STUDENT PERFORMANCE ON COMPUTER-BASED TESTS VERSUS PAPER-BASED TESTS IN INTRODUCTORY FINANCIAL ACCOUNTING: UAE EVIDENCE

Samy Garas, Zayed University
Mostafa Hassan, Qatar University

ABSTRACT

This study examines whether the use of technology-based assessment tool affects the examinations' scores of students from both genders. The study, therefore, ascertains whether the mode of student testing (computer-based or paper-based) in an introductory-level financial accounting course impacted students' scores (a direct measure of learning). In doing so, the study relies on experimental design wherein the type of examination is being controlled together with other contextual variables such as timing of the exam, instructors, and the gender of students. The sample consists of 78 students undertaking financial accounting courses at Zayed University during the summer semester. A simple difference in means statistics test shows that there is no statistically significant difference between the students' paper-based and computer-based scores. However, benchmark regression analysis showed that males performed better than females on CBT, and females outperformed males on PBT. The paper provides evidence from the Gulf region as of how technology-based assessment is affected by the gender, which matter needs to be addressed by university instructors working in Middle East universities.

Keywords: Financial Accounting, Examination Mode, Gender, United Arab Emirates.

INTRODUCTION

Assessment is an essential tool in the educational process because it measures the students' understanding (Joosten-ten Brinke et al., 2007). In higher education institutions, the assessment of academic progress takes different formats such as essays, dissertations, examinations, assignments, projects, and presentations. Sim, Holifield, & Brown (2004) had identified more than fifty different techniques used within higher education for assessment purposes; nevertheless, the most commonly used tool is an examination. As part of e-learning trend, the computer-based tests (CBT) became more prevalent than paper-based tests (PBT) in the domain of educational assessment as changes are made in assessment methodologies reflect practical changes in pedagogical methods (OECD, 2010). Hence, the CBT became dominant in different types of examinations such as standardized tests (e.g., GRE, GMAT, and SAT), accounting professional certifications (e.g., CPA, CIA, and CMA) and college examinations (Bull, 1999; Conole & Warburton, 2005). Nevertheless, the comprehensive implementation of CBT mode has been hindered by questions such as the equivalence of CBT mode to PBT mode (Akdemir & Oguz, 2008; Chua & Don, 2013).

The CBT provides several advantages such as proposing a solution to mechanize the assessment process (Triantafillou, Georgiadou, & Economides, 2008); reducing paper consumption which indirectly reduces greenhouse gases and energy consumption (DeRosa,

2007); assisting students to evaluate their strengths and weaknesses (Kaklauskas et al., 2010); and providing quantitative improvements in assessment for academics and tutors (Singleton, 1997). These improvements can be noticed in reducing preparation time and cost, enhancing the test security, analyzing the results easily, keeping record for item analysis and reliability of scoring (Gvozdenko & Chambers, 2007; Sanni & Mohammad, 2015; Singleton, Horne, & Thomas, 1999; Smith & Caputi, 2005; Tippins, 2011), increasing efficiency (Halcomb et al., 1989; Karay et al., 2015; Lee & Barua, 1999; Zakrzewski & Bull, 1998), and providing instantaneous feedback to students (Bugbee, 1996; Bull & McKenna, 2003; Butler, 2003; Erturk, Ozden & Sanli, 2004; Zandvliet & Farragher, 1997). Moreover, the CBT offer enormous scope for innovations in testing and assessment (Bennett, 1998; Chatzopoulou & Economides, 2010) and measure complex form of knowledge and reasoning which is not possible through traditional methods (Bodmann & Robinson, 2004).

On the other end, CBT might have some limitations such as lacking underlining or making notations on computer screen, looking at the computer screen for a long time, and anxiety from changing the exam mode from PBT to CBT (Butler, 2003). Nonetheless, prior research argued that students still prefer CBT more than PBT as performance assessment because CBT is more promising, credible, objective, fair, interesting, fun, fast and less difficult or stressful (Sambell, Sambell, & Sexton, 1999; Croft et al., 2001).

This study is an attempt to gain insight into student perceptions of computerized testing in an introductory financial accounting course at Zayed University in the United Arab Emirates with majority students from female gender in the Middle East, which is dominated by Arabic and Islamic culture. The study hypothesizes that once controlling for individual characteristics, student performance in each exam mode might be related to gender difference. It means male students might outperform their female counterparts in CBT mode and female students might outperform their male counterparts in PBT mode. To test these hypotheses, a regression analysis has been made on data collected from 78 students from both genders in four tests administered to four sections. The data consisted of 16 test events and a total of 9,672 data points.

The remainder of the article is organized as follows: section two reviews prior literature review and develops our hypotheses. Section three describes our research design and data collection. Section four presents the findings together with a discussion of gender impact on students' performance. The final section summarizes and concludes the paper.

LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

The acceptance and use of CBT are increasing each year. However, one unresolved problem associated with using CBT is performance bias due to examinees' differences (FairTest, 2007). The examinees' performance might be affected by computer experience and familiarity (Fulcher, 1999; McDonald, 2002). Mazzeo & Harvey (1988) and Mead & Drasgow (1993) tried to measure the examinee's performance on CBTs as well as PBTs and concluded conflicting results based on the type of test (i.e., power vs. speeded tests, or personality vs. placement tests). For example, Mazzeo & Harvey concluded that scores from CBT might not be comparable with PBT versions due to the greater ease with which answers can be indicated when the test is administered by computer. However, Mead and Drasgow focused on correlations across formats and found no effect for carefully constructed power tests, but a substantial effect for speeded tests. Also, some studies revealed the significant difference between the two testing modes based on test scores (Choi, Kim, & Boo, 2003; Scheuermann & Björnsson 2009), while other studies reported opposite results (Al-Amri, 2008; Boo, 1997). These inconsistent findings led some

scholars to suggest for conducting systematic studies on CBT on testing and to carefully check its reliability (Johnson & Green, 2004; Wang & Shin, 2010) and validity (Al-Amri, 2008; Alderson, 2000) before opting for CBTs.

Furthermore, the extant literature reported mixed results about the gender impact on the performance of the examinees. For example, Gallagher, Bridgeman, & Cahalan, (2000); and Leeson, (2006) asserted the existence of gender effect on examination mode. Also, Oduntan et al. (2015) concluded that male students outperformed female students on CBT, whereas Jalali, Zeinali, & Nobakht (2014) emphasized that female students outperformed male students in both modes. Furthermore, Wallace and Clariana (2005) found that male students outperformed their female peers on the initial assessment regardless of the test mode, whereas female students using CBT outperformed male students on the final assessment. On the contrary, Alexander et al., (2001) noted no gender differences in performance of both test modes.

In addition, several studies presented mixed findings of the performance of examinees from both genders based on their previous usage of computers. For example, Badagliacco (1990); Ogletree & Williams (1990); Shashaani (1994); and Makrakis & Sawada (1996) argued that male students express more interest than female students in computing, have greater access to computers, enjoy working with computers, and have more confidence in their ability to work with computers. Other studies stated that males had more positive perceptions than females towards the use of a digital library (Koohang, 2004) and the use of web-based instructions at an open university (Enoch & Soker, 2006). In addition, other studies stated that males were thought to be more positive towards computers than females concerning computer anxiety (Chou, 2003; Tsai, Lin, & Tsai, 2001), computer self-efficacy (Durndell & Thomson, 1997; Schaumburg, 2004) and computer attitude (Liaw, 2002). On the contrary, several studies found no significant differences in gender concerning computer anxiety, computer liking, and confidence about using computers (Clariana & Wallace, 2002; Francis, 1994), and perceived the usefulness of computers (Shashaani, 1997). In fact, Shashaani investigated the attitudes of 202 college students from both genders and concluded that students are aware of the importance of computer knowledge for obtaining jobs, saving time and work, processing data, and solving problems. Furthermore, Zhang and Espinoza (1998) reported that both genders recognized the usefulness of computers and emphasized their perception of advanced levels of computer technologies as significant predictors of the desirability of learning computing skills.

The above conclusion has been contradicted by several studies which found no difference in the assessment scores of the two modes across a variety of education levels, disciplines, and geographies such as K–12 students (Wang et al., 2008), educational psychology students (Bodmann and Robinson, 2004), GRE students (Macrander et al., 2012), high-school students and medical students in Germany (Schroeders and Wilhelm, 2010; Hochlehnert et al., 2011; Karay et al., 2015); computer science students (Karadeniz, 2009), engineering students (Özalp-Yaman and Çağiltay, 2010), education and language, and literature students in Turkey (Akdemir & Oguz, 2008); biology students and education students in Malaysia (Chua et al., 2013; Chua, 2012); and anatomy students in Australia (Meyer et al., 2016). Some studies have compared student populations from multiple disciplines and found the same results (Bayazit and Askar, 2012; Neuman and Baydoun, 1998).

The outcomes of the above studies encouraged us to investigate the gender impact on the performance of first-year college students on the two modes of examinations. Accordingly, the following hypotheses are proposed:

H1 Ceteris paribus, student performance in CBT mode is positively related with male gender.

H2 Ceteris paribus, student performance in PBT mode is positively related with female gender.

In doing so, this study contributes to the limited research in the area of CBT in accounting education (Apostolou, Blue, & Daigle, 2009). Moreover, the study contributes to the current literature by examining the differences in student performance based on the mode of the examination whether PBT or CBT. The study also tests whether these differences exist even after controlling for individual specific characteristics like gender and ability. The following section explains the process of data collection and analysis, which is followed by findings and discussion.

RESEARCH DESIGN

This section describes the methodology of research design along with the approach that has been used in collecting data from students in an introductory-level course in financial accounting to measure their performance in CBTs as well as PBTs.

METHODOLOGY

Primary data has been collected for undergraduate students at Zayed University with majority students from female gender, which is dominated by Arabic and Islamic culture. Due to cultural norms, the university campus is split into two identical campuses: one for male students and another one for female students; however, both genders have same instructors and curricula. The data had been collected from 78 undergraduate students (62 female students and 16 male students) across four sections during Summer semester of the year 2015. Students registration for Summer term courses is voluntary.

