Test Mode Familiarity and Performance - Gender and Race Comparisons of Test Scores among Computer-Literate Students in Advanced Information Systems Courses

Patricia Wallace
School of Business
The College of New Jersey
Ewing, New Jersey 08628, USA

Roy B. Clariana
College of Education
Penn State University
Malvern, PA 19355, USA

ABSTRACT

This investigation compares the effects of test mode, gender and race on paper-based versus computer-based delivery of two high stakes multiple-choice course examinations, Midterm and Final. Computer-literate students in upper-level business courses (n = 144) were randomly assigned to receive both tests either on paper or on computer. There were no significant gender effects, though males scored slightly higher than females on both tests. However, participants who received the tests on paper significantly outscored those who received the tests on computers, but this difference occurred ***only*** on the Midterm examination. Most striking, non-white females receiving the computer-based test mode scored lowest on the Midterm examination but then scored highest on the Final; all other groups maintained their relative positions from Midterm to Final. It was concluded that test mode familiarity does impact test performance. The results suggest that even computer-literate students in advanced Information Systems classes should practice using mock computer exams before taking high stakes computer-based tests, and that test mode familiarity affected non-white females most.

Keywords: Test Mode, Information Systems, Test Familiarity, Computer-Based Testing, Gender, Race.

1. INTRODUCTION

Recently, there has been a rapid expansion of computer-based testing (mainly multiple-choice format) in both higher education and the professions. For example, college credit and admissions tests include the College-Level Examination Program (CLEP), Graduate Record Examinations General Test (GRE), Test of English as a Foreign Language (TOEFL), and the Graduate Management Admission Test (GMAT). Computer-based professional licensure tests include the National Council of State Boards of Nursing (NCLEX), the United States Medical Licensing Examination (USMLE), and Praxis for new teachers from the Educational Testing Service (Bennett, 2002; Wainer and Eignor, 2000).

Logistics and access concerns as well as tradition suggest that these kinds of tests will be offered in both computer and paper form for many years to come. Which begs the question, are computer-based multiple-choice tests comparable to their paper-based "relatives"? Further, are computer-based tests "fair" for the increasingly diverse higher-education student body? And, most importantly, does test mode familiarity affect test performance? Since many courses are in transition from traditional paper-based testing to computer-based testing, the importance of identifying and understanding how test mode factors impact performance, along with gender and race differences, is essential to curriculum design and instruction.

2. BACKGROUND

2.1 Overview

Instructional design canon insists that paper-based versus computer-mediated instructional components should produce exactly equivalent results if the content and cognitive activities are the same (Clark, 1994). In most test mode effect studies, the computer-based and paper-based versions

are identical and so the cognitive activity should be the same, yet significant differences are regularly observed.
For example, using identical paper-based and computer-based multiple-choice tests, Lee and Weerakoon (2002) reported that, overall, students performed significantly better on paper than on the computer. Paper-based test scores were also greater than computer-based test scores for both mathematics and English CLEP tests (Mazzeo, Druesne, Raffeld, Checketts, and Muhlstein, 1991) and for recognizing fighter plane silhouettes (Federico, 1989). Other studies have reported no difference between computer and paper-based tests (Schaeffer, Reese, Steffen, McKinley, and Mills, 1993).

In a review of computer-based testing studies, Bunderson, Inouye, and Olsen (1989) reported nine studies that showed higher scores for paper-based tests, three studies that showed higher scores for computer-based tests, and eleven studies that showed no difference. Based on their findings, the chances of any particular test being equivalent on paper and computer are only 11 of 23, or approximately 50 percent.

How much different are computer-based versus paper-based test scores? Bunderson et al. (1989) state that even though, "...the scores on tests administered on paper were more often higher than on computer-administered tests... the score differences were generally quite small..." (p.378). Mead and Drasgow (1993) in a meta-analysis of well-designed computer versus paper-based cognitive ability tests also found that on average, paper-based test scores were slightly greater than computer-based test scores. Note that most of these comparison studies involved extensively developed and refined tests and so these results may not generalize to instructor-developed classroom tests.

