

SDRT 4/SDMT 4 Administration Mode Comparability Study



Shudong Wang, Ph.D.

Michael J. Young, Ph.D.

Thomas E. Brooks, Ph.D.

August 2004

SDRT 4/SDMT 4 Administration Mode Comparability Study**Acknowledgements**

This technical report summarizes a larger, complete work entitled *Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests*, which is the result of a collaborative effort by many Pearson Inc. staff members who should be recognized for their significant contributions. The complete work will be available from the first author in fall 2004.

Nathan Wall, Doug McAllaster, Holly Zhang, Jenny Jiang, and Agnes Stephenson, Ph.D. assisted with data analysis and input, created tables and figures, and provided quality assurance for the results.

Jenny Hoffmann and David J. Kirk provided technical writing support.

Pearson Sampling staff coordinated the research management and recruitment of participants for the study. They also supervised the administration of the tests. Sampling staff included David J. Kirk, Lucie Morales, Carlos Ramirez, Priscilla Villanueva, John Ramirez, and Betsy J. Case, Ph.D.

Craig Douglas, Stephanie Redding, and June Wilson led the development and launch of the computer-based version of the tests.

This article is a technical summary of a forthcoming research report by Pearson's Psychometric and Research Services entitled *Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests*. All supporting data and tables will be available in the final publication of the full report.

SDRT 4/SDMT 4 Administration Mode Comparability Study

Introduction

The widespread availability of computers in schools has focused the assessment community on the use of computer-based testing solutions in the classroom. There are many potential benefits and advantages associated with tests delivered via computer including immediate scoring and reporting, greater test security, flexible test administration schedules, reduced costs compared to handling paper-and-pencil test materials, the use of multimedia item types, and the ability to measure response time (Boo and Vispoel, 1998; Folk and Smith, 1998; Schmit and Ryan, 1993; Klein and Hamilton, 1999; Wang, Young, and Brooks, 2003).

However, as schools include computer-based testing in their traditionally paper-and-pencil assessment systems, concerns arise about the validity and comparability of scores from the two administration modes. This summary will discuss the results of an administration mode comparability study conducted by Pearson's Psychometric and Research Services using the *Stanford Diagnostic Reading Test*, (SDRT 4) 4th Ed. and *Stanford Diagnostic Mathematics Test*, (SDMT 4) 4th Ed. tests delivered in both computer-based (online) and paper-and-pencil formats.

The Stanford Diagnostic Reading and Mathematics Tests

The SDRT 4 and SDMT 4 were published in 1996 by Pearson Inc. (Pearson). SDMT 4 *Online Testing* delivers the same content as the paper-based version. SDRT 4 *Online Testing* delivers the same content as the paper-based version with the exception of the scanning subtest available only on the paper version. The online versions of both SDRT 4 and SDMT 4 provide teachers and students with immediate feedback and are intuitive and user-friendly. While the paper versions are user friendly as well, scoring is not instantaneous.

SDRT 4 provides group-administered diagnostic assessment of the essential components of reading to determine students' strengths and needs in grades 2 through 12. The test provides detailed coverage of reading skills, including many easy questions, so teachers can better assess and plan instruction for students who are struggling with reading.

SDRT 4/SDMT 4 Administration Mode Comparability Study

SDMT 4 measures competence in the basic concepts and skills prerequisite to success in mathematics, while emphasizing problem-solving concepts and strategies. The test identifies specific areas of difficulty for each student in grades 2 through 12 so that teachers can plan appropriate intervention.

Administration Mode Effects

The primary concern that arises when a test is administered to students in two different modes is that results might be affected by the change in delivery format. That is, are student scores from both test administration modes equivalent? For example, it may be possible for a student's level of computer anxiety or familiarity to affect test scores when compared to scores from students taking the paper-and-pencil version.

In *Guidelines for Computer-Based Tests and Interpretations* (American Psychological Association [APA], 1986), score equivalence between computer-based tests (CBT) and paper and pencil tests is defined as follows:

Scores from conventional and computer administrations may be considered equivalent when: (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) 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* (1986) also emphasize the importance of limiting influences such as computer anxiety and computer experience irrelevant to the purposes of the test.