To our best of knowledge, the sample size is big enough to draw significant conclusions. Additionally, the number of female students outweighs the number of male students because the number of female students in this university is ten times higher than male students; therefore, it seems normal to have more female students in the sample. Interestingly, other studies have also used similar small samples. For example, Ford, Vitelli, and Stuckless (1996) used a sample of 52 students; Clariana & Wallace (2002) used a sample of 105 students; Lumsden et al., (2004) used a sample of 93 students; Anakwe (2008) used a sample of 54 students. Accordingly, our sample size is big enough to generalize our findings among different academic institutions in similar cultures and regions.

The 78 students have been distributed in four sections, where three of them assigned for female students exclusively and the other one for male students. Due to Arabic and Islamic cultural norms common throughout the region, course sections at the undergraduate level are segregated by gender. The three female sections have 18, 23, and 21 students. and the male section has 16 students. The registration of students into these sections had been controlled by the registrar office which left no room for instructor bias. However, according to the university policy male sections are located in a separated campus from female students; thus, male students are forced to be seated together in one section but female students are free to choose one of the three sections. All the four sections were opened for any student from the relevant gender to register because the student's number in each section was way below the maximum number which is 35 students. Thus, there is no bias in gender. On the other hand, all students in the four sections are full-time students with age 18-22 years which eliminates the age bias because all students have either limited or no prior experience in accounting.

The first two sections used to meet for three and a half hours in the morning (0800–1130) while the other two sections used to meet for three and a half hours in the afternoon (1200–1530) five days a week for a three-week and two-day term during the holy month of Ramadan. Some of the students (male and female) were repeating the course because they had failed it previously. The introductory-level course of financial accounting is required of all students matriculated in the college of business at this university. Thus, some students from other majors find this course hard to pass in the first time; therefore, they failed it and took it again, which justifies the reason for having repeater students. Within the local area, this university is unique in awarding four credit hours for introductory financial accounting courses. In this course, some of the basic topics in accounting are covered such as the accounting cycle; financial reporting; current assets accounts such as cash, receivables, and merchandise inventory; and long-term assets accounts such as plant assets, natural resources, and intangible assets. The instructor explains the technical work of each topic as well as the best practices using the assigned textbook as well as supplementary materials such as academic articles, other textbooks, and online software.

The four sections have been taught by two faculty members of equal ranks (Assistant Professor of Accounting). Both faculties had several years of experience in teaching the introductory financial accounting course to both genders at Zayed University during the Summer semester. Hence, they closely collaborated to deliver the same materials in all sections using the same pedagogies. Moreover, they provided the same tests which are already provided to other students in other semesters. Since the financial accounting courses is a college required course and the university normally offers multi sections of the course every semester, several instructors are responsible for teaching that courses. Nevertheless, to ensure rigorous in our research we controlled for the possible effect of the instructor on the student performance by having a dummy variable for the instructors. The dummy variable takes the value of 1 for instructor X and value of 0 for instructor Y.

The tests were identical in every way except for the mode of delivery in order to assess students' understanding of the aforementioned topics. Each test includes a variety of questions namely mathematical, conceptual, and mixed questions. The mathematical questions are used to check students' understanding in computing the cost of merchandise inventory or assets depreciation or net receivable etc. The conceptual questions are used to check students' understanding about journalizing, posting, and adjusting the accounting balances. The mixed questions are used to exam students' understanding in creating one of the financial statements and posting the accounts with appropriate names and balances. These questions have been placed in different formats such as multiple choice, multiple answers, true/false, drop-down menu, and fill-in-the-blank questions.

To avoid bias in grading, students were allowed to backtrack the questions in the CBT mode in order to have the same opportunity offered in PBT mode. Moreover, answering questions was designed to be either right or wrong and long questions were split into smaller parts where each part is graded separately in order to give a score for each part if the question has multiple computations. The computation process was very simple and can be made with a simple calculator; which doesn't affect students' performance and eliminate any bias in correction. Furthermore, any answer outside the exam paper or was not entered on the computer has been ignored. Each instructor corrected the submitted answers of his sections in order to give students access to their submission to review their scores and ensure no bias has occurred in grading.

Multiple questions of each type were included on each test. Questions, formatting, and answering options were the same regardless of test mode. The first test had a relative weight of 17 percent toward the final grade. The remaining three tests had a relative weight of 25 percent for each one. The remaining 8 percent was allocated for class participation and attendance.

Students were given mock tests in both modes, CBT and PBT, which had a twofold objective: firstly, to familiarize students with types of questions that would be on the tests, and secondly, to ensure that students were familiar with technical and operational elements of both modes. All students had prior experience with both test modes through other courses in this university. All students had completed the university's colloquy (general education) program during which they received explicit instruction, and gained actual experience, in CBT.

Each instructor taught a morning section and an afternoon section. The morning sections consisted of a total of 41 female students (18 female students for instructor (1) and 23 female students for instructor (2)). The afternoon sections consisted of a total of 37 students (21 female students for instructor (1) and 16 male students for instructor (2)). The breakdown of students by section is summarized in Table (1).

