

Abstract

Purpose: To compare paper-based test (PBT) and computer-based test (CBT) scores and whether demographics, previous academic success, learning style or attitudes affected performance. **Methods:** 65 second-year optometry students completed attitude questionnaires after four midterm examinations — one PBT and one CBT in two courses taught by the same instructor in the same semester. Attitudes were analyzed via X^2 comparisons while test scores were compared via analysis of variance. **Results:** PBT and CBT scores were the same, but female, non-Hispanic and non-White students and those with a preference for PBT had higher PBT scores. **Conclusion:** While overall PBT and CBT formats appeared equivalent, we uncovered demographic trends in both test performance and attitudes that warrant further investigation.

Key Words: computer-based test, paper-based test, student performance

Background

Standardized assessments used for professional licensure and to measure pre-professional aptitude have been administered electronically for decades. Computer-based test (CBT) platforms have also been introduced in many undergraduate and professional programs, and their use has coincided with the emergence of cognitive learning theories that stress integration of classroom teaching and assessment.¹

Most health profession programs have multiple layers of learning outcomes. Specifically, optometric educators must consider individual course objectives, school or overall curriculum outcomes, and professional outcomes and attributes defined by the Association of Schools and Colleges of Optometry (ASCO),² the Accreditation Council on Optometric Education (ACOE)³ and the National Board of Examiners in Optometry (NBEO).⁴ While the 20th century paradigm of providing traditional lectures followed by static, multiple-choice paper-based tests (PBT) still exists and has value,⁵ it makes it far more challenging and time-consuming to measure student performance on these learning outcomes across a curriculum. After all, PBT formats do not automatically generate reports of student performance according to learning outcomes in a course or across the curriculum. That process requires a significant time investment and coordination between instructors after a test is given. In contrast, CBT platforms allow evaluators to digitally assign learning levels and course or program levels to test questions, providing improved feedback when attempting to drive assessment-based curriculum or instructional changes.⁶ Further, controlled studies and informal observations have shown that CBT formats provide additional advantages as well as some disadvantages for students and instructors. Additional advantages include automatic grading with nearly real-time score reporting to students⁷ and higher-resolution feedback about specific subject content.⁸ Disadvantages of CBT formats include computer eyestrain, compatibility and connectivity issues⁸ as well as difficulties creating, navigating or grading in-depth questions.⁹ Arguably, the most important barrier is the general skepticism among both instructors and students toward computerized testing until they achieve an initial positive experience.¹⁰

Previous studies of student academic performance using CBT are equivocal. At least two groups of researchers have concluded that simply administering an exam using a computerized format vs. a traditional paper- and pencil-based format had no significant effect on achievement^{11,12} while another group has observed that test scores improved with CBT formats for both static-type (multiple-choice, true/false) and interactive problem-based questions.¹³ Our institution recently implemented one such CBT platform, and computerized testing was highly encouraged (but optional) by our administration during a “hybrid” period of one academic year. At the end of that year, instructors were required to administer only computerized tests. That one-year period — when both PBT and CBT formats were used — provided us with a unique opportunity to better understand how student demographics, academic performance and attitudes influenced their transition to computerized testing. Our aim was to determine whether there was a significant difference between mean PBT and CBT scores and the extent to which various student factors and attitudes could influence performance. The results from this study could reveal different strategies for improving student performance and easing the overall transition to computer-based testing.

Methods

Subjects

We recruited 65 (39 females, 26 males) subjects with a mean age (\pm S.D.) of 24.7 (\pm 1.87) years. Due to the high proportion of Hispanics in our student population, we wanted to analyze the results not only considering age, gender and race but also Hispanic ethnicity (10 Hispanic, 55 non-Hispanic). Volunteers were eligible if they were enrolled and in good academic

Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |
2

standing (i.e., cumulative GPA of 2.0 or higher) in the Rosenberg School of Optometry's second-year class during the fall semester of 2015. All subjects were recruited via a classroom announcement by the investigators and were all paid for their participation. All subjects provided informed consent. The study protocol was approved by the institutional review board at the University of the Incarnate Word and carried out according to the guidelines set out in the 1964 Declaration of Helsinki.

Research design

During the fall semester of their second year, 65 students completed two midterm examinations in two different courses (ocular physiology and organ pathology). These courses were selected because they were offered simultaneously in the same semester by the same instructor, and all exams were delivered in the same classroom setting. This could minimize the impact of different instructors using different question styles to assess student performance. These examinations were administered as regularly scheduled assessments in each course. In the first course (organ pathology), the students took the first midterm electronically via ExamSoft (ExamSoft Worldwide, Inc., Boca Raton, Fla.) and the second midterm using a traditional paper and pencil format (graded using Scantron forms and software; Scantron Corporation, Eagan, Minn.). The order was reversed for the second course (ocular physiology). The overall crossover design is shown in **Figure 1**. Despite the possible influence of test performance on student attitudes, students were unaware of this research opportunity until after the end of the second midterms. This was done for two reasons. We were primarily interested in student performance and attitudes after their initial exposure to CBT. Students in this cohort were not previously exposed to CBT (ExamSoft) in optometry school but rather used PBT (Scantron) during their first year. In addition, we did not want to create anxiety nor prompt concerns in students about the new CBT format.

While all examinations contained only multiple-choice items (both single and multiple response), the computerized format allowed for randomized ordering of questions and answer choices (i.e., multiple exam versions) while Scantron-graded exams were administered using a single version. Highlighting and backward navigation were the only additional features enabled on ExamSoft. In addition, students had access to scratch paper regardless of the testing platform. At the conclusion of the semester, study participants' attitudes toward paper and computerized exams were measured via a custom questionnaire with 5-rating scale semantic differential items (use and analysis described in a review of best practices in assessing student attitudes¹⁴).