Studies conducted by other researchers typically find that measures obtained from tests delivered in the computer-based mode are similar to those obtained from the paper-and-pencil mode (Bergstrom, 1992; Boo and Vispoel, 1998; Bugbee, 1996; Evans, Tannehill, and Martin, 1995; Neuman and Baydoun, 1998; Wang, Newman, and Witt, 2000; Wang, Young, and Brooks, 2003).

Purpose of the Study

This Pearson research study was designed to collect empirical evidence to study the comparability and equivalence of scores obtained from the computer-based and paper-and-pencil administration modes of SDRT 4/SDMT 4. An answer was sought to the following question:

What are the effects of administration mode (computer-based or paper-and-pencil) and mode order (computer-based first or paper-and-pencil first) on the variability and magnitude of test scores?

SDRT 4/SDMT 4 Administration Mode Comparability Study**Methods****Instruments**

The SDRT 4 and SDMT 4 tests are available in six “color” levels that cover grades 2 through 12. For purposes of the study, these test levels are referred to as L1 through L6 and were administered to grades 2 through 12 as shown in Table 1. (Note that levels L4, L5, and L6 comprise two or more grades each and that ninth-graders participated in the administration of both levels L5 and L6.)

Table 1. Number of Items by Total and Subtest for SDRT 4/SDMT 4 Levels L1 through L6*

		Red (L1)	Orange (L2)	Green (L3)	Purple (L4)	Brown (L5)	Blue (L6)
		Grade 2	Grade 3	Grade 4	Grades 5–6	Grades 7–9br**	Grades 9bl–12**
Total and Subtest							
SDRT 4	Total Reading***	120	110	115	84	84	84
	Phonetic Analysis	40	30	30			
	Vocabulary	40	40	40	30	30	30
	Comprehension	40	40	45	54	54	54
SDMT 4	Total Mathematics	52	52	52	52	52	52
	Concepts & Applications	32	32	32	32	32	32
	Computation	20	20	20	20	20	20

* For computer-based and paper-and-pencil administration modes.

** “9br” denotes ninth-graders taking the Brown Level and “9bl” denotes ninth-graders taking the Blue Level.

*** The paper-and-pencil version of SDRT 4 includes a “Scanning” subtest that is not part of the computer-based version of SDRT 4 and was not administered as part of the paper-and-pencil SDRT 4 in the study.

Norm-referenced scores and criterion-referenced progress indicators are available with the SDRT 4 and SDMT 4 tests. Additionally, the SDRT 4/SDMT 4 subtests have been statistically equated to the *Stanford Achievement Test Series*, as well as to the third editions of the *Stanford Diagnostic Reading Test* and the *Stanford Diagnostic Mathematics Test*. Additional information about SDRT 4/SDMT 4 development, standardization, and scoring can be found in the corresponding *Teacher’s Manuals for Interpreting* (Pearson, 1995).

Subjects and Data Collection

In July and August 2003, students who were registered to attend grades 2 through 12 in the coming fall were invited to participate in the comparability study. Interested students were invited to participate in one or both of the SDRT 4 or SDMT 4 testing sessions. Both modes of a test (computer-based and paper-and-pencil) were administered to each participant. Financial compensation was provided to each

SDRT 4/SDMT 4 Administration Mode Comparability Study

student who registered for a SDRT 4 or SDMT 4 testing session. Some students registered for both SDRT 4 and SDMT 4 testing sessions.

Of the sample of students participating in the study, results for a total of 1,863 students taking SDRT 4 and 1,774 students taking SDMT 4 were analyzed. Among these examinees, approximately 51% were female and 49% male.

To arrive at these totals, some of the examinees' results had to be excluded from the study because they neglected to provide the correct identifying information and, as a result, their computer-based scores could not be matched with their paper-and-pencil test scores. For a full breakdown of the numbers of students (*N*-counts) who took the two modes of SDRT 4 or SDMT 4 please refer to Tables 2 through 13 in the full report. These tables also present score-related descriptive statistics for grades 2 through 12.

For the purposes of the study, examinees in each grade were randomly assigned to one of two groups. One group took the paper-and-pencil test first, followed by the same test in the computer-based mode. The other group took the computer-based test first, followed by the same test in the paper-and-pencil mode. There was a short break between the administrations of the two modes for each student. After finishing both modes of a particular test (i.e., the SDRT 4 or the SDMT 4), each examinee completed a Comparability Study Survey Form.

