8	0.69	-0.06
SA1	93.5	89.2	0.89	-0.05	OE	64.5	54.9	0.79	-0.14
SA1	65.8	39.5	0.44	-0.45					
SA1	61.9	31.1	0.40	-0.66					
SA2	53.5	34.4	0.67	-0.30					
SA2	47.7	23.7	0.59	-0.50					
Grade 8 Mathematics									
SA1	99.0	97.5	0.98	-0.01	OE	74.1	61.9	0.84	-0.18
SA1	99.0	97.9	0.98	0.00	OE	67.9	52.4	0.82	-0.20
SA1	97.2	56.8	0.57	0.00	OE	39.6	18.5	0.49	-0.74
SA1	99.2	100	1.00	0.00					
SA1	97.9	97.5	1.00	-0.01					
SA1	97.4	94.4	0.90	-0.03					
SA2	90.7	60.3	0.80	-0.05					
SA2	94.3	91.3	0.95	0.03					

Data suggest a consistent pattern by item type across tests. The highest percent of perfect agreement between computer-automated scoring and human-scoring, on average, was for dichotomous short-answer items (61-99%), followed by short answer items with 0, 1, 2 score points (47-94%); while open-ended items had the lowest percent of perfect agreement (40-74%) between scoring process. Across tests, the mean differences between computer-automated scoring and human-scoring were smaller for Mathematics than for Reading for each item type. For instance, the mean differences for dichotomous short-answer items was around 0.0 for grade 8 Mathematics, but the mean differences ranged -.07 to .35 for grade 5 Reading and -.66 to .12 for grade 8 Reading. The agreement between computer-automated scoring and human-scoring also varied from item to item for the same item format. For example, the two dichotomous short-answer items for grade 8 Reading, one item had 99% agreement with the mean difference of .04 between two scores; while the other item had 62% agreement with the mean difference of -.66 out of 2, the maximum score points. These variations might be due to the quality of the item and the clarity of the scoring rubric.

The perfect agreement between two trained readers was then compared with the agreement between computer-automated scoring and human-scoring. Data in Table 8 suggest several consistent patterns. First of all, the agreement between computer-automatic and human-scoring and the agreement between two trained raters was higher for Mathematics than for Reading. Secondly, the highest agreement was found for dichotomous short-answer items and the lowest agreement for open-ended items. Thirdly, a consistent higher agreement between the two raters than the agreement between computer-automated scoring and human-scoring, respectively, by item formats and tests.

Table 8. Comparison of Perfect Agreement between Scoring Methods

Test	SA1 (0, 2)		SA2 (0 - 2)		OE (0 - 4)	
	CAS and HS	Raters	CAS and HS	Raters	CAS and HS	Raters
Reading						
Grade 5	89.2	92.0	63.2	77.4	55.5	56.9
Grade 8	79.8	87.4	50.6	70.7	57.8	55.2
Mathematics						
Grade 8	98.3	99.8	92.5	98.8	60.5	79.4
Total	89.1	93.1	68.0	81.6	57.7	62.4

CAS: Computer-automatic scoring

HS: Human-scoring

Findings and Lessons Learned

Findings of the Study: The objectives of this pilot study were to investigate the comparability of test scores between the paper/pencil (PPT) and computer-based (CBT) test administration modes and to examine the consistency between human-scoring and computer-automated scoring on short-answer and open-ended items. Over 1,600 volunteer students participated in the study and the majority of them were low-achieving students. Participants were assigned into two groups, PPT or CBT, for grades 5 and 8 Reading, and grade 8 Mathematics. This study employed the common-item, common-administration-mode, non-equivalent groups design to equate the two administration modes. The same items from the abbreviated version of Stanford Achievement Test, 9th edition (SAT9) under the same paper/pencil administration were used as anchor to put the Delaware-developed items portion under paper/pencil and computer-based administration into same scale.

A positive CBT effect was found for all three tests, .28 for grade 5 Reading, .07 for grade 8 Reading, and .32 for grade 8 Mathematics. The CBT effects resulted in the average changes of scale scores of 11.6, 2.8, and 12.8, respectively, for grades 5 and 8 Reading, and grade 8 Mathematics. The average changes of scale scores represent different raw score points from 1 to 6. The results of equating indicated that the CBT items appeared harder than the PPT items of the same test. Consequently, students using paper/pencil version must earn more raw score points, particularly on grade 5 Reading or grade 8 Mathematics, to receive the same scale scores as their counterparts using computer-based version. The consequences become more serious when the performance levels are used for high-stakes decisions.

Although the results of this study did not support a consistent pattern by item formats, a larger CBT effect was observed for constructed-response items (including short-answer and open-ended items) than for multiple-choice items in most cases.

The results suggest a consistent pattern between computer-automated scoring and human-scoring by item type across the three tests. The highest percent of perfect agreement was found, on average, for dichotomous short-answer items, followed by short-answer items with 0, 1, 2 score points; while the open-ended items had the lowest percent of perfect agreement between the two scoring process. In addition, the rater consistency was found consistently higher than the agreement between computer-automated scoring and human-scoring across item types and tests.