2.2 Importance for Business Educators

In any given business school program, 70 percent of the students on average are white males (Gilbert, 2003) who by nature of the business course work, must become computer literate. For this reason, Cukier, Shortt, and Devine (2000) pointed out that identifying and addressing gender, race, and age differences in any component of instruction (and especially testing) is crucial.

In a study of the Graduate Record Examination delivered by computer and paper to a population of highly computer literate examinees who self-selected to take the computer-based version, Parshall and Kromrey (1993) reported that computer-based test scores on the verbal, quantitative, and analytic sections were all greater than complementary paper-based test scores. Here, gender, race, and age were associated with test mode. In general, white males did better with computer-based delivery while males in other racial groups did best with the paper-based tests, though there was no difference between females for paper or computer-based tests.

In a study involving freshman business undergraduates (n = 105) in an introductory business course, Computer Fundamentals, Clariana and Wallace (2002) examined the effects of a computer-based versus identical paper-based unit test on fundamental knowledge given early in the course sequence. Results showed that the computer-based test group significantly outperformed the paper-based test group. Gender, competitiveness, and computer familiarity were NOT related to this performance difference.

On the other hand, Wallace and Clariana (2004) have reported gender differences. Undergraduate-level freshman business majors (n = 207) enrolled in an introductory business_computer skills course were randomly assigned to either the computer or paper-based test mode for the duration of the course. They examined student performance on two separate tests, the first test near the start of the course and the second at the end of the course. Results showed that students scored significantly higher on computer versus paper administration, similar to the results of Clariana and Wallace (2002). Further, *post hoc* analysis indicated that the female group, whether tested on paper or online, scored below the males on the first test; however on the final exam, the female students in the computer-administered test group on average attained the highest scores of any group. They suggested that females gained computer savvy during this introductory-level course that mitigated the initial performance deficits observed on the first test.

The purpose of this present investigation is to appraise the comparability of identical computer- and paper-based tests of advanced-level Information Systems course content for males and females of different races. Specifically, this course, like many courses, is in transition from traditional paper-based testing to computer-based testing, and the findings from this investigation will determine if test mode and/or test mode familiarity is a factor in grade performance among highly literate computer users. Bugbee (1996) recommends that test developers must show that computer-based and paper-based versions are equivalent, or must provide scaling information to allow the two to be equated. Similarly, Clariana and Wallace (2002) state, "...it is critical to realize that computer-based and paper-based tests, even with identical items, will not necessarily produce equivalent measures of student learning outcomes. Instructors and institutions should spend the time, cost, and effort to mitigate test mode effects." Besides providing direct evidence of the adequacy of this computer-based test approach for this specific course, the findings may also contribute to the growing base of studies on the use of computer-based tests in business education.

3. METHODOLOGY

3.1 Participants

Four sections of an upper-level Information Systems course, Information Resource Management, consisting of 144 students were selected as the sample for this investigation. Two sections consisting of 72 students were randomly selected as the paper-based (traditional) test group. Two additional sections consisting of 72 students were identified as the computer-based test group. In the sample, 30 females and 42 males participated in the paper test treatment whereas 25 females and 47 males participated in the computer-administered test treatment. The smaller number of female

students (38.2%) participating in the study is directly related to the disproportionate number of female students versus male students that are enrolled in the major, Information Systems Management. To insure consistency in methodology, all four sections were scheduled as morning classes that met one day a week for a double period of classroom lecture and computer lab time. In addition, all four sections were taught by the same professor, in the same classroom and computer lab, and using the same instructional delivery to insure that differences due to instructor, class time, classroom facilities, and or computer labs were not mitigating factors.

