SUMMARY OF THE FINDINGS OF THE 1999-2000 SCREEN READING FIELD TEST

INCLUSIVE COMPREHENSIVE ASSESSMENT SYSTEM

May 2000

Pamela J. Brown, Associate Policy Scientist
Andy Augustine, Educator-in-Residence

Delaware Education
R&D
Research & Development
CENTER

Delaware Education Research & Development Center
University of Delaware
Newark, DE 19716

Publication T00-007.1

EXECUTIVE SUMMARY

The purpose of this research study was to determine if assessment items administered using screen reading software measure student learning better than assessment items in a paper and pencil format. Using a computer to present a test orally controls for standardization of administration and allows each student to complete the assessment at his/her own pace. Few published studies have used a computer to present a test orally (Burk, 1998).

In this study, 96 students completed a science assessment and 110 students completed a social studies assessment. One version was administered in the traditional paper and pencil format while the other version was administered via a computer utilizing screen reading software. The purpose of this study was to determine if the format of the assessment (screen reading vs. paper/pencil) differentially affected student performance. In order to compare student performance on the two versions of the assessment, a repeated-measures design using the general linear model (GLM) was used.

The results of the repeated-measures ANOVA revealed that for both the social studies and the science assessment, the students' reading score had a significant effect. This implies that a student's reading level confounds their assessment scores in the content areas of science and social studies. Format (screen reading versus paper/pencil) did not have a significant impact on the scores on this assessment when controlling for a student's reading ability. When selecting only "good readers," the science assessment reveals significant differences. When selecting only "poor readers," the same pattern emerges. When selecting only "average readers," there are no significant differences for either science or social studies.

While this study revealed no significant differences between the performance of students completing the pencil and paper format version versus the screen reading format when controlling for reading performance, using screen reading software as an accommodation in science for students with poor reading skills could be effective. It is likely that the limited numbers of significant results are compounded by the lack of appropriate instruction for students with poor reading skills. That is, if reading is the primary instructional method for students to learn concepts in the content areas of science and social studies, then students who performed poorly on these assessments, performed poorly because of lack of knowledge about science or social studies rather than inability to comprehend the test questions.

INTRODUCTION

The purpose of this research study was to determine if assessment items administered using screen reading software measure student learning better than assessment items in a paper and pencil format. This study is part of a larger study entitled the Inclusive Comprehensive Assessment System (ICAS) Project. The goal of the ICAS project is to evaluate various assessment methods or accommodations that maximize access to large-scale assessments by eliminating barriers in testing situations that are not relevant to the construct being measured. This study is specifically designed to evaluate the usefulness of screen reading software for assessments for students with reading difficulties as well as those without reading difficulties.

Several research studies on the K-12 student population have focused on the use of computer-based testing (CBT) which generally involves using a computer to administer a paper and pencil test. Other studies on the K-12 student population have focused on presenting the tests using audio cassettes, video cassettes, or human readers. The studies that explore the use of audio or video cassettes in a classroom permit a standard administration of the assessment. On the other hand, these devices generally are administered to an entire class of students and thus do not allow individual students to work at their own pace. Using a human reader also does not allow individual students to work at their own pace. In addition, using a human reader also presents other problems such as a lack of standardization of the assessment administration. Using a computer to present a test orally controls for standardization of administration and allows each student to complete the assessment at his/her own pace. Few published studies, however, have used a computer to present the test orally.

METHODOLOGY

Creation of the Assessments

For this study, four assessments were created and administered -- two in the area of social studies and two in the area of science. The assessments were comprised of publicly released NAEP (National Assessment of Educational Progress) items that were selected by several Delaware and Pennsylvania high school social studies and science teachers. Items on both versions of the assessment were matched for content area, process skill, and difficulty level assessed. In addition, the items were arranged in order of difficulty from the easiest to the most difficult.