Lessons Learned from the Study: Findings from previous studies (Russell and Plati, 2000; Bennett, 2001, 2002; Choi and Tinkler, 2002) indicated that the familiarity to computers may have great impact on student performance, particularly for young students. According to the results of the 2005 DSTP Annual Student Survey, the opportunity to access a home computer was still significantly lower for African American and Hispanic students than Caucasian and Asian students even though the percentage steadily increased from year to year. More minority students, students with disabilities, and students from a low-income family participated in the PPT Group of the pilot study might

suggest that those students were more comfortable to use the paper/parcel version. This issue could threaten the construct validity, especially for high-stakes assessments.

Participating teachers and students provided positive feedback about the CBT administration, such as easier shipping, handling, and distributing test booklets. Students were very interested in taking a test on the computer. However, nearly all the incidents occurred during testing, according to the Incident Reports and Observation Reports for the pilot study, were related to technology difficulties, software design, and experience of computer use for younger students. The feasibility of CBT for statewide assessment programs, in addition to other factors, largely depends on the quality of computer network system, computer literacy of teachers, and qualified technology staff in school settings.

Limitations of this study should be taken into account when considering the findings. First, the small, volunteer sample might have caused unstabled step values when using the equating approach to examine comparability of test score across administration modes. Moreover, the positive skewed frequency distributions and lack of full coverage of test score scale, due to the majority of low-achievement participants, might be the source of sampling errors. The changes of item layout and the context of few Mathematics items due to limited features and functions of the software from the paper/pencil version to the computer-based version might have contribution to the CBT effect.

References

Bennett, R. (2003). Comparability in online assessment. Paper presented at the 2004 CCSSO Large-Scale Assessment Conference, June 2004, Boston, MA.

CCSSO (2003). Computer-based testing survey. A publication of Council of Chief State School Officers.

Choi, S.W. and Tinkler, T. (2002). Evaluating comparability of paper-and pencil and computer-based assessment in a K-12 setting. Paper presented at the NCME Annual Meeting, April, 2002, New Orleans, LA.

Court, S.C. (2005). Statistical and substantive comparability: The need for between-mode-equating and an equal number of test forms. A Report submitted to the State.

Delaware Student Testing Program: Report of Student Survey Questionnaire (2005).

Ewing, M., Wiley, A. & Gillie, J.M. (2003). Moving from paper-and-pencil administration to computer-based testing: An investigation of construct equivalence and subgroup differences. Paper presented at the 2003 NCME Annual Meeting, April 2003, Chicago, IL.

Russell, M. and Plati, T. (2000). Mode of administration effects on MCAS Composition performance for grades four, eight, and ten. A Report of Findings Submitted to the Massachusetts Department of Education.

Schwarz, R.D., Rich, C. & Podrabsky, T. (2003). A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests. Paper presented at the 2003 NCME Annual Meeting, April 2003, Chicago, IL.

Wang, S.D., Young, M.J., & Brooks, T.E. (2004). Administration mode comparability study for Stanford Diagnostic Reading and Mathematics tests. Research Report, Harcourt Assessment, Inc.

Wang, S.D., Young, M.J., & Brooks, T.E. (2006). A state Stanford-9 online Writing test study: Validity evidence of computer automated scoring on web test. An unpublished Report.

Zhang, L.R and Lau, C.A. (2005). Examining the comparability of test scores between computer-based and paper/pencil testing modes. Paper presented at the 2005 CCSSO Large-Scale Assessment Conference, June 2005, San Antonio, TX.

Appendix

Summary of Student Survey for Grade 5 Reading

Question	CBT		PPT	
1. Do you have a computer at home that you can use?	N.	%	N.	%
Yes	196	70.8	148	74.0
No	45	16.2	50	25.0
No response	36	13.0	2	1.0
Total	277	100.0	200	100.0
2. How often do you use the computer at home for learning?	N.	%	N.	%
Almost every day	58	20.9	35	17.5
Once or twice a week	73	26.4	52	26.0
Once or twice a month	38	13.7	40	20.0
Never	64	23.1	70	35.0
No response	44	15.9	3	1.5
Total	277	100.0	200	100.0
3. How often do you use the computer at home for fun?	N.	%	N.	%
Almost every day	106	38.3	66	33.0
Once or twice a week	69	24.9	46	23.0
Once or twice a month	18	6.5	26	13.0
Never	38	13.7	57	28.5
No response	46	16.6	5	2.5
Total	277	100.0	200	100.0
4. Does your school have computers that you can use?	N.	%	N.	%
Yes	236	85.2	198	99.0
No	4	1.4	1	0.5
No response	37	13.4	1	0.5
Total	277	100.0	200	100.0
If you answer "no" to question 4, you may skip questions 5-7. If you answer "yes" to question 4, please answer questions 5-7.				
5. How often do you use the computer in school for learning?	N.	%	N.	%
Almost every day	40	14.4	38	19.0
Once or twice a week	108	39.0	66	33.0
Once or twice a month	53	19.1	52	26.0
Never	34	12.3	43	21.5
No response	42	15.2	1	0.5
Total	277	100.0	200	100.0