Testing for the study took place over a two-month period at two locations in San Antonio. One testing center was located at the Pearson office building; the second, at a nearby conference facility. Each testing center conducted simultaneous computer-based and paper-and-pencil testing sessions in separate rooms holding 25 examinees per session. Testing took place primarily during the day. Eight evening sessions were also offered for students in grades nine, ten, eleven, and twelve to accommodate work/school schedules.

Scoring Procedures

Scores were reported as raw scores (number correct) for each SDRT 4 and SDMT 4 subtest. The Total Reading and Total Mathematics scores are the sums of the respective subtest scores.

Experimental Design

To investigate the effects of administration mode and mode order on the variability and magnitude of test scores, the *split-plot repeated-measures design* was selected. It is the most powerful quantitative research method for testing causal hypotheses (Gall, Borg, and Gall, 1996).

The study is structured with one between-subject factor called *mode order* and one within-subject factor called *mode* with unequal group sizes. The number of subjects who took the computer-based test first (N_1) was not equal to the number of subjects

SDRT 4/SDMT 4 Administration Mode Comparability Study

who took the paper-and-pencil test first (N_2). Subjects are nested within mode order but crossed with respect to mode; that is, all examinees were tested in both modes. To control for carryover effects (memory, fatigue, and attention loss) from one test mode to the other, the mode order was counterbalanced. Mode order was treated as a fixed effect, while subjects were considered random. Figure 1 illustrates this experimental design.

Figure 1. Split Plot Repeated Measures Design with Unequal Group Size ($N_1 \neq N_2$)

		Subject	Within-Subject Factor (Mode)	
			Paper-and-Pencil Score	Computer-Based Score
Between-Subject Factor (Mode Order)	Computer-Based First	1	X_{111}	X_{121}
		2	X_{112}	X_{122}
	
		N_1	X_{11N_1}	X_{12N_1}
	Paper-and-Pencil First	1	X_{211}	X_{221}
		2	X_{212}	X_{222}
	
		N_2	X_{21N_2}	X_{22N_2}

Typically, balanced groups are used with the split plot repeated measures design. However, because of variations in the group sizes for the study, there were unequal numbers of examinees in the two mode order groups. One solution for balancing unequal groups in a split plot repeated measures design is to match the number of observations in the smaller cells with the same number of observations randomly selected from the larger cells. In this study, however, such treatment would reduce N -counts in cells already having small numbers of subjects. Although the numbers of subjects in the two mode order groups were unequal, the cell sizes in each mode were proportional to the number of subjects in mode order. Thus, the effect of non-orthogonality leading to F -tests for confounded effects was relatively minor (Glass & Hopkins, 1996; Winer, Brown, and Michels, 1991).

Data Analysis and Results

Descriptive Statistics by Grade and Test Level

N -counts and the first four moments describing the distributions of raw scores for SDRT 4/SDMT 4 totals and subtests were calculated for all grades and test levels.

SDRT 4/SDMT 4 Administration Mode Comparability Study

The four moments are: mean (average), standard deviation (measure of average deviation from the mean), skewness (degree of asymmetry of a distribution), and kurtosis (degree of peakedness of a distribution). These statistics (not all shown in this publication) were calculated for both administration modes of SDRT 4/SDMT 4 for grades 2 through 12 and test levels L4, L5 and L6. Table 2, which presents descriptive statistics for grade 6, is from the complete data set.

Table 2. Descriptive Statistics for SDRT 4/SDMT 4 Totals and Subtests for Grade 6

Test	Mode	Total/Subtest	N	Mean*	SD	Skewness	Kurtosis
Reading (SDRT 4)	CBT	Total	283	67.69	13.06	-1.44	1.73
		Vocabulary	283	24.70	4.47	-1.74	3.18
		Comprehension	283	42.99	9.32	-1.34	1.39
	P&PT	Total	283	67.22	13.49	-1.47	1.76
		Vocabulary	283	24.66	4.70	-1.88	3.65
		Comprehension	283	42.55	9.47	-1.25	1.04
Mathematics (SDMT 4)	CBT	Total	317	42.36	7.75	-1.04	0.48
		Concepts & Applications	317	26.04	5.06	-1.00	0.38
		Computation	317	16.32	3.45	-1.24	1.25
	P&PT	Total	317	42.24	8.91	-1.45	1.97
		Concepts & Applications	317	25.89	5.67	-1.33	1.48
		Computation	317	16.34	3.89	-1.54	2.09

* Based on raw scores. SD =Standard Deviation, CBT=computer-based test, and P&PT=paper-and-pencil test.