Table 1
COURSE SECTIONS

	Section 1	**Section 2**	**Section 3**	**Section 4**
Instructor	1	2	1	2
Time	Morning	Morning	Afternoon	Afternoon
Students	18 Females	23 Females	21 Females	16 Males

For each section, two of the four tests were administered in a traditional PBT mode and two of the tests were administered via computer using Blackboard Respondus Lockdown browser. The test modes alternated among sections so that two sections completed test (1) on paper and two sections completed the same test on computer. These modes were swapped for each subsequent test. Consequently, sections (1) and (3) was administered test (1) on paper, test (2) on computer, test (3) on paper, and test (4) on computer. The testing regime was opposite for sections (2) and (4). The test administration sequence is presented in Table (2). There was one week of instruction before the first test and, subsequently, one week of instruction between each test. The tests were administered on the same day for all classes.

Table 2
TEST ADMINISTRATION SEQUENCE

Exam	**Course**	**Instructor 1**	**Instructor 2**
Exam 1	Morning class	Paper	Computer
	Afternoon class	Computer	Paper
Exam 2	Morning class	Computer	Paper
	Afternoon class	Paper	Computer
Exam 3	Morning class	Paper	Computer
	Afternoon class	Computer	Paper
Exam 4	Morning class	Computer	Paper
	Afternoon class	Paper	Computer

To ensure the integrity of the assessment and mitigate examination malpractice (Sanni and Mohammad 2015) during the CBT mode, the Respondus LockDown browser function on Blackboard was deployed where questions were randomized and students were able to go back

and forth among the questions. Respondus LockDown browser is a custom web browser that prevents students from printing, copying, accessing another URL, and opening other applications while Blackboard online assessment is ongoing. When an assessment is started, students are locked into it until they submit their answers for grading (Blackboard 2016). Nevertheless, the PBT questions were not randomized because it was not feasible to randomize the questions on paper so much as on computer. Traditional invigilation techniques were used to ensure the integrity of the assessment environment during the PBT (i.e., spaced seating, faculty observation).

EMPIRICAL RESULTS

The test-performance data from the four tests administered to the four sections have been analyzed and compared. The data consisted of 16 test events and a total of 9,672 data points. Initially, we assumed that students' performance on PBT mode would be similar to their performance in CBT mode regardless of their gender, class timing, teaching instructor, and experience with course materials.

Analysis

The dependent variable examined was student performance as determined by scores on each question. Both modes of exams have the same scores for questions because the same questions in each test were given to all sections except for giving two sections the questions on computer and the other two sections on paper. Accordingly, there is no difference in the scoring system; however, we hypothesized that male students might get higher scoring in CBT mode and female students receive higher scoring on PBT mode.

The independent variables were student gender (dummy variable, which took the value of (1) for male students and (0) for female students); class timing (dummy variable, which took the value of (1) for morning class and (0) for afternoon class); test mode (dummy variable, which took the value of (1) for PBT and (0) for CBT); instructor (dummy variable); question type (two dummy variables for conceptual and mathematical questions, with mixed questions as the reference category); course repetition (dummy variable, which took the value of (1) for students repeating the course and (0) for new students); and cumulative GPA prior to the course (continuous variable). The summary statistics are presented in Tables (3A) and (3B).

Table 3A
SUMMARY STATISTICS

Variable	Obs.	Mean	Std. dev.	Min.	Max.
Male (dummy)	9672	0.205128	0.403816	0	1
Morning class (dummy)	9672	0.525641	0.499368	0	1
PBT (dummy)	9672	0.495864	0.500009	0	1
Instructor 1 (dummy)	9672	0.500000	0.500026	0	1
Score	9672	72.47457	42.99048	0	100
GPA	9672	2.812179	0.530607	1.96	3.8
Repeat (dummy)	9672	0.294872	0.456009	0	1
Conceptual (dummy)	9672	0.572581	0.49473	0	1

Table 3B PERCENTAGES OF CONCEPTUAL QUESTIONS BY EXAM	
Exam	**Conceptual**
1	0.898
2	0.130
3	0.058
4	0.657

In Table (3A) The obs column refers to the number of observations used in the regressions by STATA in the benchmark regressions. As there are 78 students in the sample, there are 78 GPAs. However, as there is more than 1 observation for each student, and there are no missing observations, there are 9672 observations used in regressions. Similarly, The 0/1 dummy measures' means indicate simply the percentage of men, percentage of students in morning classes, percentage of students taking PBTs, percentage of students with Instructor 1, and percentage of repeaters. The data has been analyzed using difference-in-means tests as well as a multiple regression analysis. Moreover, robustness checks have been conducted through clustering by student.

Difference-In-Means Tests

The analysis began with a basic difference-in-means test and the results are presented in Table (4).

Table 4 DIFFERENCE-IN-MEANS TEST			
Variable	**Group 0 mean**	**Group 1 mean**	**Two tailed t-test, p value**
Paper (= 1 for paper based exams)	71.943	73.015	0.22
Male (=1 for male students)	72.081	73.999	0.08*
Morning (=1 for morning classes)	74.228	70.892	0.00***
Repeat (=1 fpr repeat students)	73.066	71.060	0.04**
Instructor 1 (=1 if the instructor is instructor 1)	74.242	70.708	0.00***
Conceptual (=1 if the question is conceptual)	63.384	79.260	0.00***

*Significant at 10% level.
**Significant at 5% level.
***Significant at 1% level.