Participants Selected

For this study, eighteen school districts in Delaware and three school districts in Pennsylvania were contacted to participate. Eleven high schools across eight school districts throughout Delaware and two school districts in Pennsylvania agreed to participate. Consent forms were distributed to all high school seniors (n = 2,593) as well as to their parents in each of these schools. Less than one-fourth (13.6%) of the parents and students returned the consent forms after two mailings. Most parents (74.2%) who returned the consent forms gave their consent, but some of these students were unable to participate due to absenteeism or withdrawal from school. The sample included students who had reading difficulties (as measured by a standardized reading test) as well as students that did not have reading difficulties. Table 1 contains information about the reading level of the participants. For Delaware students their 10th grade DSTP reading score was used to determine their reading level.

Table 1
Reading Level of Students (as measured by national standardized tests) Who Completed the Assessment by Content Area

Content	Range of Reading Percentile	Mean Reading Percentile	Standard deviation	Total Sample Size
Science	5-99	57.23	26.88	96
Social Studies	1-99	55.08	27.08	110

Research Design

To ensure that there were no order effects, half of the students began with Version A and finished with Version B while the other half began with Version B and finished with Version A. Table 2 presents the research design used.

Administration of the Assessments

Ninety-six students completed the science assessment and 110 students completed the social studies assessment. Each version consisted of a variety of grade-appropriate multiple choice and open-ended items. One version was administered in the traditional paper and pencil format while the other version was administered via a computer utilizing screen reading software. Authorware 5.0 was the software package used for the administration of the screen reading portion of this study. All students completed both versions of the assessment so as to serve as their own control for this study. This controls for the impact of extraneous variables such as race, gender, age, and SES on the results of this study.

Table 2
Number of Students to Participate in Research Study

Content Area	Format Completed First	
	Paper/Pencil	Screen Reading
Social Studies	50	50
Version A in paper/pencil format AND Version B in screen reading format	25	25
Version A in screen reading format AND Version B in paper/pencil format	25	25
Science	50	50
Version A in paper/pencil format AND Version B in screen reading format	25	25
Version A in screen reading format AND Version B in paper/pencil format	25	25
Total	100	100

Screen reading software permitted the student to listen via a headset to the test items as they were displayed on the computer screen. Each student could choose to listen to any assessment item multiple times. Students selected an answer for the multiple-choice items by using the mouse to click on option A, B, C, or D. For the open-ended items, students typed their answer into a text box on the screen.

Each correct response to a multiple choice item received one point while the open-ended item was scored using a 3-point or 4-point rubric. A total score was calculated by summing the scores received for each item on the assessment. Table 3 provides a summary of the type of items on each assessment administered.

The purpose of this study was to determine if the format of the assessment (screen reading vs. paper/pencil) differentially affected student performance. In order to compare student performance on the two versions of the assessment, a repeated-measures design using the general linear model (GLM) was used. The within- subjects factor was the students' scores on the assessments while the order in which they took the assessments (version and format) were the between-subjects factors. The percentile rank on the reading portion of a national standardized test served as the covariate. Furthermore, a series of t-tests were used to explore score differences based on format and version, and a regression analysis was conducted to determine if a student's reading score was useful in predicting a student's science or social studies assessment score.

Table 3
Description of Mathematics and Science Assessments Administered

	Version	Type of Items	Number of Items	Total Score Possible
Social Studies	A	Open-Ended	5	15
		Multiple Choice	13	13
	B	Open-Ended	5	16
		Multiple Choice	12	12
Science	A	Open-Ended	2	6
		Multiple Choice	31	31
	B	Open-Ended	2	6
		Multiple Choice	30	30

Scoring Process for the Open-Ended Items

Each open-ended item was scored by a rater using the rubric that accompanied the NAEP assessment item. The raters for the items had strong backgrounds in the appropriate content area. Since the rubrics were straightforward (see Figures 1 & 2), only one rater was used to score each item. However, to control for bias, the same rater scored all assessments for a particular item.

Figure 1.
Example of Scoring Guide for a Science Item

3 = Complete - student response describes two ways in which heart disease can be prevented, such as those below.

2 = Partial - student response describes one way in which heart disease can be prevented.