Score Means

For SDRT 4, the differences between the mean scores for totals and subtests administered in either mode are less than 1.00 raw score point for all grades and test levels with the following exception: For grade 2 only, the difference between the Total Reading score means for the computer-based and paper-and-pencil modes is 2.37 raw score points and the difference between the Comprehension subtest score means for the two modes is 1.40 raw score points.

For SDMT 4, the differences between the mean scores for totals and subtests administered in either mode are less than 1.00 raw score point for all grades and test levels.

Note that for the exceptions where there were differences in mean scores, equivalent score interpretations between modes can be obtained through equating methods.

SDRT 4/SDMT 4 Administration Mode Comparability Study***Standard Deviation, Skewness, and Kurtosis***

For SDRT 4 and SDMT 4, standard deviation, skewness, and kurtosis values for totals and subtests are essentially equal in value when modes are compared for all grades and test levels.

In short, the descriptive statistics calculated for all grades and test levels demonstrate that scores obtained from either administration mode of SDRT 4/SDMT 4 (with the exception noted above) are essentially equal in value. For example, a sixth-grader's score on the computer-based Computation subtest can be directly compared to a sixth-grader's score on the same subtest taken in paper-and-pencil mode.

Differences in mean scores based on mode and mode order were determined by inferential statistics and are discussed under *Analysis of Variance (ANOVA) Results by Grade* and *Analysis of Variance (ANOVA) Results by Test Level*.

Variability of Test Scores

Reliability coefficients (Cronbach's coefficients alpha) were calculated for SDRT 4/SDMT 4 totals and subtests for each mode order and for both mode orders combined (overall) within each administration mode. These data were calculated for grades 2 through 12 and test levels L4, L5, and L6. (Table 3, which presents reliability coefficients for grade 6, is from the complete data set.)

SDRT 4/SDMT 4 Administration Mode Comparability Study**Table 3. Reliability Coefficients for SDRT 4/SDMT 4 by Mode and Mode Order for Grade 6**

Subject	Mode	Total/Subtest	No. Items	Cronbach's Coefficient Alpha		
				CBT First	P&PT First	Overall
Reading (SDRT 4)				(N=149)	(N=134)	(N=283)
	CBT	Total	84	0.95	0.94	0.94
		Vocabulary	30	0.84	0.83	0.83
		Comprehension	54	0.93	0.92	0.93
	P&PT	Total	84	0.95	0.94	0.95
		Vocabulary	30	0.86	0.83	0.85
		Comprehension	54	0.93	0.93	0.93
Mathematics (SDMT 4)				(N=181)	(N=136)	(N=317)
	CBT	Total	52	0.88	0.91	0.89
		Concepts & Applications	32	0.83	0.87	0.85
		Computation	20	0.80	0.80	0.80
	P&PT	Total	52	0.93	0.91	0.92
		Concepts & Applications	32	0.89	0.87	0.88
		Computation	20	0.88	0.81	0.85

CBT=computer-based test and P&PT=paper-and-pencil test.

The reliability coefficient presented here is an index of change in the relative standing of students in a group from one administration to another (Nitko, 2004), the prominent point being that the higher the coefficient alpha value, the lower the expectation that students' scores will differ from one administration to the next. Hence, a high reliability coefficient means that the test is a reliable measure each time it is used regardless of the administration mode.

For SDRT 4, coefficients alpha for Reading *totals* by mode, mode order, and overall were 0.89 or higher for all grades and test levels. Coefficients alpha for *subtest* scores by mode, mode order, and overall were 0.78 or higher for all grades and test levels with the exception of the Vocabulary subtest at grades 2 and 9, where they were somewhat lower, especially for the second-grade students taking the computer-based test first. The lower subtest values are not unexpected because at grade 2, the *N*-counts were relatively low and computer unfamiliarity can be high and can negatively affect scores. For grade 9, the lower reliability coefficients may be due to the small number of items in the Vocabulary subtest relative to the Comprehension subtest as well as low *N*-counts.