With our semantic differential items, respondents did not indicate a level of agreement with a statement (as in Likert scaled items); rather, they chose a position on a scaled line that connoted a preference for one method or the other. For example, survey question #1 asked subjects to indicate whether they were "...more stressed taking the test using the computer-based test" or "...more stressed taking the test on paper." Whether the direction (+ or -) corresponded to CBT or PBT preference was randomly assigned for each question. However, for analysis, "-2" (on the 5-rating scale) indicated a strong preference for the PBT (more stressed by the CBT) and a "+2" indicated a strong preference for CBT (more stressed by the PBT). Responses of "-1" or "+1" indicated less endorsement, and a "0" indicated no preference or endorsement at all. Students also completed a 16-item VARK questionnaire (VARK; available online at <http://vark-learn.com/the-vark-questionnaire/>) to determine the level to which they are visual (V), aural (A), read/write (R) or kinesthetic (K) learners, as different learning styles have been associated with student performance.¹⁵

Analysis

Our first goal was to determine the relationship between the CBT and PBT scores. We examined scores across the four examinations via repeated measures analysis of variance (ANOVA) with course and examination mode as fixed factors. We then pooled all the exam scores into a single variable representing each subject's difference between their average CBT

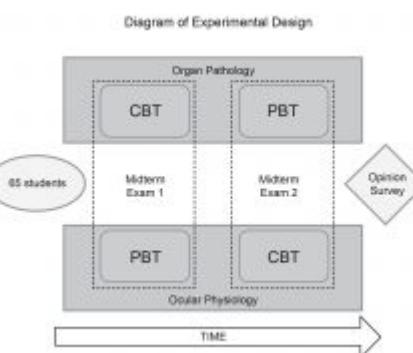


Figure 1. Subjects took two midterms in two separate courses (organ pathology and ocular physiology). In organ pathology, the first midterm was administered electronically and the second midterm using a traditional paper and pencil format. The order was reversed for the second course. After both midterms, attitudes toward the examination formats were measured via a custom questionnaire.

[Click to enlarge](#)

Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |

and PBT scores. These differences were also analyzed by ANOVA with gender, race (White vs. non-White) and ethnicity (Hispanic vs. non-Hispanic) as fixed factors. The difference scores were also regressed on continuous variables such as age and academic parameters (Optometry Admission Test [OAT] scores, undergraduate grade point averages [GPAs], GPA thus far in the optometry program, and learning style and attitudes). Distributions of subjects by demographics are shown in **Table 1**.

We determined the level of endorsement (either for the CBT or PBT format) by comparing the medians for each survey question to zero via a Wilcoxon signed rank test. Ad hoc comparisons of students performing best on PBT or CBT were also performed, and these are described in detail in the Results section. Statistical significance was defined as a p-value less than 0.05 for all testing. All analyses were performed using SPSS (IBM Corporation, Armonk, N.Y.) and Excel (Microsoft Inc., Redmond, Wash.).

Results

All 65 second-year students consented to participation. The range of the difference between the computer-based and paper-based scores (C-P, for convenience) was -18.6 to 13.9 (refer to **Table 2** for summary of scores). One male subject scored 34.5 points (4.5 standard deviations from the mean difference) better on the PBT. His performance and survey data were excluded from all analyses, as his performance bias for paper-based testing was considered an outlier. All test scores for the remaining subjects ($n = 64$) were analyzed using repeated measures ANOVA. The main effect of course [$F(1,57) = 11.43$, $p = 0.001$] — with students scoring 3.4% higher in ocular physiology — and the interaction of course and exam mode [$F(1,57) = 6.607$, $p = 0.013$] were both significant, but there was no main effect of exam mode [$F(1,57) = 1.740$, $p = 0.192$]. So, while there was a bias (C-P = -2.30) in our subjects toward better performance on the PBT, the difference was insignificant.

TABLE 2
Exam Scores (PBT, CBT and difference) by Gender, Ethnicity and Race*

	All subjects			Hispanic subjects			Non-Hispanic subjects		
	All	White	Non-White	All	White	Non-White	All	White	Non-White
PBT	83.9 (7.4)	85.1 (7.2)	82.2 (7.5)	79.5 (7.9)	82.2 (5.8)	73.2 (9.7)	84.8 (7.1)	85.0 (7.4)	82.4 (6.8)
Male	81.4 (7.8)	83.5 (6.0)	78.1 (8.7)	76.1 (8.8)	84.5 (1.8)	87.8 (0.1)	82.4 (7.2)	83.4 (7.1)	85.8 (7.8)
Female	85.6 (8.7)	86.2 (7.9)	84.2 (8.6)	81.8 (8.3)	83.3 (8.8)	— ^a	86.2 (8.7)	87.5 (7.3)	84.8 (9.8)
CBT	81.6 (7.8)	84.1 (6.1)	78.0 (8.6)	77.7 (18.8)	83.0 (8.0)	69.1 (2.9)	82.4 (7.1)	84.3 (6.2)	79.7 (7.8)
Male	79.7 (8.2)	83.9 (7.4)	73.4 (8.2)	73.0 (13.4)	84.5 (9.7)	83.4 (0.4)	80.8 (8.2)	83.9 (7.5)	79.8 (7.8)
Female	82.9 (6.5)	84.2 (5.3)	80.0 (7.8)	80.1 (7.5)	82.5 (5.3)	— ^a	84.4 (8.4)	84.7 (5.2)	81.7 (7.3)
C-P	-2.30 (8.8)	-1.02 (6.4)	-4.17 (8.8)	-1.87 (7.1)	0.82 (5.6)	-0.17 (6.8)	-2.38 (8.8)	-1.43 (6.6)	-3.65 (6.5)
Male	-1.62 (7.3)	0.44 (7.1)	-4.78 (8.7)	-2.53 (7.1)	-0.82 (12)	-4.23 (0.0)	-1.52 (7.5)	0.91 (7.8)	-4.82 (7.8)
Female	-2.73 (8.2)	-1.96 (5.8)	-3.84 (8.7)	-1.70 (7.7)	1.16 (3.8)	— ^a	-2.92 (8.8)	-2.83 (6.2)	-2.20 (8.1)

^aMean values (SD) reported. *Data not shown for single subject.