Question	CBT		PPT	
6. How often do you use the computer in math class?	N.	%	N.	%
Almost every day	15	5.4	4	2.0
Once or twice a week	24	8.7	19	9.5
Once or twice a month	22	7.9	16	8.0
Never	175	63.2	161	80.5
No Response	41	14.8	0	0.0
Total	277	100.0	200	100.0
7. How often do you use the computer in reading class?	N.	%	N.	%
Almost every day	21	7.6	15	7.5
Once or twice a week	43	15.5	31	15.5
Once or twice a month	44	15.9	40	20.0
Never	124	44.8	114	57.0
No Response	45	16.2	0	0.0
Total	277	100.0	200	100.0

Summary of Student Survey for Grade 8 Reading

Question	CBT		PPT	
1. Do you have a computer at home that you can use?	N	%	N	%
Yes	116	74.4	16	69.6
No	26	16.7	7	30.4
No response	14	9.0	0	0.0
Total	156	100.0	23	100.0
2. How often do you use the computer at home for learning?	N	%	N	%
Almost every day	33	21.2	2	8.7
Once or twice a week	35	22.4	5	21.7
Once or twice a month	39	25.0	9	39.1
Never	33	21.2	7	30.4
No response	16	10.3	0	0.0
Total	156	100.0	23	100.0
3. How often do you use the computer at home for fun?	N	%	N	%
Almost every day	65	41.7	9	39.1
Once or twice a week	38	24.4	4	17.4
Once or twice a month	11	7.1	3	13.0
Never	27	17.3	7	30.4
No response	15	9.6	0	0.0
Total	156	100.0	23	100.0
4. Does your school have computers that you can use?	N	%	N	%
Yes	132	84.6	23	100.0
No	7	4.5	0	0.0
No response	17	10.9	0	0.0
Total	156	100.0	23	100.0
If you answer "no" to question 4, you may skip questions 5-7. If you answer "yes" to question 4, please answer questions 5-7.				
5. How often do you use the computer in school for learning?	N	%	N	%
Almost every day	35	22.4	4	17.4
Once or twice a week	23	14.7	1	4.3
Once or twice a month	62	39.7	14	60.9
Never	14	9.0	4	17.4
No response	22	14.1	0	0.0
Total	156	100.0	23	100.0

Question	CBT		PPT	
6. How often do you use the computer in math class?	N	%	N	%
Almost every day	6	3.8	1	4.3
Once or twice a week	9	5.8	1	4.3
Once or twice a month	10	6.4	12	52.2
Never	111	71.2	9	39.1
No Response	20	12.8	0	0.0
Total	156	100.0	23	100.0
7. How often do you use the computer in reading class?	N	%	N	%
Almost every day	8	5.1	0	0.0
Once or twice a week	11	7.1	0	0.0
Once or twice a month	32	20.5	3	13.0
Never	83	53.2	20	87.0
No Response	22	14.1	0	0.0
Total	156	100.0	23	100.0

Summary of Student Survey for Grade 8 Mathematics

Question	CBT		PPT	
1. Do you have a computer at home that you can use?	N	%	N	%
Yes	264	68.2	211	76.7
No	67	17.3	64	23.3
No response	56	14.5	0	0.0
Total	387	100.0	275	100.0
2. How often do you use the computer at home for learning?	N	%	N	%
Almost every day	65	16.8	68	24.7
Once or twice a week	85	22.0	57	20.7
Once or twice a month	80	20.7	60	21.8
Never	97	25.1	87	31.6
No response	60	15.5	3	1.1
Total	387	100.0	275	100.0
3. How often do you use the computer at home for fun?	N	%	N	%
Almost every day	167	43.2	126	45.8
Once or twice a week	72	18.6	60	21.8
Once or twice a month	24	6.2	22	8.0
Never	66	17.1	64	23.3
No response	58	15.0	3	1.1
Total	387	100.0	275	100.0
4. Does your school have computers that you can use?	N	%	N	%
Yes	315	81.4	269	97.8
No	14	3.6	4	1.5
No response	58	15.0	2	0.7
Total	387	100.0	275	100.0
If you answer "no" to question 4, you may skip questions 5-7. If you answer "yes" to question 4, please answer questions 5-7.				
5. How often do you use the computer in school for learning?	N	%	N	%
Almost every day	61	15.8	44	16.0
Once or twice a week	69	17.8	49	17.8
Once or twice a month	130	33.6	104	37.8
Never	52	13.4	63	22.9
No response	75	19.4	15	5.5
Total	387	100.0	275	100.0

Question	CBT		PPT	
6. How often do you use the computer in math class?	N	%	N	%
Almost every day	16	4.1	19	6.9
Once or twice a week	25	6.5	25	9.1
Once or twice a month	38	9.8	21	7.6
Never	238	61.5	194	70.5
No Response	70	18.1	16	5.8
Total	387	100.0	275	100.0
7. How often do you use the computer in reading class?	N	%	N	%
Almost every day	8	2.1	11	4.0
Once or twice a week	15	3.9	28	10.2
Once or twice a month	73	18.9	53	19.3
Never	217	56.1	169	61.5
No Response	74	19.1	14	5.1
Total	387	100.0	275	100.0