Without any additional controls, we found no statistically significant difference between the two modes in terms of the students' results. However, many factors seem to have affected test scores. The male students (compared to female students) and non-repeating students (compared to repeating students) had statistically significantly higher test scores. Similarly, students in the morning sections performed statistically significantly worse than those in the afternoon sections. Overall, students scored significantly lower on the mathematical or mixed questions compared with conceptual questions, and significantly worse under instructor (1) than instructor (2). We should also note that the aforementioned differences-in-means were all significant at the 1 percent level, with the exception of gender difference, which was significant at the 10 percent level, and the repeating student effect, which was only significant at the 5 percent level.

Regressions

The above framework did not allow for controlling multiple factors simultaneously; therefore, a regression analysis has been made. The type of collected data allowed us to control for unobserved heterogeneity in three dimensions: the student dimension, the question dimension, and the test dimension. Given the presence of exam and question invariant variables like GPA, We opted for the random effects estimations (generalized least squares). The findings of the regressions are presented in Table (5).

Table 5
REGRESSIONS

Dependent Variable: Score in Question out of 100	**(1)**	**(2)**	**(3)**
GPA	16.713***	15.311***	14.933***
	(-8.61)	(8.38)	(8.09)
Paper Based (Dummy)	4.207***	6.499***	-3.553***
	(3.83)	(5.9)	(-2.89)
Male (Dummy)	-2.636	4.773*	-10.476***
	(-1.12)	(1.82)	(-4.00)
Repeat (Dummy)	-0.312	1.799	0.723
	(-0.14)	(0.83)	(0.34)
Conceptual Question (Dummy)	-0.23816	-0.23839	-0.23791
	(-0.96)	(-0.97)	(-0.97)
Paper*Male	-9.546***	-16.177***	-2.770***
	(-17.81)	(-22.39)	(-3.80)
Paper*GPA	-0.24437	0.802**	0.802**
	(-0.68)	(2.2)	(2.2)
Paper*Repeat	3.682***	2.608***	2.608***
	(8.63)	(6.07)	(6.07)
Paper*Conceptual	1.353***	1.354***	1.351***
	(5.82)	(5.88)	(5.86)
Exam 2 (Dummy)	-32.133***	-33.767***	-30.417***
	(-94.22)	(-94.10)	(-84.28)
Exam 3 (Dummy)	-11.063***	-11.063***	-11.064***
	(-31.83)	(-32.13)	(-32.13)
Exam 4 (Dummy)	-10.172***	-11.807***	-8.455***
	(-38.53)	(-40.98)	(-29.09)
Instructor 1 (Dummy)		9.676***	
		(4.41)	
Paper*Instructor1		-6.706***	
		(-13.54)	
Morning (Dummy)			-10.188***
			(-4.69)
Paper*Morning			6.705***
			(13.54)
_Constant	26.860***	24.533***	37.773***
	(4.59)	(4.47)	(6.32)
Number of Obs	9672	9672	9672
Wald Statistic		15742.75	15747.09

z-statistics are in parentheses. (*** Significant at 1 percent level; ** Significant at 5 percent level; *Significant at 10 percent level)

The dependent variable for each and every regression is the score on question out of 100. As partial credit was given, the aforementioned variable is a continuous variable. The independent variables include the grade point average of the student at the start of the semester, as well as a number of dummy variables which take the binary values of 1 or 0. In addition, some interaction terms have been used to control for the existence of joint effects, which includes our main variable of interest, paper*male, that investigates whether males (females) do better (worse) on PBT. Finally, a dummy variable for every exam was used to control for exam-specific effects. Given the presence of high multicollinearity between the morning dummy and instructor dummy, these variables are omitted in the first specification, and then later added one-by-one to second and third specifications. Our empirical results showed that male students performed better than female students on CBT mode (Regression 1-3) which supports our first hypothesis.

H1 Ceteris paribus, student performance in CBT mode is positively related with male gender.

Furthermore, the empirical results showed that female students performed better than male students on PBT mode (Regression 1-3) which supports our second hypothesis.

H2 Ceteris paribus, student performance in PBT mode is positively related with female gender.

Moreover, students entering the course with higher GPAs had higher scores regardless of the test mode (Regressions 1-3). Students in the morning sections performed worse than students in the afternoon sections regardless of the test mode (Regression 3). We should also note that all students did significantly better on conceptual questions with PBT, and finally, repeating students also did better on PBT (Regressions 1-3).