PBT = paper-based test; CBT = computer-based test; C-P = CBT - PBT scores

Table 2. Click to enlarge
TABLE 1
Distribution and Age of Subjects by Gender, Ethnicity and Race (White vs. non-White)*

Overall (N=64)	All subjects			Hispanic subjects			Non-Hispanic subjects		
	All	White	Non-White	All	White	Non-White	All	White	Non-White
N (%)	64 (100%)	38 (59%)	26 (41%)	18 (19%)	7 (11%)	11 (17%)	54 (84%)	31 (49%)	23 (36%)
Male	25 (39%)	15 (23%)	10 (16%)	4 (8.2%)	2 (3.1%)	2 (3.1%)	21 (33%)	13 (20%)	8 (12%)
Female	39 (61%)	23 (39%)	16 (25%)	8 (9.4%)	5 (7.8%)	11 (18%)	33 (52%)	18 (28%)	15 (23%)
Age	24.7 (1.8)	24.9 (1.9)	24.7 (1.8)	24.5 (1.6)	24.7 (2.9)	24.6 (1.8)	24.8 (2.0)	24.7 (1.8)	
Male	24.5 (1.7)	24.9 (1.5)	24.0 (2.8)	24.9 (2.2)	25.0 (1.4)	24.5 (3.5)	24.5 (1.7)	24.9 (1.6)	23.9 (1.8)
Female	24.9 (2.0)	24.7 (2.1)	24.3 (1.2)	24.2 (1.3)	— ^a	— ^a	24.9 (2.1)	24.8 (2.3)	25.1 (1.8)

^aMean values (SD) for age are reported. *Data not shown for single subject.

Table 1. Click to enlarge

For the two formats to be considered equivalent, it has been suggested that the two testing methods must yield similar dispersions and overall distributions of scores.¹⁶ The scores from all examinations were distributed normally ($p > 0.05$ on Anderson Darling testing), though the distributions were shaped somewhat differently (**Figure 2**). The distribution most distinct from the others was for the ocular physiology PBT, where no student received a failing grade (<65%). However, 4.7% (3 of 64) of students failed the CBT in ocular physiology. This trend was reversed for organ pathology, where only 6.3% (4 of 64) failed the CBT, but 10.9% (7 of 64) failed the PBT. There does not appear then to be a systematic difference in test scores based on test platform. We then combined the two course test scores for each platform, and the relationship between the average scores from each platform is shown in **Figure 3**. Computer-based and paper-based scores were highly positively correlated ($r = 0.62$, $p < 0.001$), providing additional evidence that the two platforms are essentially equivalent in assessing student performance.

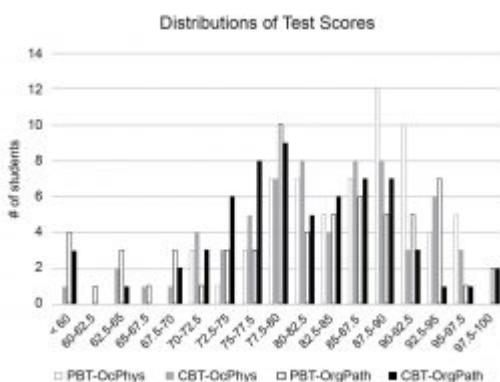
Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |
4


Figure 2. Distributions of paper-based test (PBT) and computer-based test (CBT) scores. The shapes of the distributions were somewhat different among all four midterm examinations, but all were distributed normally. [Click to enlarge](#)

It has been argued, however, that any two methods that test the same parameter and provide wide variability in scores will almost always be positively correlated.¹⁷ It is better for us to calculate the difference (C-P) and plot this difference against the overall average of both methods.¹⁸ This difference (or Bland-Altman) plot is shown in **Figure 4**. The mean difference (-2.30) is represented by the solid line. When the difference between the two methods was regressed on the average of all exams, we saw no relationship ($r=0.06$, $p>0.9$). That is, mean overall test performance did not systematically predict a performance bias for either testing platform. The differences were distributed normally (Kolmogorov-Smirnov statistic=0.10, $p>0.05$) with a median of -3.50 and 25th and 75th percentiles of -6.63 and 2.25, respectively. This slightly positively skewed distribution was seen when we computed the reference interval (or 95% limit of agreement) for the two testing methods (mean ± 1.96 standard deviation of the difference). These values (10.3, -15.3) are shown in Figure 4 as thick dashed lines and indicated significant agreement between the two methods for all but six subjects (four performing significantly better on the CBT and two significantly better on the PBT).

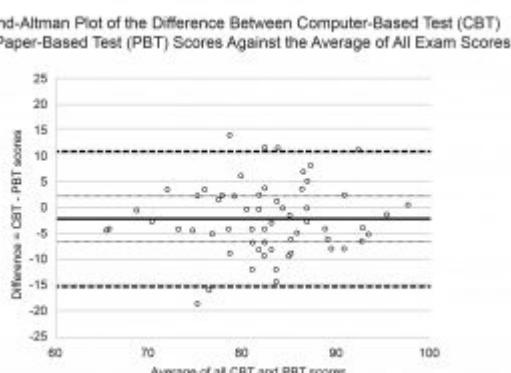


Figure 4. Positive differences represent better mean performance on the CBT. The solid black line represents the mean difference between CBT and PBT scores (-2.30). The dotted lines represent the 25th percentile (-6.63) and 75th percentile (2.25). The heavy dashed lines represent the 95% confidence intervals below (mean -1.96 S.E.; -15.27) and above the mean (mean +1.96 S.E.; 10.57). [Click to enlarge](#)

