DOCUMENT RESUME

ED 357 081 TM 019 860

AUTHOR Siskind, Theresa G.; And Others
TITLE The Instructional Validity of Computer Administered Tests.
PUB DATE Mar 92
NOTE 21p.; Paper presented at the Annual Meetings of the Eastern Educational Research Association (15th, Hilton Head, SC, March 5-9, 1992) and the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Testing; Comprehension; Computer Assisted Instruction; *Computer Assisted Testing; High Schools; *High School Students; *Instructional Effectiveness; Performance Tests; Scores; *Teaching Methods; Test Format; Test Items; *Test Validity
IDENTIFIERS Item Characteristic Function; *Paper and Pencil Tests

ABSTRACT

The instructional validity of computer administered tests was studied with a focus on whether differences in test scores and item behavior are a function of instructional mode (computer versus non-computer). In the first of 3 studies, performance test scores for approximately 400 high school students in 1990-91 for tasks accomplished with the computer were correlated with objective paper-and-pencil test scores for comprehension of the same tasks. In the second study, a comparison was made between test scores of 77 high school students in 1991-92 taking an objective test via computer and scores on the same test in paper-and-pencil format. In the third study, the item characteristics of tests given in a computer format were compared to the item characteristics of the same tests presented in a paper-and-pencil format. Data from the first two studies were used. Findings from the three studies indicate that performance tests for computing are not equivalent to objective paper-and-pencil tests even when the content is the same. Test scores and item statistics for tests in a paper-and-pencil format and computer format do not differ for content taught in a combined computer and lecture mode. Whether students not taught by computer are hampered by computer-based tests is a logical extension of this research. Seven tables present study findings. (SLD)

The Instructional Validity of Computer Administered Tests

Theresa G. Siskind
Assistant Professor
The Citadel
(803) 792-7824

Elsie C. Andrews
Summerville High School, Gregg Campus

Susan Kovas
The Citadel

Paper presented at the annual meeting of the Eastern Educational Research Association, March 6, 1992.

The Instructional Validity of Computer Administered Tests

Previous studies have indicated that computer tests require less administration time than their paper-and-pencil counterparts, provide immediate feedback to students, and allow teachers to compute test statistics easily (Fletcher & Collins, 1986; Marso & Pigge, 1987; Olsen, Weiss, & Langford, 1989; Russell, Peace, & Mellsop, 1986; Stiggins, Conklin, & Bridgeford cited in Hsu & Yu, 1989). Computer tests have also been touted as less sloppy than the typical teacher made test (Fleming & Chambers cited in Marso & Pigge, 1987).

Research about differences in test scores for computer versions and paper-and-pencil test versions has resulted in conflicting findings (Bunderson, Inouye, & Olsen, 1989; Mazzeo & Harvey, 1988). In cases where paper-and-pencil tests resulted in higher scores, the limitations of the computer format may have influenced the scores. For example, scores on mathematics tests of ratio and proportion (Ronau & Battista, 1988) and mathematical reasoning tests (Lee & Hopkins, 1985; Lee, Moreno & Sympson, 1984) seem to be somewhat dependent upon "scratchwork" space - a commodity not easily accommodated by computer testing.

The biggest drawback cited for computer tests is the inability, in some applications, to review and revise responses (Fletcher & Collins, 1986; Wise & Plake, 1989).

Although computer anxiety appears to have little or no relationship to test performance (Seymour & Others, 1986; Wise, Barnes, Harvey, & Plake, 1989), it seems reasonable that

familiarity with the hardware would have some impact on examinee performance. The notion of instructional validity, which emphasizes the match between instruction and testing, would seem to be an important consideration with computer testing. And, further, it seems that instructional validity might be influenced by the "format" of instruction as well as the content.

The purpose of the present paper is to investigate the instructional validity of computer administered tests. The focus of the present study differs from other studies of computer testing in that the primary question is whether differences in test scores and item behavior are a function of instructional mode (computer versus non-computer).

Methods

The current study actually encompasses three studies/approaches to curricular validity. In the first study, performance test scores (for tasks accomplished via computer) are correlated with objective paper-and-pencil test scores (testing comprehension of the same tasks as accomplished on the computer). In the second study, a comparison is made between the test scores of pupils taking an objective test via computer and pupils taking the same test in paper-and-pencil format. And in the third study, the item characteristics of tests given in computer format are compared to the item characteristics of the same tests presented in paper-and-pencil format.