DISCUSSION

The findings of the present study showed similarities as well as differences in several aspects of students' performance based on gender, prior experience, class timing, and instructors as compared with the findings of other studies conducted in other regions. Firstly, our results corroborate the previous studies of Fulcher (1999) and McDonald (2002) which confirm that students prior experience with computer positively affect their performance on CBT mode. Moreover, we partially support the previous studies of Badagliacco (1990); Ogletree & Williams (1990); Shashaani (1994); and Makrakis & Sawada (1996), which concluded that male students have more confidence in their ability to work with computers. Secondly, our findings align with Gallagher et al., (2000), and Leeson (2006) who emphasized the existence of gender impact on examination mode. Our study showed that, when controlling for gender, male students outperformed female students on CBT, while female students outperformed male students on PBT. Moreover, our results corroborate Oduntan (2015) which emphasized that male students outperformed female students in CBT mode. Thirdly, our findings from a Middle Eastern university corroborated prior findings with respect to the validity (Al-Amri, 2008; Alderson, 2000) and reliability (Johnson & Green, 2004; Wang & Shin, 2010) of CBT compared to PBT. Fourthly, the outcomes of our study revealed the instructor's impact on student performance which corroborates the outcomes of Apostolou et al., (2009).

Although not central to the research question, our analysis demonstrated that students who were repeating the course scored better on PBT than they did on CBT. Students in the morning sections, in general, performed worse than their peers in the afternoon sections. It is

important to note, however, that the generalizability of the findings of this study is limited by the cultural norms in students allocation among sections and relatively small sample size. As such, replicating the study within the institution across courses, faculty members, and disciplines to control for some of the unique results we encountered is warranted prior to the wholesale adoption of CBT. We also suggest that the timing of classes be examined to determine whether this does in fact impact test performance or whether this result was anomalous due to seasonal shifts in the students' extracurricular life activities (e.g., the holy month of Ramadan).

CONCLUSION

This study is an attempt for examining the equivalency between CBT and PBT as alternative modes of examination. In this attempt, we hypothesized that male students might outperform female students in CBT and vice-versa. These hypotheses have been tested using the examinations scores in four different sections of financial accounting course at Zayed University where male students are separated from female due to cultural norms. The number of students is 78 undergraduate students comprised of 62 female and 16 male across the four sections with 16 test events during summer semester. Interestingly, the student performance in CBT mode was positively related to male gender, while student performance in PBT mode was positively related with female gender. Furthermore, students with prior experience in accounting, and the class timing, as well as the instructors, affected students' performance from both genders.

This study contributes to the literature in two ways. Firstly, while prior research has examined the gender impact on students' performance using a sample drawn from developed countries, this study takes the first step to fill evident gap in literature by exploring a unique setting that is often ignored by many scholars. This helps to widen our knowledge of accounting education in the Middle Eastern culture. Secondly, the prior inconsistent findings in research led some scholars to suggest for conducting systematic studies on CBT on testing and to carefully check its reliability and validity before opting for CBTs. Hence, the findings of our study suggest that concerns about the validity and reliability of CBT compared to PBT seem unwarranted. CBT has been shown repeatedly to be an equally valid and reliable test mode for student assessment in the academic environment.

This study is subject to a number of limitations. First, this study is limited by specific cultural norms which enforce the gender segregation in education. Secondly, the sample size is slightly small and has more female students than male students because the majority of students in this university are female students. Thirdly, the study was made during the summer semester which has a relatively shorter period than a normal semester (Fall/Spring). As such, replicating the study within the institution across courses, faculty members, and disciplines to control for some of the unique results we encountered is warranted prior to generalizing the adoption of CBT.

REFERENCES

Akdemir, O. & A. Oguz (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education, 51*(3), 1198-1204.

Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. Essex Graduate Student Papers in Language & Linguistics, 10, 22-44.

Alderson, J. (2000). Technology in testing: The present and the future. *System, 28*(4), 593-603.

Alexander, M., Bartlett, J., Truell, A. & Ouwenga, K. (2001). Testing in a computer technology course: An investigation of equivalency in performance between online and paper and pencil methods. *Journal of Career and Technical Education*, *18*(1), 69-80.

Anakwe, B. (2008). Comparison of Student Performance in Paper-Based Versus Computer-Based Testing. *Journal of Education for Business*, *84*(1), 13-17.

Apostolou, B., Blue, M. & Daigle, R. (2009). Student perceptions about computerized testing in introductory managerial accounting. *Journal of Accounting Education, 27*(2), 59-70.

Badagliacco, J. (1990). Gender and race differences in computing attitudes and experience. *Social Science Computer Review*, *8*(1), 42-64.

Bayazit, A. & Aşkar, **V.** (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Education Review*, 13(2), 219-226.

Bennett, R. (1998). *Reinventing assessment: speculations on the future of large-scale educational testing*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Bodmann, S. & Robinson, D. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51-60.

Boo, J. (1997). *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished doctoral dissertation, University of Iowa.

Bugbee Jr, A. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3), 282-299.

Bull, J. & McKenna, C. (2003). *A Blueprint for Computer-Assisted Assessment*. New York: Routledge.

Bull, J. (1999). Computer-assisted assessment: Impact on higher education institutions. *Educational Technology & Society*, *2*(3), 123-126.

Butler, D. (2003). *The Technology Source archives at the University of North Carolina, January /February.* Retrieved on June 26, 2017 from http://technologysource.org/article/impact_ of _computer-based_testing _on_student _ attitudes_and_behavior/.