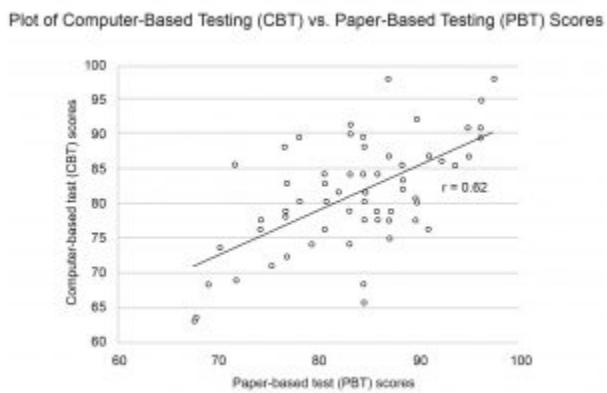


Figure 3. Computer-based test scores and paper-based test scores were highly positively correlated.
[Click to enlarge](#)

While the testing administration methods appeared to be equivalent, we were still interested in what factors may be responsible for the trend in our subjects to perform better on paper. We pooled all the exam scores into a single variable (C-P), which was analyzed by ANOVA with gender, race (White vs. non-White) and ethnicity (Hispanic vs. non-Hispanic) as fixed factors. We found a trend for the mean bias towards PBT to be more in non-White subjects than in White subjects [$F(1,57)=3.093$, $p=0.053$]. When analyzed separately by one-sample t-tests, the bias toward PBT in non-White subjects was significantly less than zero [$t(57)=-3.22$, $p<0.005$] but not significant [$t(57)=-0.98$, $p=0.33$] in White subjects. The effects of gender [$F(1,57)=0.609$, $p=0.439$] and ethnicity [$F(1,57)=0.272$, $p=0.604$] were not significant. However, while neither gender nor ethnicity played a significant role on ANOVA, there was a difference when analyzed separately. That is, the mean difference for female subjects (-2.73) was significantly less than zero on a one-sample t-test [$t(57)=-2.75$, $p=0.01$]. This was not the case for male subjects [mean: -1.62; $t(57)=-1.11$, $p=0.28$]. Similarly, the mean for non-Hispanic subjects (-2.38) was significantly less than zero [$t(57)=-2.65$, $p=0.01$] but not for Hispanic subjects [mean: -1.87; $t(57)=-0.84$, $p=0.42$]. So — when analyzed separately — female, non-Hispanic and non-White subjects performed significantly better on paper-based tests than computer-based tests.

We then regressed the difference (C-P) onto the following predictors: age, OAT scores (academic average, total science and

Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |
5

reading comprehension), undergraduate GPAs (total and math and science), GPAs in the first three semesters of our professional program, and learning styles. We found no significant correlations.

No validated questionnaire for determining attitudes toward paper and computerized examination formats is currently available. Therefore, we performed a factor analysis (via principal components) and found that the extracted communalities for all the survey question responses were significant (ranging from 0.65 to 0.96) and contributed to a single component that explained more than 37% of the variance in responses. It seems, therefore, that the survey questions represented, at least to some extent, student preference for one testing platform or the other. Therefore, we considered them all in our analysis.

The survey responses were not distributed normally, so we used a non-parametric test (Wilcoxon signed rank test) to determine whether the median responses were different than zero. A significant test and negative median indicated an overall preference (or endorsement) for the PBT format. If the median was positive, the endorsement was for the CBT. These results are summarized in **Table 3** and are shown for overall responses as well as broken down by gender and ethnicity (Hispanic or non-Hispanic). Subjects overall endorsed the PBT on 15 questions and the CBT on five questions. The results were identical for the 54 subjects who identified as non-Hispanic and similar for female subjects who endorsed paper examinations on two additional questions. Male subjects endorsed the computerized format for four questions but only endorsed paper for one question. Hispanic subjects endorsed paper testing for two questions and computer-based tests for one question.

Some commonly endorsed items for paper were questions 1 (less stressed by paper exam), 5 (more positive experience with paper) and 22 (prefer to mark tentative answers and go back before recording final answer). Endorsements for computer-based testing included questions 4 (prefer the feedback after taking the CBT), 8 (understand the need for CBT) and 18 (taking CBT prepares better for national licensure exams). To test whether these endorsements correlated in some way with a performance bias toward the paper or computerized formats, we created trichotomous variables for all survey questions, considering any negative response an endorsement for PBT, any positive response an endorsement for CBT, and any zero response to be "no opinion." Because we are really only interested in whether these endorsements predicted significantly better performance on one testing platform or the other, we only analyzed subjects that scored at least 2.25 points higher on the computer-based test (those who performed best on the CBT) or at least 6.63 points better on paper (those who performed best on PBT; refer to Figure 4). We then calculated the likelihood (via X^2 analysis) that the distribution (number of PBT endorsements, no opinion, or CBT endorsements) was the same between these two performance groups.

There were significant findings for eight survey questions representing perceived difficulty (question 2), "mood" toward platforms (questions 9 and 17), overall preference (questions 6, 10, 13 and 15), and which platform better prepares for national licensure exams (question 18). Four representative significant findings are shown in **Figure 5**. To check for a "goodness of fit" (as in X^2 analysis), we sorted the scaled predictors into ordinal variables by assigning the lower quartile as "1", second quartile as "2", third quartile as "3", and the upper quartile as "4." Only the distributions of current optometry school GPA differed significantly between students performing best on PBT or CBT ($p=0.027$; **Figure 6**).