Study 1: The Correlation of Performance and Objective Tests

Subjects. The sample consisted of approximately 400 high school students who were enrolled in a semester course, "Introduction to Computers," during the 1990-91 school year. Twenty different classes taught by two different teachers comprised the sample.

Procedure. The two teachers planned together so that instruction and testing were standardized across the teacher and class combinations. During the second half of the course, the computer applications of word processing, databases and spread sheets were taught. Instruction was given by guiding the students through the computer applications. Each student worked on a computer while the teachers instructed students about the procedures. The students were tested on each of these applications in two ways. One test was an objective, paper-and-pencil test assessing students' knowledge of the application. The other test was a computer, performance test in which the student actually had to apply the same concepts tested on the objective test. The two tests were administered during consecutive class periods with the paper-and-pencil test scheduled first. The scores on these pairs of tests were correlated and the results appear in Table 1. Means and standard deviations for the tests are given in Table 2.

Insert Tables 1 and 2 about here

Findings and Discussion. It is interesting to note that the correlations between the pairs of tests measuring like content (0.24, 0.39, 0.46), with the exception of the last unit on spread sheets, is generally weaker than the correlations between the same types of tests. The objective test pairs showed correlations of 0.48, 0.51 and 0.61 while the applications pairs displayed correlations of 0.45, 0.42 and 0.42. One of the higher correlations (0.53) is between the database performance test and the spreadsheets objective test.

These data provided limited support for the notion that objective tests and performance tests for computer applications do not measure the same skills. If one's goal is for the students to be able to execute the application, an application test would appear to be a more accurate measure.

Study 2: A Comparison of Test Scores across Formats

Subjects. The sample consisted of 87 high school students who were enrolled in a semester course, "Introduction to Computers," during the fall semester of the 1991-92 school year. The students were instructed by the same teacher during four different class periods.

Based on a survey administered to 77 of the subjects near the end of the semester, one was in eighth grade, 43 were in ninth grade and 33 were tenth graders. The table below provides the gender and ethnic breakdown of the sample.

	Black	White	Other
Male	5	37	2
Female	15	16	2

Although enrolled in an introductory computer class, 62 of the 77 subjects indicated that they had used computers previously. Thirty-eight had computers at home. When queried about their interest in using computers, 32 responded that they always found computers interesting to use and 26 said they frequently found computer use interesting. Conversely, eight indicated it was always a chore to use the computer and 14 said that it was frequently a chore. Twenty-three of the 77 students indicated that they would always make a better grade on a computer test than a paper-and-pencil test while nine said they would never make a better grade on a computer test. Thirty-five said they would always like to take tests like the Scholastic Aptitude Test (SAT) on the computer.

Procedures. During the first part of the course, students were introduced to the history of computing, computer functions and basic computer operations. The type of material covered ranged from classes of computers to types of hardware to operating systems. Instruction was provided through lecture and computer-based tutorials.

Students were tested six times on this material (Chapter 1, Chapter 2, Chapter 3, DOS, Chapters 4&5, Chapter 7). For five of the six tests, students were assigned (a priori) to either a

condition of computer testing or a condition of paper-and-pencil testing. The tests in the two conditions were identical and differed only in the mode of administration. The remaining test administration was standardized across students. All students took the Chapter 2 test in the paper-and-pencil format.

All tests were taken in the regular classroom and students sat in their usual seats. Students who took the computer versions of tests used the same assigned computer that they worked with daily. Students taking the computer tests could change their answers at any point during testing and could review all of their answers at the end of the test. All tests were hand-scored by the teacher and all students received their scores at the same time. (No one received immediate feedback.) Overall scores were compared between the conditions.

Instruments. To establish content validity of the tests, two of the tests (Chapter 4&5 and Chapter 7) were compared to the list of instructional objectives provided by the teacher. An independent observer found a 100% match between the test content and instructional content. Reliability of the Chapter 7 test was computed at .97 using a split-half procedure corrected with the Spearman-Brown prophecy formula. Inter-rater reliability was verified by an independent observer who re-scored a sampling of all of the tests. While most of the test questions were objective, there were a few short essay questions. The teacher's key included specific instructions so that these portions could be accurately evaluated by the independent scorer.