SDRT 4/SDMT 4 Administration Mode Comparability Study

For SDMT 4, coefficients alpha for Mathematics *totals* by mode, mode order, and overall were 0.88 or higher for all grades and test levels except grade 9, where they were slightly lower. Coefficients alpha for *subtest* scores by mode, mode order, and overall were 0.80 or higher for all grades and test levels with the exception of grades 9 and 10. For grade 9, the reliability coefficients were somewhat lower, especially for the Computation subtest and for the group taking the paper-and-pencil test first. At grade 10, the reliability coefficients for the Computation subtest were slightly lower than 0.80. Of the SDRT 4/SDMT 4 subtests, Computation is the shortest (20 items) and, therefore, relatively lower reliability coefficient values are not unexpected.

In summary, reliability coefficients for SDRT 4/SDMT 4 totals and subtests were moderately high and generally not variable regardless of which mode or mode order was employed.

Magnitude of Test Scores

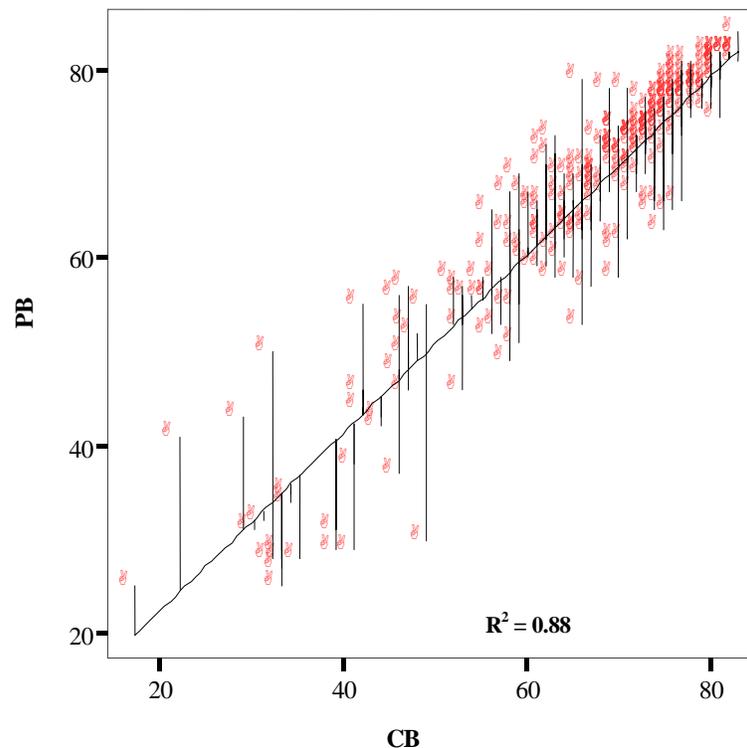
Four questions related to the magnitude of scores were posed:

1. Does a linear relationship exist between scores obtained from the computer-based administration mode versus the paper-and-pencil administration mode?
2. How does the ranking of computer-based scores compare with the ranking of paper-and-pencil test scores?
3. What differences, if any, exist between the means of total scores from the same computer-based and paper-and-pencil tests?
4. What impact, if any, does administration mode order have on mean test scores?

Scatter plots, evaluation of rank-order correlation coefficients, and analysis of variance (ANOVA) tables were used to answer these questions.

Linear and Rank Order Correlations

Scatter plots, including coefficients of determination (R^2), illustrating the linear relationship between computer-based and paper-and-pencil total scores were calculated for all grades and test levels. A visual inspection of all scatter plots shows that for all grades and test levels, the relationships between scores from the two modes are linear, positive, and range from moderate to high. (Figure 2, which presents a scatter plot and R^2 for grade 6, is from the complete data set.)