TABLE 3
Summary of Median (mean) Survey Responses for all Questions

Question # (Text)	Overall	Male	Female	Hispanic	Non-Hispanic
1. More stressed to when platform?	-1.0 (0.88)*	0.0 (0.94)	-1.8 (0.00)*	-1.6 (0.46)	-1.0 (0.83)*
2. Which platform easier?	0.0 (0.33)*	0.0 (0.12)	-1.8 (0.02)*	0.0 (0.90)	-0.0 (0.29)*
3. More satisfied with my score using...	0.0 (0.40)*	0.0 (0.08)	-1.8 (0.08)*	0.0 (0.40)	0.0 (0.42)*
4. Preferred the feedback using...	1.0 (0.84)*	2.0 (1.28)*	1.8 (0.72)*	1.8 (0.73)	1.0 (0.88)*
5. More positive experience for my info...	-1.0 (0.44)*	0.0 (0.08)	1.8 (0.77)*	1.8 (0.44)	-0.8 (0.48)*
6. Would prefer taking all tests using...	0.0 (0.12)	1.0 (0.48)	1.8 (0.48)*	0.0 (0.48)	-0.0 (0.07)*
7. More beneficial to school's evaluations...	0.0 (0.08)	0.0 (0.18)	0.0 (0.18)	0.0 (0.18)	0.0 (0.02)
8. Do (do not) see the need for CBT	1.0 (0.84)*	1.0 (0.72)	1.8 (1.08)*	1.8 (0.72)	1.0 (0.81)*
9. Happy (angry) about using CBT	0.0 (0.88)*	1.0 (0.84)*	0.0 (0.80)*	0.0 (0.84)	0.0 (0.87)*
10. Better to administer all tests using...	0.0 (0.00)	0.0 (0.18)	0.0 (0.01)	0.0 (0.00)	0.0 (0.00)
11. I am more afraid of tests using...	0.0 (0.33)*	0.0 (0.08)	0.0 (0.09)*	0.0 (0.33)	0.0 (0.17)*
12. This is a better overall test platform	0.0 (0.34)*	0.0 (0.26)	0.0 (0.44)*	0.0 (0.26)	0.0 (0.27)*
13. I would like to take all tests using...	0.0 (0.17)	0.0 (0.04)	0.0 (0.31)	0.0 (0.43)	0.0 (0.13)
14. I prepare more for this platform	0.0 (0.00)	0.0 (0.08)	0.0 (0.10)	0.0 (0.10)	0.0 (0.02)
15. Offered choice, would take all tests using...	-0.0 (0.28)	0.0 (0.28)	-1.8 (0.46)	0.0 (0.28)	-0.0 (0.28)
16. Experience more consistent with...	0.0 (0.34)*	0.0 (0.24)	0.0 (0.41)*	0.0 (0.39)	0.0 (0.35)*
17. Glad if stayed with paper (switch to...)	0.0 (0.08)	0.0 (0.08)	0.0 (0.35)	0.0 (0.40)	0.0 (0.05)
18. Better prepares me for licensing exams	1.0 (0.86)*	1.0 (0.58)	1.8 (1.05)*	2.0 (1.48)*	1.0 (0.86)*
19. I retain information more using...	0.0 (0.38)*	0.0 (0.01)	0.0 (0.58)*	0.0 (0.48)	0.0 (0.38)*
20. I am better prepared when use...	0.0 (0.08)	0.0 (0.01)	0.0 (0.31)	0.0 (0.10)	0.0 (0.12)
21. The timing of the exam better with...	0.0 (0.02)	0.0 (0.06)	0.0 (0.26)	0.0 (0.16)	0.0 (0.06)
22. During tests, arrive at answers as I go...	-1.0 (0.80)*	-1.0 (0.82)	-1.8 (0.48)*	0.0 (0.80)	-1.0 (0.86)*
23. I am comfortable with technology...	1.0 (1.08)*	2.0 (1.28)*	1.8 (1.07)*	1.8 (0.90)	1.0 (1.02)*
24. More comfortable platform (less eyestrain, etc.)	-1.0 (0.84)*	0.0 (0.46)	-2.0 (1.13)*	-0.0 (0.78)	-1.0 (0.81)*
25. I have no apprehension with electronic testing	0.0 (0.33)	0.0 (0.28)	-1.0 (0.38)	-1.0 (0.10)	0.0 (0.22)
26. Prefer tests in lecture hall (paper) or in learning rooms (paper & pencil)	2.0 (0.18)	0.0 (0.24)	0.0 (0.46)	-2.0 (0.18)	0.0 (0.08)
27. I learned more from the test using...	0.0 (0.84)*	0.0 (0.82)	0.0 (0.64)*	0.0 (0.66)	0.0 (0.84)*
28. Multiple response questions easier to navigate using...	0.0 (0.27)	0.0 (0.01)	-1.0 (0.41)	-1.0 (0.30)	0.0 (0.22)
29. Heard about it more from paper (paper exams, exams)	-2.0 (1.16)*	-10 (0.36)	-2.0 (1.07)	-2.0 (1.28)*	2.0 (1.16)*
30. Demonstrating exams prior to test helps me prepare (unbiased me)	0.0 (0.19)	0.0 (0.18)	0.0 (0.41)	0.0 (0.20)	0.0 (0.18)
31. Better helped me test course self-esteem	0.0 (0.25)*	0.0 (0.04)	0.0 (0.58)*	0.0 (0.50)	0.0 (0.24)*
32. CBT process is just fine the way it is (needs significant improvement)	0.0 (0.34)*	0.0 (0.08)	1.0 (0.82)*	0.0 (0.00)	-1.0 (0.41)*
33. Was easier to read the exams using...	0.0 (0.88)*	0.0 (0.08)	-1.0 (0.08)	-1.0 (0.88)	0.0 (0.54)*
34. This platform was fairer	0.0 (0.18)	0.0 (0.04)	0.0 (0.29)	0.0 (0.18)	0.0 (0.20)