Findings and Discussion. Initial analyses comparing the results on the computer versions of the tests with the results on the paper-and-pencil versions seemed to be confounded by class period so an additional series of two-way analyses were performed using both class period and type of test as the independent variables. Table 3 provides the means and standard deviations for these comparisons while Table 4 presents the findings from the general linear models analysis. The DOS test was eliminated from analysis because students were allowed to retake this test due to low scores. Not all low-scoring students, however, availed themselves of this opportunity.

Insert Tables 3 and 4 about here

Although Test 2 was taken by all students in paper-pencil format, it was included in the analysis to test the hypothesis that different class periods performed differently on the tests. For Tests 2 and 3 significant differences were found by class period. Although not statistically significant at the .05 level, differences were found on Tests 1 and 4&5 as well. Students in the last period class appear to perform less well in general than students in the other classes.

The significant interaction on Test 3 is probably due primarily to class period differences and to spurious findings due to the two late (make-up) examinees that took paper-and-pencil tests in the fourth and fifth period classes. The

significant difference for type of test on Test 4&5 is similarly affected by the late examinees.

In summary, there appear to be no true differences in the test scores of examinees who take tests in the computer format and examinees who take tests in the paper-and-pencil format in classes where instruction is delivered by a combination of computer and lecture. This does not preclude differences that might occur if instruction were given in one mode only.

Study 3: Item Analysis Across Test Formats

Procedures. In this study, a more indepth analysis of two of the tests from Study 2 was performed. Utilizing traditional and one-parameter latent trait analyses, the response patterns for students taking the computer version of Tests 4&5 and 7 were compared to the patterns for the students completing the tests in paper format.

Instruments. The test on Chapters 4&5 assessed students knowledge about computer input and output devices and media. In addition to key terms introduced in the chapter, the 42-item test required students to identify types of input and output devices and media, to be familiar with the data entry process, and to be familiar with careers related to data entry and input/output control. Thirty-five of the 42 items were objective in nature.

The test on Chapter 7 assessed students knowledge about microcomputer systems. The 47-item test required students to distinguish between personal and professional microcomputers, to identify input and output devices used with microcomputers, to

understand the components and functions of the Central Processing Unit, and to identify storage devices used with microcomputers. Forty of the 47 items were objective.

Findings and Discussion. For each of the tests, an item analysis was performed on the objective items using the Iteman and Rascal programs of the MicroCat system, version 3.0. Table 5 reports a summary of traditional item statistics for the tests, and Tables 6 and 7 report the Rasch item difficulty values.

Insert Tables 5-7 about here

Although differences in item statistics were not tested for significance, it is apparent that there are no meaningful differences across the types of tests. These data tend to substantiate the findings of Study 2.

Conclusions

Despite limitations in the designs, these three studies represent initial attempts to evaluate computer based tests in terms of instructional validity. In an introduction to computing class, much of the instruction utilizes the computer format. It seems appropriate to test students using the computer format as well. In light of the instructional mode, is the computer testing mode superior? The findings from these three studies indicate that (1) performance tests for computing are not equivalent to objective, paper-and-pencil tests even when the content is the same, and (2) test scores and item statistics for

tests taken in paper-and-pencil format and computer format do not differ for content taught in a combined computer and lecture mode. While not covered in the present studies, a logical extension of this question is the question of whether or not students who have not been instructed via computer are hampered by computer-based tests.

References

Bunderson, C. V., Inouye, D. K, & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed, pp. 367-408). New York: Macmillan.

Fletcher, P. & Collins, M. A. J. (1986). Computer-administered versus written tests - advantages and disadvantages. Journal of Computers in Mathematics and Science Teaching, 6(2), 38-43.

Hsu, T. & Yu, L. (1989). Using computers to analyze item data response. Educational Measurement: Issues and Practice, 8(3), 21-28.