SDRT 4/SDMT 4 Administration Mode Comparability Study**Figure 2. Scatter Plot of SDRT 4 Scores for Grade 6**

PB=paper based and CB=computer based

The R^2 values, which represent the proportion of variation explained by the linear relationship, range from 0.84 to 0.92 for all SDRT 4 grades, except for grade 2, for which R^2 was 0.76, and for test levels L4, L5, and L6, for which the R^2 values were close to 0.80. For SDMT 4, R^2 values were 0.83 or higher, with the exception of those for grades 2 through 6 and test level L4, for which R^2 values ranged from 0.68 to 0.79.

Both Pearson's correlation coefficient (r_P) and Spearman's rank order correlation coefficients (r_S) were used to analyze the relationship between quantitative variables (i.e., scores from the computer-based mode and the paper-and-pencil mode). Although the r_P and r_S for the same variable within the same mode are usually close in value, they differ from one administration mode to the other due to differences in the null hypothesis being tested. Pearson's correlation coefficient measures the strength of the linear association between two variables and, therefore, will be greatly affected by the presence of outliers. On the other hand, Spearman's rank order correlation coefficient measures the rank order relationship between two variables. It does not use observed data; rather, it uses the ranks of the data without making an assumption about the distribution of scores.

SDRT 4/SDMT 4 Administration Mode Comparability Study

Pearson's correlation coefficients (r_p) were calculated to determine consistency of test scores (i.e., test reliability) on totals and subtests across administration modes.

Because examinees took the computer-based mode shortly after taking the paper-and-pencil mode, or vice versa, elapsed time between the two testing sessions could be controlled. If there is a linear relationship when scores from the modes are compared, the r_p values should be high. If the test is measuring the same strengths and weakness across administration modes, and if scores from both modes place examinees in similar rank order, the r_s values should also be high.

Results show that r_s and r_p were statistically significant at an alpha (α) level of 0.01 across grade and mode. In general, the magnitudes of r_s and r_p between the same totals/subtests for the different modes are around 0.90. These were larger than the magnitudes of r_s and r_p between different totals/subtests for the same mode or the magnitudes of r_s and r_p between different totals/subtests for the different modes. Similar patterns existed among all SDRT 4/SDMT 4 totals/subtests across the grades and levels.

Pearson's correlation coefficients and Spearman's rank order correlation coefficients were calculated for all grades and test levels. Table 4, which presents r_p and r_s values for grade 6 SDRT 4 total and subtest scores when administration modes are compared, is from the complete data set.

Table 4. Pearson's Coefficients of Equivalence (Upper Shaded Cells) and Spearman's Rank Order Correlation Coefficients (Lower Shaded Cells) for SDRT 4 Totals and Subtests—Grade 6

Mode	Total/ Subtest	CBT			P&PT		
		TT	VO	CO	TT	VO	CO
CBT	TT	1.00	0.89	0.98	0.91	0.80	0.90
	VO	0.84	1.00	0.77	0.83	0.85	0.76
	CO	0.97	0.71	1.00	0.88	0.71	0.89
P&PT	TT	0.91	0.78	0.89	1.00	0.90	0.98
	VO	0.76	0.87	0.65	0.82	1.00	0.79
	CO	0.88	0.68	0.90	0.98	0.69	1.00

CBT=computer-based test, P&PT=paper-and-pencil test, TT=Reading Total, VO=Vocabulary subtest, and CO=Comprehension subtest.

Pearson's correlation coefficients are, for the most part, 0.80 or higher for all SDRT 4/SDMT 4 totals and subtests. Pearson's correlation coefficients for the

SDRT 4/SDMT 4 Administration Mode Comparability Study

Comprehension subtest at grade 2 and the Computation subtest at grades 3 through 6, grade 12, and test level L4 were somewhat lower than .80.

With few exceptions, Spearman's rank order correlation coefficients are 0.80 or higher, with most being in the 0.85–0.95 range. Somewhat lower Spearman's coefficients were associated with the grade 2 Comprehension subtest and the Computation subtest at grades 3 through 6, grades 8 and 9, and test level L4.

Overall, the scatter plots, Pearson's correlation coefficients (r_p), and Spearman's rank order correlation coefficients (r_s) indicate that the relationships between SDRT 4/SDMT 4 total scores for the two administration modes were linear and that the degree of agreement between rank order of scores for the computer-based and paper-and-pencil modes was reasonably high. Regardless of mode order, grade, and test level, examinees scoring highly on the computer-based test also tended to score highly on the paper-and-pencil test.