3.2 Course Content

Information Resource Management is an upper-level required course in the Information Systems Management major in the AACSB accredited Business School used in this study. This course provides a comprehensive overview of the field of computer user support. Students are introduced to the spectrum of services provided to computer users and participate in an internship at the college's Information Technology Department. In addition, students are exposed to the interpersonal, communications, and problem-solving skills required in information systems positions. Thus, junior and senior-level business students with extensive computer experience are enrolled in Information Resource Management. Due to the nature of the major, students have a high-level of computer literacy and are comfortable working with computers and various types of software.

The syllabus distributed to the students on the first day of class identified two required tests, a Midterm and a Final exam. Each test was worth 25 percent of the overall semester course grade; thus, the testing portion of the course contributed a total of 50 percent of the student's semester grade computation.

3.3 ExamView Testing Software

The ExamView test generator software used in this study was provided by the textbook publisher, Course Technology (a division of Thomson Learning). The Exam View software allows instructors to create and print both paper tests and online tests. The option one chooses depends on your particular testing needs. Using the Exam View software, the Midterm and Final examinations were created by the instructor by selecting questions from the test bank. The Midterm exam included the first six chapters of course material while the Final exam, likewise, covered six chapters. Both tests consisted of 100 multiple-choice questions each with four answer alternatives. It should be noted that the Final test was not a comprehensive examination but was identical in length and complexity to the Midterm examination.

3.4 Test Procedure

On the paper test, six or seven questions were printed on each page. Students read each question and then filled in the circle of the letter selected (A. B, C, or D) of the answer choice on an Opscan™ answer sheet. Students could review and change previously answered questions before ending the test by cleanly erasing and then rewriting their choice on the answer sheet. With the computer-administered version, students received one question per screen. Students clicked on the letter of the correct answer choice and then proceeded to the next question. Students could review and change previously answered questions before ending the computer-based testing. Thus, the Midterm and Final examinations were identical for both groups of students; the only difference was the mode of administration. In addition, the same instructor taught all students insuring that class content and coverage were consistent among all four sections used in this study.

4. RESULTS

This investigation used a posttest only design; means and standard deviations are shown in Table 1. The Midterm and Final posttest data were analyzed by 2 x 2 x 2 x 2 repeated measures of Analysis of Variance (ANOVA). The ANOVA included Test Format (paper or computer), Gender (male or female), and Race (white or non-white) as the treatment main effects, along with the repeated measure, Test Sequence (Midterm and Final).

A significant main effect was observed for Test Format, $F(1, 136) = 10.930$, $MSe = 81.959$, and $p < 0.001$ as shown in Table 2. The combined mean for the paper-based test group was significantly greater ($\bar{X} = 82.7$) than that of the computer-based test group ($\bar{X} = 78.5$). The remaining between subjects' effects was not significant.

For the within subjects' effects, significant interactions were observed between Test Format and Test Sequence, between Test Format, Test Sequence, and Gender, between Test Sequence, Gender, and Race. All three of these interactions are subsumed within the significant complex four-way interaction of Test Sequence, Test Format, Gender, and Race, $F(1, 136) = 8.040$, $MSe = 22.900$, and $p < 0.005$.

This complex four-way interaction shown in Figure 1 may be visualized as two snap shots in time, and so as follow-up analyses, two separate univariate 2 (Test Format) x 2 (Gender) x 2 (Race) Analysis of Variance (ANOVA) were conducted, one for Midterm and one for Final examination data. For the Midterm examination data, Test Format was significant, $F(1, 143) = 35.489$, $MSe = 49.852$, $p < .001$. The paper test mode Midterm mean ($\bar{X} = 85.1$) was significantly

Table 1: Means and Standard Deviations

	Paper			Computer		
	N	Mid	Final	N	Mid	Final
Female						
Non-white	10	80.7	75.8	6	73.0	83.3
		(8.2)	(7.0)		(8.4)	(10.2)
White	20	86.5	80.3	19	77.1	79.0
		(4.9)	(6.3)		(7.2)	(7.3)
Male						
Non-white	14	84.1	81.1	7	76.6	74.6
		(7.2)	(5.5)		(9.0)	(14.7)
White	28	86.1	81.5	40	77.0	81.6
		(7.4)	(6.2)		(6.8)	(7.4)