1 = Unsatisfactory/Incorrect - student response shows no understanding of how heart disease can be prevented.

Credited responses include: getting more exercise, regular exercise; reducing stress/relaxing; eating less saturated fat/avoiding greasy food

Figure 2.
Example of Scoring Guide for a Social Studies Item

3 = Appropriate - These answers explain the link between a factor and suburbanization, citing specifics or elaborating on the explanation.

2 = Partial - These answers suggest a linkage between a factor and suburbanization, but it is vague and lacks specifics.

1 = Inappropriate - These answers do not address the linkage between any factors and the growth of suburbs.

Credited responses could include:

- automobiles and highways enabled people to move further away from places where they work and shopped, encouraging the growth of communities (suburbs) at some distance from workplace, from which people can commute.
- tax deductions enabled more people to buy homes, which led to the rapid growth of suburban areas (sprawl)

Reliability Analysis

In the tables below is a summary of the reliability statistics for the two versions of the social studies and the science assessments. Reliability statistics are given for each assessment as a whole as well as for the multiple-choice questions only. Since there are fewer items on the social studies assessment than the science assessment, one would expect lower reliability statistics on the social studies assessment.

Reliability Statistics (Coefficient Alpha) for the Social Studies Assessment

	Version A	Version B
Multiple choice items only	.64	.43
All assessment items	.79	.71

Reliability Statistics (Coefficient Alpha) for the Science Assessment

	Version A	Version B
Multiple choice items only	.87	.83
All assessment items	.87	.83

Because of the low reliability of the multiple-choice items on the social studies assessments, additional analyses of the multiple-choice section of the assessment were conducted for the science assessment, but not the social studies assessment.

FINDINGS

The results of the repeated-measures ANOVA revealed that for both the social studies and the science assessment, the students' reading score had a significant effect. This implies that a student's level of ability in reading confounds their assessment scores in the content areas of science and social studies. Format (screen reading versus paper/pencil) did not have a significant impact on the scores on this assessment when controlling for a student's reading ability. The results of these tests are shown in Tables 4 - 5.

For the science assessment, however, there was also a significant interaction between the performance on the assessment and the format of the assessment (see Figure 1). The interaction

Table 4
ANOVA for a Repeated-Measures Design for the Social Studies Assessment

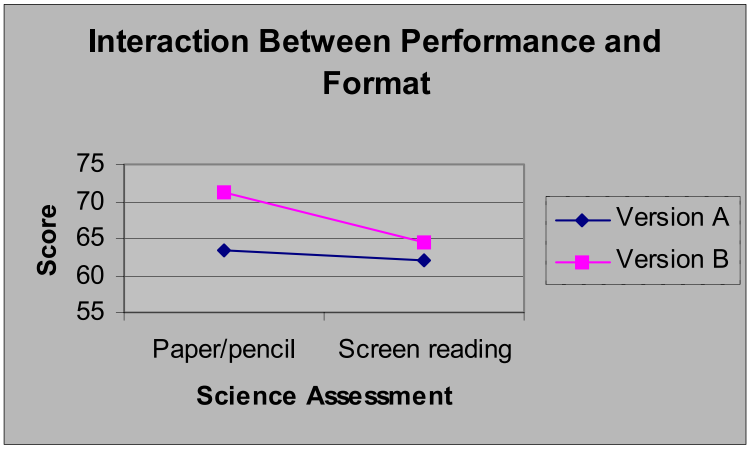
Source	df	F
	Between Subjects	
Reading Percentile (R)	1	36.110**
Order of Format (F)	1	1.216
Order of Version (V)	1	.051
V*F	1	1.264
error	80	(22.831)
	Within Subjects	
Total Score (TS)	1	.012
TS*R	1	1.074
TS*F	1	.750
TS*V	1	3.084
TS*V*F	1	.993
error	80	(9.737)

Table 5
ANOVA for a Repeated-Measures Design for the Science Assessment

Source	df	F
Between Subjects		
Reading Percentile (R)	1	25.77**
Order of Format (F)	1	3.70
Order of Version (V)	1	.011
V*F	1	.880
error	81	(390.84)
Within Subjects		
Total Score (TS)	1	.10
TS*R	1	1.73
TS*F	1	6.13*
TS*V	1	.78
TS*V*F	1	.03
error	81	(81.35)

Note. Values enclosed in parentheses represent mean square errors.
* p < .05 ** p < .01

Figure 1



indicates that students performed better using the paper-pencil format than the screen reading format for version B regardless of the order completed.

