Online or Paper: Does Delivery Affect Results?

Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests

Shudong Wang, Ph.D.

June 2004
Acknowledgements

This article was the result of a collaborative effort by many Harcourt Assessment, Inc. staff members who should be recognized for their significant contributions.

Michael J. Young, Ph.D. and Thomas E. Brooks, Ph.D. participated in the preparation of the experimental design, conducted data and psychometric analyses, and reviewed the article.

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 provided technical writing support.

Harcourt Assessment 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 Ramírez, Priscilla Villanueva, John Ramírez, 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 summary of a forthcoming research report by Harcourt Assessment’s Psychometrics 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.
Online or Paper: Does Delivery Affect Results?

Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests

Introduction

About the Stanford Diagnostic Reading and Mathematics Tests

The Stanford Diagnostic Reading Test, Fourth Edition (SDRT 4) and Stanford Diagnostic Mathematics Test, Fourth Edition (SDMT 4) were adapted to a computer-based, online format in 2003. 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. Both provide teachers with immediate feedback. The online presentations are intuitive and user-friendly.

SDRT 4 provides group-administered diagnostic assessment of the essential components of reading in order to determine students’ strengths and needs in grades 1.5 through 13 (beginning of college). The test includes detailed coverage of reading skills so teachers can better evaluate and plan instruction for students who are struggling with reading.

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

Computer-based Testing

The widespread availability of computers in the last decade 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 and greater test security. Additional advantages include flexible test administration schedules, reduced costs compared to handling paper-based test materials, the use of multimedia item types, and the ability to measure response time.

As more schools obtain Internet access, online tests will become increasingly popular. In addition to the advantages of computer-based testing, online testing offers database expansion, on-demand reporting, and immediate feedback and updates.
However, as schools include computer-based testing in their traditionally paper-based assessment systems, concerns arise about the validity and comparability of scores from the two administration modes. This article examines the results of an administration mode comparability study conducted by Harcourt Assessment in support of this product.

**Administration Mode Effects**

A 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-based version. Research studies conducted by others typically find that measures obtained from tests delivered in the computer-based mode are similar to those obtained from the paper-based mode (Bergstrom, 1992; Boo and Vispoel, 1998; Bugbee, 1996; Evans, Tannehill, and Martin, 1995; Neuman and Baydoun, 1998; Wang, Newman, and Witt, 2000).

According to *Guidelines for Computer-Based Tests and Interpretations* (American Psychological Association [APA], 1986), score comparability or equivalence between computer-based tests and paper-based 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.

**Purpose of Study**

This Harcourt Assessment research study was designed to collect empirical evidence related to the comparability and equivalence of scores obtained from the computer-based and paper-based administration modes of SDRT 4/SDMT 4.

Answers were sought to the following question:

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

**Methods**

**Instruments**

The framework of the SDRT 4 and SDMT 4 tests, including totals and subtests, is summarized in Table 1. The tests are offered in six levels that cover grades 1.5 through
13. Each level is named a specific color for ease of use by the instructor. For purposes of this study, these test levels are referred to as L1 through L6 and were administered to grades 2 (G2) through 12 (G12) as shown in Table 1.

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

<table>
<thead>
<tr>
<th></th>
<th>Red (L1)</th>
<th>Orange (L2)</th>
<th>Green (L3)</th>
<th>Purple (L4)</th>
<th>Brown (L5)</th>
<th>Blue (L6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SDRT 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Reading**</td>
<td>120</td>
<td>110</td>
<td>115</td>
<td>84</td>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>Phonetic Analysis</td>
<td>40</td>
<td>30</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Comprehension</td>
<td>40</td>
<td>40</td>
<td>45</td>
<td>54</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td><strong>SDMT 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Mathematics</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Concepts &amp; Applications</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Computation</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

* Computer-based and paper-based administration modes.
** The paper-based 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-based SDRT 4 in this study.

Norm-referenced and content-referenced scores 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 SDRT 4/SDMT 4. Additional information about SDRT 4/SDMT 4 development, standardization, and scoring can be found in the corresponding Teacher’s Manuals for Interpreting (Harcourt Assessment, 1995).

**Subjects and Data Collection**

Data from a total of 1,863 students taking SDRT 4 and 1,774 students taking SDMT 4 were analyzed to compare subtest and total subject area scores. Among these examinees, approximately 51% were female and 49% male. For the purposes of this study, examinees in each grade were randomly assigned to one of two groups. One group took the paper-based 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-based mode. There was a short break between the administrations of the two test modes for each student.

**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 one of the most powerful quantitative research methods for testing causal hypotheses (Gall, Borg, and Gall, 1996).
Data Analysis and Results

**Descriptive Statistics by Grade and Test Level**

Statistics describing the distributions of raw scores for SDRT 4/SDMT 4 totals and subtests were calculated. For SDRT 4, the differences between the mean scores for totals and subtests administered in either mode are less than 1.00 of the total raw score point for all grades and test levels with the exception of grade 2, where the difference is likely due to computer unfamiliarity for these young children. For SDMT 4, the differences between the mean scores for totals and subtests administered in either mode are less than 1.00 of the total raw score point for all grades and test levels. Where such differences exist, equating of test results will allow for equivalent score interpretations for computer-based and paper-based administrations.

In summary, the descriptive statistics calculated for all grades and test levels (with the exception noted above), demonstrate that scores obtained from either administration modes of SDRT 4/SDMT 4 are equivalent.

**Variability of Test Scores**

Reliability coefficients (Cronbach’s coefficients alpha) were calculated for SDRT 4/SDMT 4 totals and subtests. The reliability coefficient is an index of changes in the relative standing of the students in a group from one administration to another (Nitko, 2001). With respect to this study, the greater the reliability coefficient, the smaller the likelihood that students’ scores will differ from one administration to the next. That is, the test is reliable.

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**

Scatter plots, Pearson’s Correlation Coefficient ($r_p$), and Spearman’s rank order correlation coefficients ($r_S$) were calculated to determine consistency of test scores (i.e., reliability) on totals and subtests across administration modes. Overall, the scatter plots, Pearson’s Correlation Coefficient ($r_p$), 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-based 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-based test.

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. A probability value ($p$) was
calculated to determine if any statistically significant difference exist in the means of the total scores based on administration mode, mode order, or mode x mode order interactions. For SDRT 4 and SDMT 4, there were no 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 all grades and test levels, except L6, were comparable across administration mode and mode order, and there were no mode x mode order interactions for any grade or test level.

Summary and Conclusion

The results 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. The findings from this study support the use of the new computer based, online version of SDRT 4/SDMT 4. 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 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-based administrations.

References