Table 2: Analysis of Variance

Source	SS	df	MS	F	Sig.
Test Format (TF)	895.81	1	895.81	10.93	0.001
Gender (G)	36.53	1	36.53	0.44	0.505
Race (R)	306.01	1	306.01	3.73	0.055
TF * G	118.41	1	118.41	1.44	0.231
TF * R	23.47	1	23.47	0.28	0.593
G * R	0.02	1	0.02	0.00	0.986
TF * G * R	185.40	1	185.40	2.26	0.135
Error	11146.45	136	81.95		
Within-Ss					
Test Sequence (TS)	11.76	1	11.76	0.51	0.475
TS * TF	873.48	1	873.48	38.14	0.000
TS * G	28.40	1	28.40	1.24	0.267
TS * R	18.65	1	18.65	0.81	0.368
TS * TF * G	135.35	1	135.35	5.91	0.016
TS * TF * R	1.16	1	1.16	0.05	0.822
TS * G * R	167.10	1	167.10	7.29	0.008
TS * TF * G * R	184.11	1	184.11	8.04	0.005
Error (TS)	3114.36	136	22.90		

greater than the computer test mode Midterm mean ($\underline{X}$ = 76.6). Also Race was significant, $\underline{F}(1, 143) = 4.772$, $\underline{MSe}$ = 49.852, $\underline{p} < .05$. The white examinees' Midterm mean ($\underline{X}$ = 81.2) was significantly greater than non-white examinees' Midterm mean ($\underline{X}$ = 79.9). No other Midterm factors or interactions were significant. For the Final examination data, the main effects (Test Format, Gender, and Race) were not significant. The two-way interaction of Test Format and Gender was significant, $\underline{F}(1, 143) = 4.608$, $\underline{MSe} = 55.007$, $\underline{p} < .05$; and is subsumed in the three-way interaction of Test Format, Gender, and Race, $\underline{F}(1, 143) = 6.718$, $\underline{MSe} = 55.007$, $\underline{p} < .05$.

A Least Significant Difference (LSD) analysis approach was used to follow-up the significance of this three-way interaction. This analysis reveals that, first, non-white females in the computer test mode treatment outscored non-white males in the computer test mode treatment. In addition, this study revealed that while there were no significant gender effects, male students (combined white and nonwhite mean) scored slightly higher than female students (combined white and nonwhite mean) on both tests. Next, white males in the computer test mode treatment outscored both non-white males in the computer test mode treatment and non-white females in the paper test mode treatment. Finally, white males in the paper test mode treatment outscored both non-white males in the computer test mode treatment and non-white females in the paper test mode treatment. In general terms, the paper-based test groups outperformed the computer-based test groups on the Midterm but not on the Final examination. Specifically, the computer-based test groups relatively improved on the Final (see Figure 1). Further, all groups maintained their relative positions from Midterm to Final examination except for one notable exception. Specifically, non-white females in the computer-based test treatment scored lowest of all groups on the Midterm and highest of all groups on the Final.

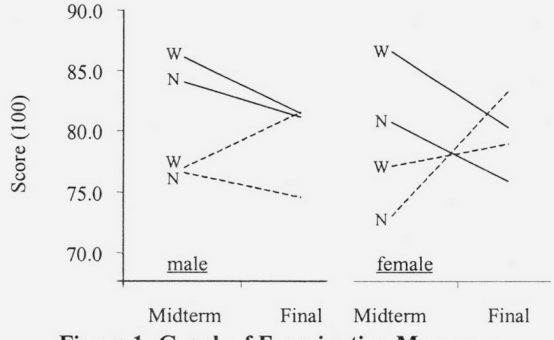

Figure 1: Graph of Examination Means-- Broken Out by Test Mode
(Key: computer – dashed; paper – solid), gender (male – M; female – F), and race (white – W; nonwhite – N).