To illuminate these findings, separate t-tests were conducted for “good readers, “average readers,” and “poor readers” in both science and social studies. For this study “good readers” are defined as those students who score at or above the 70th percentile. “Average readers” are defined as students who score above the 50th and below the 70th percentile. “Poor readers” are defined as those students who score at or below the 50th percentile. Initially, a t-test was also conducted between the two scores on the assessments. This test was to identify if there were any differences between the two versions (A and B) or the format (paper/pencil and screen reading) of the assessments. These results are presented in Tables 6 - 9.

The initial t-test showed that there are significant differences in scores by version in both science and social studies. This difference may due to inequivalency between the two versions or an interaction between version and format of the assessments.

When selecting only “good readers” for this same analysis, the science assessment still reveals significant differences; however, for social studies there are no significant differences between the two scores. When selecting only “poor readers,” the same pattern emerges. The science assessment shows significant differences between the two scores. When selecting only “average readers,” there are no significant differences for either science or social studies. So in conclusion, all readers did at least as well, or in most cases better, on Version B of the science assessment.

Table 6
Overall Paired Samples T-test Results

Content	Mean Difference	Standard Error of Mean	df	t
Science	-5.04	1.32	92	-3.83**
Social Studies	-3.53	1.63	91	-2.16*

Table 7
Paired Samples T-test Results for “Good Readers”

Content	Mean Difference	Standard Error of Mean	df	t
Science	-9.88	1.77	26	-5.60**
Social Studies	-4.81	3.70	25	-1.30

Table 8
Paired Samples T-test Results for “Poor Readers”

Content	Mean Difference	Standard Error of Mean	df	t
Science	-4.40	2.17	34	-2.03*
Social Studies	-2.86	2.54	34	-1.13

Table 9
Paired Samples T-test Results for “Average Readers”

Content	Mean Difference	Standard Error of Mean	df	t
Science	-.01	3.16	23	-.004
Social Studies	-3.27	2.99	23	-1.10

* $p < .05$
** $p < .01$

The regression analysis revealed that for the social studies assessment as well as the science assessment, the students reading score was a significant predictor of their performance. Those students who had high reading scores tended to score well on these assessments regardless of the format. In the case of the social studies assessment, this regression model predicts almost 27% of the variance of the scores. With the science assessment, this model predicts about 19% of the variance of the scores. The results of these analyses are presented in Tables 10 and 11.

Table 10

Summary of Regression Analysis for Variables Predicting Total Score on Social Studies Assessment

Variable	B	SE B	ß
Reading Percentile	.073	.015	.470**
Version	-3.02	3.74	-.08
Format	-1.16	.82	-.14

Note. $R^2 = .266$, ** $p < .01$

Table 11

Summary of Regression Analysis for Variables Predicting Total Score on Science Assessment

Variable	B	SE B	ß
Reading Percentile	.260	.060	.437**
Version	-.89	3.24	-.03
Format	-2.69	3.20	-.08

Note. $R^2 = .190$, ** $p < .01$

Table 12
Mean Score Percentages (and Standard Deviation) on Assessments

	Paper and Pencil Version	Screen Reading Version
Social Studies	63.48 (15.22)	59.75 (17.94)
Science	65.49 (17.43)	65.24 (17.83)

Additional Exploratory Analysis of the Multiple-Choice Items on the Science Assessment

Descriptive statistics were calculated for the multiple-choice section of the science assessment. These results are shown in the tables below.