Greens are shaded from -2 (full endorsement of PBT to +2 (full endorsement of CBT)

* CBT vs. PBT endorsements; median significantly different from zero at the $\alpha=0.05$ level

Table 3. Click to enlarge

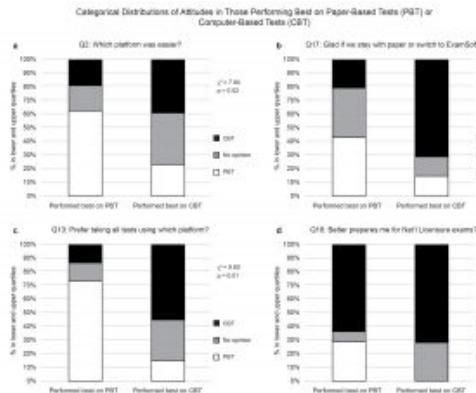
Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |


Figure 5. Each column represents the students who performed best (75th percentile) on the PBT (left; n=14) or the CBT (right; n=14). The different bands represent proportions whose attitudes favored CBT (black bands), PBT (light bands) or no preference (gray bands). The graphs shown are representative of significant findings regarding perceived difficulty (a), "mood" toward platforms (b), overall preference (c), and which platform better prepares them for national licensure exams (d). [Click to enlarge](#)

Discussion

This study found little difference in the performance of second-year optometry students on a computer-based test (CBT) compared with a paper-based-test (PBT) covering topics in organ pathology and ocular physiology. Overall, there was a trend toward higher mean scores on the PBT. In addition, an opinion survey found student attitudes favored the PBT more highly than the CBT.

Our examination performance findings are consistent with some but not all prior studies that have found little difference in student performance between CBT and PBT (see Bugbee¹⁹ and Vrabel²⁰ for reviews of the literature). Our results are consistent with those of Boevé et al.,²¹ who conducted a study with a crossover design somewhat like ours. In their undergraduate psychology course, one-half of the class (n=199) took their midterm exam on a computer while the other half of the class took the exam using paper and pencil. The groups were switched for the final examination. The researchers found no significant difference in the mean number of questions answered correctly between the computer-based and paper-based modes for both the midterm and final exam. In another recent study, Karay et al.²² evaluated the performance of 266 medical students on a standardized 200-question multiple-choice examination. Students were randomly assigned to take the exam on a computer or on paper. There was no significant difference in exam score between the groups, but students taking a PBT needed significantly more time to complete the test.

Perhaps the trend towards PBT performance in our study was influenced by real or perceived difficulty differences between the courses or the individual examinations used in our design. There was a statistically significant difference between the mean scores for ocular physiology (84.5%) and organ pathology (81.1%). This difference could be explained by an enhanced interest in the ocular physiology topics among optometry students. However, the disparity between the mean PBT and CBT scores was larger (and statistically significant on paired sample t-tests) in the ocular physiology course (86.4% for PBT vs. 81.1% for the CBT; p=0.003) than in the organ pathology course where the PBT and CBT scores (81.5% and 80.6%, respectively) were essentially equivalent (p=0.55). It is quite possible that the PBT in the ocular physiology course covered material that was easier to master or had been introduced in a previous course. Regardless of the cause, our ability to draw inferences concerning any PBT bias may be weakened by the main effect of course and its interaction with exam mode on scores.

It has been suggested that males find technology more appealing and thus are more self-confident using computers.²³ In addition, there are indications that girls achieve less well than boys on computer-based problem-solving tasks.²⁴ This has led to a concern that females may perform worse on CBT than males. We did find that females performed significantly better on PBT than CBT, while for males there was no significant difference between exam modes. Our results agree with Jeong²⁵ who studied test scores of Korean grade-school children and found the CBT scores of female students to be significantly lower than their PBT scores in three of four subjects studied. This similarity should be considered in light of the notably different cohorts. Indeed, others have found no gender differences in more applicable investigations. For example, Clariana and Wallace²⁶ compared student performance on a 100-question multiple-choice examination. Fifty-four

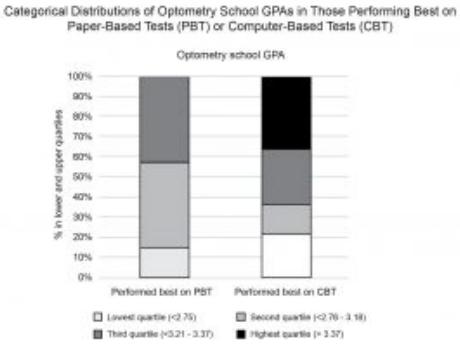


Figure 6. Each column represents the students who performed best (75th percentile) on the PBT (left; n=14) or the CBT (right; n=14). When examining all academic predictors, only distributions of current optometry school GPA differed significantly between these two groups.

[Click to enlarge](#)

Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |

7

college students took the examination on a computer while 51 students took it on paper. No attempt was made to match the two groups for gender, academic achievement or any other variable, and the investigators found that gender was not significantly associated with computer vs. paper test mode effects.

Ethnic background and race are additional factors that may influence exam mode performance. It has been reported that grade-school children from an ethnic-minority background may have less exposure to computers both in the classroom and at home than children from the majority population.²³ While our study subjects overwhelmingly endorsed computerized testing in reporting they were “comfortable with technology,” computer familiarity is one factor that has been identified as potentially influencing performance on CBT.²⁷ In our study, 10 students self-identified as being of Hispanic heritage and 54 who were non-Hispanic. Interestingly, we found that Hispanic students performed equally well on the CBT and PBT, while non-Hispanic subjects performed significantly better on the PBT. The 26 subjects who self-identified as non-White did perform significantly better on the PBT, but we found no performance bias among subjects who identified as White. These equivocal findings may indicate that disparities in computer familiarity may not be as much of a determining factor for this generation of students, at least not at the professional-school level.