Lee, J. A., & Hopkins, L. (1985, March). The effects of training on computerized aptitude test performance and anxiety. Paper presented at the 56th annual meeting of the Eastern Psychological Association, Baltimore, MD. (ERIC Document Reproduction Service No. ED 246 093)

Lee, J. A., Moreno, E. F., & Sympson, J. R. (1984, April). The effects of mode of test administration on test performance. Paper presented at the 55th annual meeting of the Eastern Psychological Association, Baltimore, MD. (ERIC Document Reproduction Service No. ED 263 889)

Marso, R. N., & Pigge, F. L. (1987, October). Teacher-made tests and testing: Classroom resources, guidelines, and practices. Paper presented at the annual meeting of the

Midwestern Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 291 781)

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

Olsen, B., Weiss, J., & Langford, D. (1989). Using computerized testing to improve learning. Principal, 69(2), 13-15.

Ronau, R. N., & Battista, M. T. (1988). Microcomputer versus paper-and-pencil testing of student errors in ratio and proportion. Journal of Computers in Mathematics and Science Teaching, 7(3), 33-38.

Russell, G. K. G., Peace, K. A., & Mellsop, G. W. (1986). The reliability of a micro-computer administration of the MMPI. Journal of Clinical Psychology, 42(1), 120-122.

Seymour, S. L., et al. (1986, January). Microcomputers and continuing motivation. Paper presented at the annual convention of the Association for Educational Communications and Technology, Las Vegas, NV. (ERIC Document Reproduction Service No. ED 267 791)

Wise, S. L., & Plake, B. S. (1989). Using computers to analyze item data response. Educational Measurement: Issues and Practice, 8(3), 5-10.

Table 1

Intercorrelation Matrix and Number of Subjects Tested (N)
for Objective (O) and Performance (P) Test Scores
on Word Processing (WP), Databases (DB) and Spread Sheets (SS)

	WP-O	WP-P	DB-O	DB-P	SS-O	SS-P
WP-O	1.00	0.24	0.48	0.33	0.51	0.17
	(418)	(390)	(396)	(348)	(389)	(354)
WP-P		1.00	0.35	0.45	0.49	0.42
		(399)	(379)	(342)	(374)	(351)
DB-O			1.00	0.39	0.61	0.30
			(407)	(350)	(385)	(351)
DB-P				1.00	0.53	0.42
				(356)	(342)	(326)
SS-O					1.00	0.46
					(399)	(350)
SS-P						1.00
						(362)

Table 2

Descriptive Statistics

for Objective (O) and Performance (P) Test Scores

on Word Processing (WP), Databases (DB) and Spread Sheets (SS)

Test	N	Mean	Standard Deviation
WP-O	418	84.82	13.77
WP-P	399	89.20	13.80
DB-O	407	81.08	15.26
DB-P	356	84.21	21.34
SS-O	399	79.72	17.19
SS-P	362	89.47	18.88

Table 3

Means, Standard Deviations, and Numbers of Students for Tests 1, 2, 3, 4&5, and 7 by Class Period and Type of Test

		PERIOD 2		PERIOD 3		PERIOD 4		PERIOD 6	
TEST		PP	COMP	PP	COMP	PP	COMP	PP	COMP
1	X	80.0	76.3	80.4		66.3	73.4	70.7	71.0
	SD	4.2	14.7	12.9		16.3	19.2	19.2	17.4
	N	2	22	24		10	9	10	9
2	X	73.6		76.1		78.1		61.7	
	SD	12.9		14.3		10.8		21.6	
	N	23		24		19		19	
3	X	85.9		85.4		96.0	76.1	34.0	77.8
	SD	4.2		12.9			23.7		15.4
	N	22		23		1	15	1	17
4&5	X	88.6		90.8			89.4	70.5	89.6
	SD	12.2		11.4			8.4	.7	9.6
	N	23		22			15	2	16
7	X	77.3		61.5	75.5		76.8	70.5	
	SD	14.7		2.1	17.0		16.3	18.3	
	N	22		2	20		16	16	

Table 4

Results of Analysis of Test Scores

Based on Type of Test and Class Period

Test	Sum of Squares	df	Mean Square	F	p
Test 1					
Period	1748.43	3	582.81	2.40	0.07
Type Test	14.28	1	14.28	0.06	0.81
Interaction	196.20	2	98.10	0.40	0.67
Test 2					
Period	3121.47	3	1040.50	4.47	0.006
Test 3					
Period	2530.20	3	863.40	3.41	0.02
Type Test	268.70	1	268.70	1.09	0.30
Interaction	1904.95	1	1904.95	7.71	0.007
Test 4&5					
Period	757.01	3	252.34	2.21	0.09
Type Test	650.25	1	650.25	5.70	0.02
Test 7					
Period	765.63	3	255.21	0.95	0.42
Type Test	353.82	1	353.82	1.32	0.25