Version A of the Science Assessment

Format	Sample size	Minimum percent earned	Maximum percent earned	Mean	Standard Deviation
Paper & Pencil	45	17.0	96.7	69.9	20.0
Screen Reading	47	0.0	100.0	70.5	19.9

Version B of the Science Assessment

Format	Sample size	Minimum percent earned	Maximum percent earned	Mean	Standard Deviation
Paper & Pencil	48	22.6	93.6	60.9	17.4
Screen Reading	46	22.6	100.0	59.9	19.0

Due to the extremely small differences between the means using the pencil & paper format compared to the screen reading format, no t-tests were necessary to identify any significant differences. However, due the large variance in the scores within a format, additional analyses were conducted to identify if any differences existed when selecting only "good readers," "average readers," or "poor readers." As described earlier, "good readers" are defined as those students who scored at or above the 70th percentile on a standardized test. "Poor readers" are defined as those students who scored at or below the 50th percentile. "Average readers" are defined as those students who scored above the 50th percentile, but below the 70th percentile. The results of these analyses by sub-group of student are listed in the following two tables.

<u>Science Assessment -- Version A</u>

	Good Readers		Average Readers		Poor Readers	
Format	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)
Paper & Pencil	12	80.3 (13.9)	11	72.4 (19.7)	20	62.5 (20.1)
Screen Reading	14	77.4 (17.4)	11*	76.4 (9.6)	15	68.0 (10.6)

* Two outliers removed before calculating mean and standard deviation.

Science Assessment -- Version B

	Good Readers		Average Readers		Poor Readers	
Format	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)
Paper & Pencil	15	68.0 (16.6)	13	63.8 (18.06)	15	52.9 (16.2)
Screen Reading	13	65.8 (20.6)	11	62.8 (18.1)	20	54.8 (15.6)

Although there are no significant differences between the performance on the paper & pencil format compared to the screen reading format, poor readers did score higher using the screen reading format (5.5% on version A and 1.9% on version B). Good readers, however, scored higher using the paper/pencil format (2.9% on version A and 2.2% on version B). Overall, regardless of the format, good readers performed better than average readers who performed better than poor readers.

Correlational Analyses for the Science Assessment

Before calculating the Pearson product-moment correlations, scatterplots for each version of the assessment by format were graphed. The percent of items answered correctly on the multiple-choice section of the assessment were plotted on the x-axis while the reading percentile was plotted on the y-axis. A reference line was drawn on the x-axis at 25% to determine the floor of the assessment, i.e., the score a student would receive by chance alone since there were four options in most cases for the multiple choice items. Two references lines were drawn on the y-axis to divide the cases into "good readers," "average readers," and "poor readers" as previously defined. These four scatterplots are presented on the following two pages.

On version A using the screen reading format, two outliers were identified. These two cases clearly do not fall within the pattern shown by the other cases. Therefore, these two outliers were removed before calculating the correlations.

Pearson Product-Moment Correlations were calculated between reading percentile score and total score on the multiple-choice section for each version of the science assessment. These statistics are presented in the table below. For both versions of the assessment, the correlations were in the positive direction and were significantly different from zero at .01 for the screen reading version and at .001 for the paper/pencil version. However, for both version A and B, these correlation coefficients did not differ significantly based on the format used.

Version	Format	Correlation Coefficient	sample size
A	Paper/Pencil	.527	43
A	Screen Reading	.444	40
B	Paper/Pencil	.472	43
B	Screen Reading	.412	44

Therefore, there is a significant relationship between a student's reading performance and his/her score on the science assessment. While there is not a significant difference between the correlation coefficient for the paper & pencil format compared to the screen reading format, the coefficient for the paper & pencil format is slightly larger implying a stronger relationship.