No other individual student characteristic was significantly associated with exam mode performance in our analysis, in contrast with previous investigations. Watson²⁷ found that academically higher performing students benefited most from a computer-assisted learning program, while Clariana and Wallace²⁶ reported that higher attaining students performed significantly better on a CBT than a PBT, or conversely, PBT hindered the performance of high-attaining students more than low-attaining students. Our results may have been limited by a relatively low number of subjects; both previous studies involved nearly double the number subjects.

Following the second midterm examination, and after both midterm exam scores had been revealed to the students, each student completed a questionnaire. Responses from students with the greatest exam mode performance difference were analyzed for significant differences in their responses to survey questions. In general, we found that students indicated a strong preference for the exam mode that they performed best on (Figure 5c). When asked which exam mode best prepares them for national licensure exams (which are CBT in optometry), nearly equal proportions of students acknowledged that CBTs were superior. This is not a surprising result, as NBEO Parts I and II (of III) are administered electronically. However, no student performing best on CBT acknowledged that PBT would better prepare them for licensing exams, and the overall distributions were significantly different (Figure 5d). These positive associations between performance and exam mode preference are not surprising. For example, it should be intuitive that more students who performed best on the paper-based test would think it was easier (Figure 5a) and that we should stay with that testing platform (Figure 5b). Our results differ from those of Washburn et al.¹³ who used a very similar design and found that students overwhelmingly preferred paper-pencil over computer-based assessments. However, the students in their study performed significantly better on the computer-based test.

We found a trend suggesting that differences in test anxiety may have contributed to differences in performance between the exam modes. While Washburn et al.¹³ found no association between test anxiety and test performance, it has been previously suggested that students with a lower comfort level with computers may experience greater test anxiety, and subsequent lower performance, with CBT.²⁸ In a study of 131 college undergraduate volunteers randomly assigned to computerized or paper-and-pencil versions of a battery of personality tests, Lankford et al.²⁹ found that female students and those with higher computer anxiety reported more depression when the test was administered on a computer. It is suggested that computer anxiety, like test anxiety in general, is not dependent on the degree of computer experience.²⁸ In our study, distribution differences in responses to questions such as “I was more stressed taking the test using...” and “I am more afraid of tests using...” approached significance, suggesting that anxiety levels were associated with exam mode performance difference. Perhaps our small study population and their foreknowledge of their exam performance prior to completing the questionnaire influenced our ability to detect a statistically significant correlation between anxiety and exam mode performance.

Strengths of our study include the 100% student participation rate, the crossover experimental design, and our access to student academic records, including undergraduate GPA and optometry school entrance exam performance. Furthermore, each student completed a VARK learning preference survey providing insight into their learning style.

Disadvantages that limit inferences from our study include the small sample size of only 64 students and a much smaller representation of Hispanic than non-Hispanic subjects. A further disadvantage is that to gain admission to optometry school students must perform well on the OAT, which is a CBT. Students that perform poorly on CBTs would not be expected to gain admission to optometry school. Although we found that OAT scores did not predict better performance on the CBT, we are dealing with a student population that is self-selected for good performance on CBT. One other

disadvantage is that the two courses that were part of this experiment (organ pathology and ocular physiology) are not exactly equivalent. While the course material is similar (both covering foundational biologic principles) and were taught concurrently by the same instructor to the same students, there are differences in subject matter that may have influenced exam performance other than exam mode alone. For example, one course was 3 credit hours while the other was 2 credit hours. However, both courses were a continuation of first-year courses, and we believe that the difficulty of the course content was comparable throughout the semester within each course.

While some investigators have found no effect of question order on test performance,^{30,31} others have demonstrated effects on scores³² and score distributions.³³ In the current study, PBTs were administered in one version with a fixed order of questions and answer choices. There were, however, multiple versions of CBTs, which limits our ability to draw inferences about differences between the two formats. Lastly, we need to acknowledge that a better investigational approach may have been to administer each of the four midterms in both formats: PBT and CBT. Students could have been randomized into CBT and PBT for each exam and swapped for the second in each course. We decided against this approach for instructional reasons as students may have considered it unfair to take their first CBT while others in their class were being evaluated over the same material with the more familiar PBT.

Conclusions

In summary, we found no statistically significant difference in overall performance between PBT and CBT in this group of healthcare professional students. Females, non-White students, and non-Hispanic students performed significantly better on PBT. We also found that performance differences between the two formats predicted student perceptions of difficulty, preference and utility of the formats. In addition, we found trends suggesting that test anxiety may contribute to poor CBT performance among some students. These trends warrant further investigation. We plan to conduct a follow-up study on this same cohort of students to examine how their attitudes and beliefs toward CBT may have evolved with increased familiarity with this mode of exam administration.

Acknowledgement

This research was supported by an Innovation in Teaching grant from the American Academy of Optometry Foundation and Johnson & Johnson Vision.

References

1. Shepard LA. The role of assessment in a learning culture. *Educational Researcher*. 2000;29(7):4-14.
2. Association of Schools and Colleges of Optometry (ASCO). Attributes of Students Graduating from Schools and Colleges of Optometry: A 2011 Report from the Association of Schools and Colleges of Optometry. 2011 [cited Aug 26, 2019]; Available from: <https://optometriceducation.org/faculty-and-administrators/asco-policy-guidelines-and-reports/>.
3. Accreditation Council on Optometric Education (ACOE). Professional Optometric Degree Standards. 2016 [cited Aug 26, 2019]; Available from:
<https://www.aoa.org/optometrists/for-educators/accreditation-council-on-optometric-education/accreditation-resources-and-guidance/optometric-degree-programs->
4. National Board of Examiners in Optometry (NBEO). Exam Content Outlines. 2019 [cited Aug 26, 2019]; Available from: https://www.optometry.org/exam_content.cfm
5. Shavelson RJ. Assessing student learning responsibly: from history to an audacious proposal. *Change*. 2007; 39(1):26-33.
6. Pellegrino JW. The evolution of educational assessment: considering the past and imagining the future. The sixth annual William H. Angoff Memorial Lecture. Presented at: Educational Testing Service; Nov 17, 1999; Princeton, New Jersey. 2004.
7. Kuikka M, Kitola M, Laakso M. Challenges when introducing electronic exam. *Res Learn Technol*. 2014;22:22817 (e-copy). doi: 10.3402/rlt.v22.22817
8. Bussieres J-F, Metras M-E, Leclerc G. Use of Moodle, ExamSoft, and Twitter in a first-year pharmacy course. *Am J Pharm Educ*. 2012;76(5):94.