Table 5

Comparison of Traditional Item Statistics for Tests 4&5 and 7

by Type of Test (Computer or Paper-and-Pencil)

Test 4&5	Paper-and-Pencil	Computer
N of Items	35	35
N of Examinees	40	29
Mean	31.725	30.586
Variance	11.749	11.691
Std. Dev.	3.428	3.419
Skew	-2.503	-0.784
Kurtosis	8.640	0.161
Minimum	16.000	21.000
Maximum	35.000	35.000
Median	32.000	31.000
Alpha	0.791	0.728
SEM	1.568	1.784
Mean P	0.906	0.874
Mean Item-Tot	0.418	0.874
Mean Biserial	0.651	0.533

Test 7	Paper-and-Pencil	Computer
N of Items	40	40
N of Examinees	37	35
Mean	31.108	31.371
Variance	40.205	33.776
Std. Dev.	6.341	5.812
Skew	-0.750	-0.328
Kurtosis	-0.095	-1.262
Minimum	15.000	21.000
Maximum	40.000	39.000
Median	31.000	33.000
Alpha	0.869	0.845
SEM	2.293	2.286
Mean P	0.778	0.784
Mean Item-Tot	0.407	0.361
Mean Biserial	0.583	0.536

Table 6

Rasch Item Difficulty Estimates for Test 4&5

Comparing Paper-and-Pencil and Computer Tests

Test Item	Paper-and-Pencil	Computer
1	--Deleted--	0.229
2	--Deleted--	-0.595
3	0.374	-0.127
4	-1.579	-1.340
5	-1.579	-0.595
6	0.624	-0.595
7	0.624	-1.340
8	-0.782*	0.229
9	-0.782	-1.340
10	0.079	1.010
11	0.374	0.779
12	--Deleted--	-1.340
13	--Deleted--	-0.595*
14	1.850	2.173
15	-0.782	--Deleted--
16	--Deleted--	--Deleted--
17	--Deleted--	1.224
18	1.222	1.806
19	0.843	1.426
20	-1.579	-0.595
21	0.079	0.523
22	0.374	1.010
23	--Deleted--	--Deleted--
24	0.079	-1.340
25	1.041	-0.127
26	0.079	--Deleted--
27	1.222	1.010
28	-1.579	-0.595
29	-0.782	-0.595*
30	0.079	0.523*
31	0.843	0.229
32	0.624	-0.127
33	-0.782	-0.595
34	1.391	1.010
35	-1.579	-1.340

*Significant Pearson chi-square lack of fit

Table 7

Rasch Item Difficulty Estimates for Test 7

Comparing Paper-and-Pencil and Computer Tests

Test Item	Paper-and-Pencil	Computer
1	-0.348	0.172
2	--Deleted--	-1.357
3	-2.442*	-1.357
4	1.053	1.055
5	-1.696	-2.089
6	-1.696	-2.089
7	--Deleted--	-2.089
8	-1.696	-2.089
9	0.752	0.172
10	-0.882	-2.089
11	-0.128*	-0.564
12	-0.882	-0.904
13	0.594	0.172
14	-0.882*	-0.904
15	-0.882	0.727*
16	-1.231	-0.904
17	0.752	0.553
18	-0.348	-0.564
19	1.053	1.211
20	0.752	0.553
21	1.198	1.055
22	1.198	1.516
23	1.198	1.516
24	-0.128	-0.285
25	-0.882	0.369
26	0.430	0.553
27	1.341	1.816
28	0.752	1.365
29	-0.882*	1.357
30	0.256	0.172
31	0.071	0.727
32	-0.882	-0.564
33	-0.128	-0.904
34	0.904	0.894
35	0.256	1.055
36	1.625	1.365
37	1.911	1.966*
38	-0.595	0.172
39	0.256	1.516
40	0.256	-0.564

*Significant Pearson chi-square lack of fit