5. CONCLUSIONS

The central focus of this investigation was to determine if test mode (paper-based versus computer-based tests), test mode familiarity, gender and/or race has an impact on test results of computer-literate students enrolled in upper-level Information Systems courses.

This investigation found that test mode did not consistently impact performance on both examinations since the paper-based test groups outperformed the computer-based test groups on the Midterm but not on the Final. These research findings concur with those of McLaren (2004) who after comparing five semesters of online and traditional sections of an undergraduate Business Statistics course found that "accomplishment of learning objectives is independent of the mode of instruction." Likewise, Wallace (2000) reported that, "no significant differences in knowledge gain were found between the control and online groups thereby suggesting that online learning presents a viable alternative instructional delivery tool and an appropriate medium for learning." Since both McLaren (2004) and Wallace (2000) used online and traditional testing modes, the finding that test mode does not consistently impact performance indicates consensus with these prior studies.

In addition, this study revealed that while there were no significant gender effects, male students scored slightly higher than female students on both tests. This may be due to lack of experience of females with computerized tests. For example, Vogel (1994) reported that "level of computer anxiety has complex effects on performance on computer administered sections of the Graduate Record Examination." Likewise, Wallace and Clariana (2004) concluded that class

experiences—such as practice exams—could mitigate such performance differences between males and females.

In the present investigation, test mode familiarity does impact test performance. Participants who received the tests on paper significantly outscored those who received the tests on computers, but this difference occurred only on the Midterm examination. Most striking, non-white females receiving the computer-based test mode scored lowest on the Midterm examination but then scored highest on the Final, while all other groups maintained their relative positions from Midterm to Final. Thus, this improvement in performance from the Midterm to the Final for this group may be attributed to test mode familiarity. Parshall and Kromrey (1993) concur that such experience improves test performance.

6. RECOMMENDATIONS

Based on the data from this investigation, it was concluded that even computer-literate students in advanced business classes benefit from *test mode familiarity*. Thus, it is recommended that *all* students should practice using mock computer exams before taking high stakes computer-based tests. This recommendation is particularly important in light of the findings from this study since the natural assumption is that students enrolled in advanced Information Systems courses are highly computer literate and, therefore, may not need to be exposed to such practice examinations. Thus, it is recommended that this misconception be corrected, and that the findings from this study regarding the benefits of *test mode familiarity* be communicated to instructors of Information Systems courses.

Furthermore, since test mode familiarity affects non-white females most, it is highly recommended that these students become familiar with the testing software to eliminate any gender and/or race differences that may occur in computerized testing. Most importantly, it is recommended that all students be exposed to practice tests to insure familiarity with the software before taking tests that have a profound impact on course grades. Thus, computer-literacy alone does not insure that students will perform well on computerized tests; test mode familiarity enables all students, regardless of gender or race, to eliminate any test mode obstacles.

7. REFERENCES

Bennett, Randy E. (2002), "Inexorable and Inevitable: The Continuing Story of Technology and Assessment." The Journal of Technology, Learning, and Assessment, Volume 1, Number 1 · June 2002. Retrieved November 18, 2003, from http://www.bc.edu/research/intasc/jtla/journal/pdf/v1n1_jtla.pdf

Bugbee, Alan C. Jr. (1996), "The Equivalence of Paper-and-Pencil and Computer-Based Testing." Journal of Research on Computing in Education, 28 (3), pp. 282-299.

Bunderson, C. Victor, Daniel K. Inouye, and Olsen, J.B. (1989), "The Four Generations of Computerized Educational Measurement." In R. L. Linn (ed.), Educational Measurement, pp. 367-407, Washington, DC: American Council on Education.