SUMMARY

This study revealed no significant differences between the performance of students completing the pencil and paper format version versus the screen reading format when controlling for reading performance. However, using screen reading software as an accommodation in science for students with poor reading skills may still be effective. While it appears that using screen reading software does not eliminate the reliance on reading, there is evidence to suggest that it may reduce its impact. However, it is likely that the limited numbers of significant results are compounded by the lack of appropriate instruction for students with poor reading skills. That is, if reading is the primary instructional method for students to learn concepts in the content areas of science and social studies, then students may perform poorly on assessments because of a lack of knowledge about science or social studies rather than an inability to comprehend the test questions. To tease out this factor (primary method of instruction), one would need to secure a sample of students who have been instructed using methods that do not require the students to learn primarily by reading, such as instruction using primarily hands-on activities.

Perhaps with social studies, reading was so confounded with their score that any version differences were undetectable, irrelevant, or nonexistent. In science, reading was important, but not so important that version differences could not be detected. Thus, building science assessment forms carefully based on process skills and difficulty level may not be sufficient to claim form equivalence.

References

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential items performance for nine handicapped groups. *Journal of Education Measurement,* 24(1), 44-55.

Burk, M. (1998, October). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities.* Paper presented at Office of Special Education Program Cross Project Meeting "Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century.

Curtis, H. A., & Kropp, R. P. (1961). A comparison of scores obtained by administering a test normally and visually. *Journal of Experimental Education,* 29, 249-260.

Epsin, C. A., & Sindelar, P. T. (1988). Auditory feedback and writing: Learning disabled and nondisabled students. *Exceptional Children,* 55, 45-51.

Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of nondisabled students under silent reading and reading plus listening modes of administrations. *Applied Measurement,* 6, 307-320.

Hasselbring, T. S., & Crossland, C. L. (1982). Application of microcomputer technology to spelling assessment of learning disabled students. *Learning Disability Quarterly,* 5, 80-82.

Helwig, R., Tedesco, M., Health, B., Tindal, G., & Almond, P. (1998). *The relationship between reading ability and performance on a video accommodated math problem-solving test.* Manuscript submitted for publication, University of Oregon.

Horton, S. V., & Lovitt, T. C. (1994). A comparison of two methods of administering group reading inventories to diverse learners. *Remedial and Special Education,* 15, 378-390.

Keene, S., & Davey, B. (1987). Effects of computer-presented text on LD adolescents' reading behaviors. *Learning Disability Quarterly,* 10, 283-290.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing.

Miller, S. (1998). *The relationship between language simplification of math word problems and performance for students with disabilities.* Unpublished master's project, University of Oregon, Eugene, OR.

Swain, C. R. (1997). A comparison of computer-administered test and a paper and pencil test using normally achieving and mathematically disabled young children (Doctoral dissertation, University of North Texas, 1997). *Dissertation Abstracts International,* 58, 0158.

Tachibana, K. K. (1986). Standardized testing modifications for learning disabled college students in Florida (modality) (Doctoral dissertation, University of Miami, 1986). *Dissertation Abstracts International,* 47, 0125.

Tindal, G., Almond, P., Heath, B., & Tedesco, M. (1998). *Single subject research using audio cassette read aloud in math.* Manuscript submitted for publication, University of Oregon.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations.* Unpublished manuscript with Mid-South Regional Resource Center, University of Kentucky.

Tindal, G., Glasgow, A., Helwig, B., Hollebeck, K., & Heath, B. (1998). *Accommodations in large scale tests for students with disabilities: An investigation of reading math tests using video technology.* Unpublished manuscript with Council of Chief State School Officers, Washington, DC.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilites on large-scale tests: An emperical study of student response and test administration demands. *Exceptional Children,* 64(4), 439-450.

Trimbal, S. (1998). *Performance trends and use of accommodations on a statewide assessment* (Maryland/Kentucky State Assessment Series Rep. No. 3). Minneapolis, MN: National Center on Educational Outcomes.

Varnhagen, S., & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly,* 7, 266-270.

Watkins, M. W., & Kush, J. C. (1988). Assessment of academic skills of learning disabled students with classroom microcomputers. *School Psychology Review,* 17, 81-88.

Westin, T. (April, 1999). *The validity of oral presentation in testing.* Montreal, Canada: American Educational Research Association.