Chatzopoulou, D. & Economides, A. (2010). Adaptive assessment of student's knowledge in programming courses. *Journal of Computer Assisted Learning*, *26*(4), 258-269.

Choi, I., Kim, K. & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.

Chou, C. (2003). Incidences and correlates of internet anxiety among high school teachers in Taiwan. *Computers in Human Behavior*, 19, 731-749.

Chua, Y. & Don, Z. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29*(5), 1889-1895.

Chua, Y. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, 28(5), 1580-1586.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593-602.

Conole, G. & Warburton, B. (2005). A Review of computer-assisted assessment. *Research in Learning Technology*, 13(1), 17-31.

Croft, A., Danson, M., Dawson, B. & Ward, J. (2001). Experiences of using computer-assisted assessment in engineering mathematics. *Computers and Education*, 27, 53-66.

DeRosa, J. (2007). *The green PDF: Reducing greenhouse gas emissions one ream at a time.* Retrieved on June 24, 2017 from http://www.scribd.com/doc/60779195/The-Green-PDF-Revolution.

Durndell, A. & Thomson, K. (1997). Gender and computing: A decade of change? *Computers & Education*, *28*(1), 1-9.

Enoch, Y. & Soker, Z. (2006). Age, gender, ethnicity and the digital divide: University students' use of web-based instruction. *Open Learning*, *21*(2), 99-110.

Erturk, I., Ozden, Y. & Sanli, R. (2004). Students' perceptions of online assessment: A case study. *Journal of Distance Education*, *19*(2), 77-92.

FairTest: *The National Center for Fair and Open Testing, (2007). Computerized Testing: More Questions than Answers*. Retrieved on June 26, 2017 from http://www.fairtest.org/ computerized-testing-more-questions-answers

Ford, B., Vitelli, R. & Stuckless, N. (1996). The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior*, *12*(1), 159-166.

Francis, L. (1994). The relationship between computer-related attitudes and gender stereotyping of computer use. *Computers and Education*, 22, 283-289.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal*, *53*(4), 289-299.
Gallagher, A., Bridgeman, B. & Cahalan, C. (2000). The effect of computer-based tests on racial/ethnic, gender, and language groups. *ETS Research Report Series*, *2000*(1), 1-17.
Gvozdenko, E. & Chambers, D. (2007). Beyond test accuracy: Benefits of measuring response time in computerised testing. *Australasian Journal of Educational Technology*, *23*(4).
Halcomb, C., Chatfield, D., Stewart, B., Stokes, M., Cruse, B. & Weimer, J. (1989). A computer-based instructional management system for general psychology. *Teaching of Psychology*, 16(3), 148-151.
Hochlehnert, A., Brass, K., Moeltner, A. & Juenger, J. (2011). Does medical students' preference of test format (computer-based vs. paper-based) have an influence on performance? *BMC Medical Education*, *11*(1), 89.
Jalali, S., Zeinali, M. & Nobakht, A. (2014). Effect of goal orientation on EFL learners' performances in CBT and PBT across gender. *Procedia-Social and Behavioral Sciences*, 98, 727-734.
Johnson, N. & Green, S. (2004). On-line assessment: The impact of mode on students performance. *Proceedings of the British Educational Research Association Annual Conference (BERA)*, Manchester, UK.
Joosten-ten Brinke, D., Van Bruggen, J., Hermans, H., Burgers, J., Giesbers, B., Koper, R. & Latour, I. (2007). Modeling assessment for re-use of traditional and new types of assessment. *Computers in Human Behavior*, *23*(6), 2721-2741.
Kaklauskas, A., Zavadskas, E., Pruskus, V., Vlasenko, A., Seniut, M. & Kaklauskas, G. (2010). Biometric and intelligent self-assessment of student progress system. *Computers & Education*, *55*(2), 821-833.
Karadeniz, Ş. (2009). The impacts of paper, web and mobile based assessment on students' achievement and perceptions. *Scientific Research and Essays*, *4*(10), 984-991.
Karay, Y., Schauber, S., Stosch, C. & Schüttpelz-Brauns, K. (2015). Computer versus paper: Does it make any difference in test performance? *Teaching and Learning in Medicine*, *27*(1), 57-62.
Koohang, A. (2004). Students' perceptions toward the use of the digital library in weekly web-based distance learning assignments portion of a hybrid program. *British Journal of Educational Technology*, *35*(5), 617-626.
Lee, B. & Barua, A. (1999). An integrated assessment of productivity and efficiency impacts of information technology investments: Old data, new analysis and evidence. *Journal of Productivity Analysis*, *12*(1), 21-43.
Leeson, H. (2006). The mode effect: A literature review of human and technological issues. *International Journal of Testing*, 6(1), 1-24.
Liaw, S. (2002). An Internet survey for perceptions of computers and the World Wide Web: Relationship, prediction, and difference. *Computers in Human Behavior*, 18, 17–35.
Lumsden, J., Sampson, J., Reardon, R., Lenz, J. & Peterson, G. (2004). A comparison study of the paper-and-pencil, personal computer, and internet versions of holland's self-directed search. *Measurement and Evaluation in Counseling and Development*, *37*(2), 85-94.
Macrander, C., Manansala, R., Rawson, S. & Han, J. (2012). The difference in performance between computer and paper administered tests in a stressful environment. *Journal of Advanced Student Science* 1 (Spring). Retrieved Jun. 26, 2017, from http://jass.neuro. wisc.edu/2012/01/Lab%20602%20Group%2014%20 Physiology%20435%20Final%20Paper.pdf
Makrakis, V. & Sawada, T. (1996). Gender, computers and other school subjects among Japanese and Swedish students. *Computers in Education*, *26*(4), 225-231.
Mazzeo, J. & Harvey, A. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Report No. 88-8). Princeton, NJ: Educational Testing Service.
McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education Journal*, 39(3), 299-312.
Mead, A. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.
Meyer, A., Innes, S., Stomski, N. & Armson, A. (2016). Student performance on practical gross anatomy examinations is not affected by assessment modality. *Anatomical Sciences Education*, *9*(2), 111-120.
Neuman, G. & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, *22*(1), 71-83.
Oduntan, O., Ojuawo, O. & Oduntan, E. (2015). A comparative analysis of student performance in paper pencil test (PPT) and computer-based test (CBT) examination system. *Research Journal of Educational Studies and Review*, *1*(1), 24-29.