Analysis of Variance (ANOVA) Results by Grade

For the SDRT 4 and SDMT 4, the differences in the means of total test scores across administration modes were determined by conducting an analysis of variance (ANOVA) based on the split plot repeated measures design of the study. Each examinee took the same form of either SDRT 4 or SDMT 4 in both administration modes. *Mode order* (computer-based first or paper-and-pencil first) was also analyzed as a between-subjects factor.

ANOVA results were calculated for SDRT 4 Reading and SDMT 4 Mathematics total scores for grades 2 through 12. Table 5, which shows results for SDRT 4 grade 6, is from the complete data set.

Table 5. ANOVA Results for SDRT 4 Grade 6

Source of Variation	<i>df</i>	Type III SS	MS	F	<i>p</i>
Between Subjects					
Mode Order	1	471.85	471.85	1.42	0.23
Error	278	92518.55	332.80		
Within Subjects					
Mode	1	1.83	1.83	0.17	0.68
Mode Order Interactions	1	1.39	1.39	0.13	0.72
Error (Mode)	278	2958.78	10.64		

df=degrees of freedom, Type III SS=Type III sum of squares, MS=mean square, F=statistics, and *p*=probability value.

SDRT 4/SDMT 4 Administration Mode Comparability Study

A probability value (p) equal to or greater than an alpha level (α) of 0.01 indicates that no statistically significant difference exists in the means of the total scores based on administration mode, mode order, or mode x mode order interactions.

The values for p show that for all SDRT 4 grades, no statistically significant differences exist in the means of the total Reading scores based on administration mode or mode order with the exception of grade 2, where there was a difference based on administration mode. The results suggest that regardless of mode or mode order, student performance on SDRT 4 would not be different (with the single exception noted) for any grade.

For SDMT 4, no statistically significant differences exist in the means of the total Mathematics scores based on administration mode or mode order with the exception of grade 4, where there was a difference based on administration mode. However, the difference between grade 4 mean total scores was less than one raw score point. These ANOVA results again suggest that regardless of mode or mode order, student performance on SDMT 4 would not be different (with the single exception noted) for any grade.

For both SDRT 4 and SDMT 4, no statistically significant differences in the means of total test scores were found based on mode x mode order interactions for any grade.

Analysis of Variance (ANOVA) Results by Test Level

ANOVA results were also calculated for SDRT 4 and SDMT 4 test levels L4 through L6. The data are aggregated from the different grades that correspond to the test level (see Table 1). For SDRT 4 and SDMT 4, there were no statistically significant differences in the means of total test scores based on administration mode, mode order, or mode x mode order interactions, with the following exceptions: (1) for SDRT 4 test level L6, the differences in the means based on administration mode and mode order were statistically significant, and (2) for SDMT 4 test level L6, the differences in the means based on mode x mode order interactions were statistically significant.

Overall, the ANOVA results indicate that SDRT 4/SDMT 4 scores for most of the grades and test levels were comparable across administration mode and mode order. Additionally, there were no statistically significant differences in the means of the scores based on mode x mode order interactions for any grade or test level.

Summary and Conclusions

The study examined empirically collected data to determine the comparability of scores resulting from computer-based and paper-and-pencil administrations of the SDRT 4 and SDMT 4 tests. The study focused on the effects of administration mode and mode order on test score variability and magnitude. The results of the study provide strong, broad-based evidence of the reliability and comparability of SDRT 4 and SDMT 4 scores for all grades and levels regardless of the administration mode.

SDRT 4/SDMT 4 Administration Mode Comparability Study

The major findings are as follows:

1. In general, raw scores from the different administration modes of SDRT 4/SDMT 4 had nearly equal or similar values for the first four moments: mean, standard deviation, skewness, and kurtosis.
2. All differences between the means of scores obtained from the computer-based and paper-and-pencil administration modes were less than one raw score point, with the exception of SDRT 4 grade 2. The difference between modes was greater for grade 2 due to expected higher levels of unfamiliarity with computers and a small N -count.
3. In general, reliability coefficients for SDRT 4/SDMT 4 totals and subtests were moderately high and generally not variable regardless of the administration mode or mode order. Larger numbers of items per subtest corresponded to larger corresponding reliability coefficients.
4. Overall, the scatter plots, Pearson's coefficients of equivalence (r_P), and Spearman's rank order correlation coefficients (r_S) indicate that the relationships between SDRT 4/SDMT 4 total scores for the two administration modes were linear and that the degree of agreement between rank order of scores for the computer-based and paper-and-pencil modes was reasonably high. Regardless of mode order, grade, and test level examinees scoring high on the computer-based test also tended to score high on the paper-and-pencil test.
5. In general, SDRT 4/SDMT 4 score means for most of the grades and test levels were comparable across administration mode and mode order, and there were no mode x mode order interactions.

The findings from this study support the comparison of scores from the computer-based and paper-and-pencil formats of the SDRT 4/SDMT 4 tests. Differences in scores based on administration mode do not appear to exceed expected random errors for most SDRT 4/SDMT 4 subtests across all grades and test levels. For the few exceptions where there were significant differences in mean scores, the test results obtained from the different administration modes will be equated to allow for equivalent score interpretations for computer-based and paper-and-pencil administrations.

This article is a technical summary of a forthcoming research report by Pearson's Psychometric and Research Services entitled *Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests*. All supporting data and tables will be available in the final publication of the full report.

References

- American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Boo, J., & Vispoel, W. P. (1998, April). *Computer versus paper-pencil assessment of educational development: Score comparability and examinee preference*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.
- Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5(3), 225–264.
- Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3), 282–299.
- Evans, L. D., Tannehill, R., & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments, & Computers*, 27(2), 162–165.
- Folk, V. G., & Smith, R. (1998, September). *Model for delivery of computer-based tests*. Paper presented at the ETS-sponsored colloquium on Computer-Based Tests: Building the Foundation for Future Assessments, Philadelphia, PA.
- Francis, L. J., and Evans, T. E. (1995). The reliability and validity of the Bath County Computer Attitude Scale. *Journal of Educational Computing Research*, 12(2), 135–146.
- Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). New York: Longman Publishers.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in educational and psychology* (3rd ed.). Needham Heights, MA: Allyn and Bacon.
- Harris, D. J., & Gao, X. (2003). *A conceptual synthesis of context effects*. Paper presented at the annual of the American Educational Research Association, Chicago, IL.
- Klein, S. P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions*. (Research Report IP-182). Santa Monica, CA: RAND.

SDRT 4/SDMT 4 Administration Mode Comparability Study

- Lankford, S. J., Bell, R. W., & Elias, J. W. (1994). Computerized versus standard personality measures: Equivalency, computer anxiety, and gender differences. *Computers in Human Behavior, 10*, 497–510.
- Massoud, S. L. (1991). Computer attitudes and computer knowledge of adult students. *Journal of Educational Computing Research, 7*(3), 269–291.
- McInerney, V., McInerney, D. M., & Sinclair, K. (1994). Student teachers, computer anxiety and computer experience. *Journal of Educational Computing Research, 11*(1), 27–50.
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*, 71–83.
- Nitko, A. J. (2004). *Educational assessment of students*. (4th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.
- Parshall, Cynthia G., & Jeffrey D. Kromrey. (1993, April). *Computer testing versus paper-and-pencil: An analysis of examinee characteristics associated with mode effect*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.
- Schmit, M. J., & Ryan, A. M. (1993). Test-taking disposition: A missing link? *Journal of Applied Psychology, 77*, 624–637.
- Wang, S., Newman, L., & Witt, E. A. (2000). *AT&T aptitude test equivalence study: A comparison of computer and paper-and-pencil employment examinations*. (Technical Report). Bala Cynwyd, PA: Harcourt Assessment System, Inc.
- Wang, S., Young, M. J., & Brooks, T. (June, 2003). *Validity evidence for the computer automated scoring of a web-administered assessment: The South Dakota/Stanford 9 online writing pilot*. (Technical Report). San Antonio, TX: Pearson Inc.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. (3rd ed.). New York: McGraw-Hill.

Additional copies of this and related documents are available from:

Pearson Inc.

19500 Bulverde Rd.

San Antonio, TX 78259

1-800-211-8378

1-877-576-1816 (fax)

<http://www.pearsonassess.com>