Clariana, Roy B. and Patricia E. Wallace (2002), Paper-based versus computer-based assessment: Key factors associated with the test mode effect. British Journal of Educational Technology, 33 (5), 595-604.

Clark, Richard E. (1994), "Media Will Never Influence Learning." Educational Technology, Research and Development, 42 (2), 21-29.

Cukier, Wendy, Denise Shortt, and Irene Devine (2000), "Gender and Information Technology: Implications of Definitions," Journal of Information Systems Education, Vol. 13, No. 1, 2000, pp. 7-15.

Federico, Pat-Anthony (1989), "Computer-Based and Paper-Based Measurement of Recognition Performance." Navy Personnel Research and Development Center Report NPRDC-TR-89-7. (Available from ERIC: ED 306 308)

Gilbert, Nedda (2003), "Female Appeal: Business Schools Get Up to Speed". Princeton B-School Review. Retrieved November 18, 2003, from http://www.princetonreview.com/mba/research/articles/women/appeal.asp

Lee, Gary and Weerakoon, Patricia (2002), "Student performance in Computer-Based Versus Paper-and Pen Multiple-Choice Tests." Presented at the Annual Conference of Higher Education Research and Development Society of Australasia, Perth, Australia, July 3-6, 2002. Retrieved November 18, 2003, from http://www.ecu.edu.au/conferences/herdsa/main/

Mazzeo, John, Barry Druesne, Paul C. Raffeld, Keith T. Checketts, and Alan Muhlstein (1991), "Comparability of Computer and Paper-and-Pencil Scores for Two CLEP General Examinations." College Board Report No. 91-5. (Available from ERIC: ED 344 902)

McLaren, Constance H. (2004), "A Comparison of Student Persistence and Performance in Online and Classroom Business Statistics Experiences." Decision Sciences 2 (1), pp. 1-10.

Mead, Alan D. and Fritz Drasgow (1993), Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. Psychological Bulletin, 114, 449-458.

Parshall, Cynthia G. and Jeffrey Kromrey (1993), "Computer Testing versus Paper-and-Pencil: An Analysis of Examinee Characteristics Associated with Mode Effect." A paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA, April. (Available from ERIC: ED 363 272)

Schaeffer, Gary A., Clyde M. Reese, Manfred Steffen, Robert L. McKinley, and Craig N. Mills (1993), "Field Test of a Computer-Based GRE General Test." Educational Testing Service RR-93-07. (Available from ERIC: ED 385 588)

Vogel, Lora Ann (1994), "Explaining Performance on Paper-and-pencil Versus Computer Mode of Administration for the Verbal Section of the Graduate Record Exam," Journal of Educational Computing Research, Vol. 11, No. 4, 1994, pp. 121-128.

Wainer, Howard and Daniel Eignor (2000), "Caveats, Pitfalls, and Unexpected Consequences of Implementing Large-Scale Computerized Testing." In H. Wainer (Ed.), Computerized adaptive testing: A primer (2nd ed.). Mahwah, NJ: Erlbaum.

Wallace, Patricia E. (1999), "A Comparison of the Effectiveness of Traditional Instruction Versus Online Instruction when Learning to use a Spreadsheet," Penn State University, August, 1999.

Wallace, Patricia E. and Roy B. Clariana (2004), "Gender Differences in Computer-Administered Versus Paper-Based Tests." International Journal of Instructional Media, *32* (2), in press.

AUTHOR BIOGRAPHIES

Patricia Wallace teaches courses in Information Systems and Technology as an Associate Professor in the School of Business at The College of New Jersey (TCNJ). She earned her doctorate from Temple University and completed post-doctoral study in Management Science and Information Systems at Penn State University.

Roy Clariana is an Associate Professor in the Instructional Systems program in the College of Education at the Pennsylvania State University. He earned his doctorate from Memphis State University.

Information Systems & Computing
Academic Professionals

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the Journal of Information Systems Education have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 1055-3096