9. Cook J, Jenkins V. Getting started with e-assessment. Bath: University of Bath. 2010.
10. Craven P. History and challenges of e-assessment: The 'Cambridge Approach' perspective – e-assessment research and development 1989-2009. Cambridge: Cambridge University. 2009.
11. Millsap CM. Comparison of computer testing versus traditional paper-and-pencil testing [dissertation]. Denton (TX): University of North Texas. 2000.
12. Capay M, Magdin M, Mesarosava M. Enhancement of e-testing possibilities with the elements of interactivity reflecting the students' attitude to electronic testing. Paper presented at: Proceedings of the 10th European Conference on e-Learning; Nov 4-5; Brighton UK. 2011.
13. Washburn S, Herman J, Stewart R. Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams. *Adv Physiol Educ.* 2017;41:548-555.
14. Lovelace M, Brickman P. 2013. Best practices for measuring students' attitudes toward learning science. *CBE Life Sci Educ.* 2013 Winter;12(4):606-617.
15. Akhlaghi N, Mirkazemi H, Jafarzade M, Akhlaghi N. Does learning style preferences influence academic performance among dental students in Isfahan, Iran? *J Educ Eval Health Prof.* 2018 Mar 24;15:8. doi:10.3352/jeehp.2018.15.8
16. American Psychological Association. Guidelines for computer-based tests and interpretations. Washington, DC; 1986.
17. Bland J, Altman D. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8(2):135-160.
18. Bland J, Altman D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;327(8476):307-310.
19. Bugbee AC. The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education.* 1996;28(3):282-299.
20. Vrabel M. Computerized versus paper-and-pencil testing methods for a nursing certification examination: a review of the literature. *Comput Inform Nurs.* 2004;22(2):94-98.
21. Boevé AJ, Meijer RR, Albers CJ, Beetsma Y, Bosker RJ. Introducing computer-based testing in high-stakes exams in higher education: results of a field experiment. *PLoS ONE.* 2015;10(12):e0143616. doi: 10.1371/journal.pone.0143616.
22. Karay Y, Schauber SK, Stosch C, Schuttpelz-Brauns K. Computer versus paper—does it make any difference in test performance? *Teach Learn Med.* 2015;27(1):57-62.
23. Volman M, van Eck E, Heemskerk I, Kuiper E. New technologies, new differences. Gender and ethnic differences in pupils' use of ICT in primary and secondary education. *Computers & Education.* 2005;45(1):35-55.
24. Barbieri MS, Light PH. Interaction, gender, and performance on a computer-based problem solving task. *Learn Instr.* 1992;2(3):199-213.
25. Jeong H. A comparative study of scores on computer-based tests and paper-based tests. *Beh Inform Technol.* 2014;33(4):410-422.
26. Clariana R, Wallace P. Paper-based versus computer-based assessment: key factors associated with the test mode effect. *Brit J Educ Technol.* 2002;33(5):593-602.
27. Watson B. Key factors affecting conceptual gains from CAL materials. *Brit J Educ Technol.* 2001;32:587-593.
28. Marcoulides GA. The relationship between computer anxiety and computer achievement. *Educ Comput Res.* 1988;4(2):177-187.
29. Lankford JS, Bell RW, Elias JW. Computerized versus standard personality measures: equivalency, computer anxiety, and gender differences. *Computers in Human Behavior.* 1994;10(4):497-510.
30. Monk JJ, Stallings WM. Effects of item order on test scores. *J Ed Research.* 1970;63(10):463-465.
31. Gohmann SF, Spector LC. Test scrambling and student performance. *J Econ Ed.* 1989;20(3):235-238.

Student Performance and Attitudes During the Transition from Paper to Computer-Based Examinations |

10

32. Gruber RA. Sequencing exam questions relative to topic presentation. *J Acct Ed.* 1987;5(1):77-86.
33. Carlson JL, Ostrosky AL. Item sequence and student performance on multiple-choice exams: further evidence. *J Econ Ed.* 1992;23(3):232-235.

 Follow

 Share

Brian K. Foutch, OD, PhD, [foutch@uiwtx.edu] is an Associate Professor at the University of the Incarnate Word (UIW) Rosenberg School of Optometry where he teaches epidemiology, public health and neuroanatomy. He also precepts third- and fourth-year optometry interns in Primary Care clinics at the UIW Eye Institute.

Lourdes Fortepiani, MD, PhD, is an Associate Professor at the University of the Incarnate Word Rosenberg School of Optometry where she teaches foundational sciences including physiology, systemic disease, ocular physiology and ocular correlates of systemic disease. She is also an Adjunct Assistant Professor in the Department of Cellular and Integrative Physiology at UT Health San Antonio.

Richard Trevino, OD, is an Associate Professor at the University of the Incarnate Word (UIW) Rosenberg School of Optometry where he teaches advanced clinical procedures, glaucoma and posterior segment diagnosis and management. He also serves as Chief of the Ocular Disease Service at the UIW Eye Institute.