OECD. (2010). *PISA Computer-based assessment of student skills in science*. Retrieved Jun. 15, 2017, from http://www. oecd.org/ publishing/corrigenda

Ogletree, S. & Williams, S. (1990). Sex and sex-typing effects on computer attitudes and aptitude. *Sex Roles*, 23(11/12), 703-712.

Özalp-Yaman, Ş. & Çağiltay, N. (2010). Paper-based versus computer-based testing in engineering education. *IEEE EDUCON Education Engineering*, 1631-1637.

Sambell, K., Sambell, A. & Sexton, G. (1999). Student perceptions of the learning benefits of computer-assisted assessment: a case study in electronic engineering. In Brown, S., Race. P. & Bull. J. (Eds.), *Computer assisted assessment in higher education*. London: Kogan Page.

Sanni, A. & Mohammad, M. (2015). Computer based testing (CBT): An assessment of student perception of JAMB UTME in Nigeria. *Computer* 6 (2). Retrieved Jun. 26, 2017, from http://www.cisdijournal.net/uploads/ V6N2P3_-_CISDIAR_JOURNAL.pdf

Schaumburg, H. (2004). Laptops in der Schule–ein Weg zur Ü berwindung des Digital Divide zwischen Jungen und Mädchen? (Laptop computers in the classroom: A way to overcome the technological gender gap among students?). *Zeitschrift fur Medienpsychologie*, 16, 142-154.

Scheuermann, F. & Björnsson, J. (2009). The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing. *Joint Research Centre*, Italy.

Schroeders, U. & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*(4), 284-292.

Shashaani, L. (1994). Gender differences in computer experience and its influence on computer attitudes. *Journal of Educational Computing Research*, *11*(4), 347-367.

Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. *Journal of Educational Computing Research*, *16*(1), 37-51.

Sim, G., Holifield, P. & Bandrown, M. (2004). Implementation of computer assisted assessment: Lessons from the literature. *Research in Learning Technology*, *12*(3), 217-233.

Singleton, C. (1997). Computer-based assessment of reading. In Beech, J.R. & Singleton, C.H. (Eds.), *Psychological assessment of reading*. London: Routledge.

Singleton, C., Horne, J. & Thomas, K. (1999). Computerised baseline assessment of literacy. *Journal of Research in Reading*, *22*(1), 67-80.

Smith, B. & Caputi, P. (2005). Cognitive interference model of computer anxiety: Implications for computer based assessment. *Computers in Human Behavior*, 21, 713-728.

Tippins, N. (2011). Overview of technology-enhanced assessments. In *Technology-Enhanced Assessment of Talent*, edited by Tippins, N.T., Adler, S. & Kraut, A.I., 1-18. San Francisco, CA: Jossey-Bass.

Triantafillou, E., Georgiadou, E. & Economides, A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers and Education*, *50*(4), 1319-1330.

Tsai, C., Lin, S. & Tsai, M. (2001). Developing an internet attitude scale for high school students. *Computers and Education*, 37, 41-51.

Wallace, P. & Clariana, R. (2005). Gender differences in computer-administered versus paper-based tests. *International Journal of Instructional Media*, *32*(2), 171.

Wang, H. & Shin, C. (2010). Comparability of computerized adaptive and paper-pencil tests. *Measurement and Research Service Bulletin*, 13, 1-7.

Wang, S., Jiao, H., Young, M., Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68(1), 5-24.

Zakrzewski, S. & Bull, J. (1998). The mass implementation and evaluation of computer-based assessments. *Assessment and Evaluation in Higher Education*, 23(2), 141-152.

Zandvliet, D. & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, *29*(4), 423-438.

Zhang, Y. & Espinoza, S. (1998). Relationships among computer self-efficacy, attitudes toward computers, and desirability of learning computing skills. *Journal of Research on Computing in Education*, *30*(4), 420-438.