IMPACT OF COMPUTER-PREDICATED EDUCATIONAL ACHIEVEMENT TEST ON TEST PERFORMANCE AND TEST TAKERS' MOTIVATION

Amarathunga, P. A. B. H.[1] and Pathiratne, S[2]

[1]*Faculty of Business Studies and Finance, Wayamba University of Sri Lanka*
buddhini@wyb.ac.lk
[2]*Faculty of Computing, ESOFT Metro Campus, Sri Lanka*

Abstract

There has been an incrementing interest in recent years in developing and utilizing computer-predicated tests in edifying assessment. To supersede paper-predicated tests with computer-predicated ones, the standards for developing computerized-assessment (International Test Commission., 2004) requires equipollent test scores to be established for the incipient computer-predicated test and the conventional paper-predicated test. However, in most test mode commensurability studies, the genuine test items used have been identical, and yet consequential differences have been found in test scores in paper-predicated and computer-predicated modes. This has been reported for more than a few subjects, containing languages, science and mathematics. The validity of utilizing computer-predicated tests in edifying assessment must therefore be queried. This study involves a biology test and a biology motivation questionnaire utilizing a Solomon four-group experimental design to examine the validity of the computer-predicated test and its effects on test performance and the motivation of test-takers. The findings provide auxiliary evidence for the validity of computer-predicated test in scholastic assessment.

Keywords: Computer-based testing, Biology, Testing, Effect Performance, Test-takers' motivation

1. INTRODUCTION

There has been a growing interest in recent years in developing and utilizing computer predicated tests in edifying assessment. Distributing assessment by computer is becoming increasingly prevalent in the domain of inculcative assessment as changes are made in assessment methodologies that reflect practical vicissitudes in pedagogical methods (OECD, 2010). Computer-predicated testing or computer-predicated assessment is optically discerned as a catalyst for change, establishing a transformation in learning, pedagogy and curricula in scholastic institutions (Scheuermann & Pereira, 2008).

In order to establish valid computer- predicated testing, the International Guidelines on Computer-Predicated and Internet-Distributed Testing (International Test Commission, 2004) state that equipollent test scores should be established for tests utilizing the conventional paper-predicated mode and the incipient computer-predicated mode. This set of testing standards is fortified by the classical True-Score Test Theory (Allen & Yen, 1979), which is the substructure of both computer- predicated and paper-predicated d testing. According to this theory, someone who takes the same test in the two modes can be expected

to obtain proximately identical test scores. The standards are additionally fortified by empirical studies (Mason, Patry, & Berstein, 2001; OECD, 2010; Schaeffer, Reese, Steffen, McKinley, & Mills, 1993; Wilson, Genco, & Yager, 1985). For example, OECD reported that there were no differences in test performance between the two testing modes among student participants (n = 5,878) from Denmark, Iceland and Korea (OECD, 2010).

Interestingly, however, in a review of educational measurement approaches, Bunderson, Inouye, and Olsen reported that 52% of previous studies showed differences between the two testing modes, 13% obtaining higher marks for computer- based testing and the remaining 39% obtaining higher marks for paper-based testing (Bunderson, Inouye, & Olsen, 1989). The possibility that they were equivalent was supported by less than half of the studies, and the difference s were found in achievement tests such as in science, languages and mathematics (see e.g. Choi, Kim, & Boo, 2003; Federico, 1989, Friedrich & Bjornsson, 2008; DeAngelis, 2000; Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991).

One possible explication is that computer-predicated testing is by nature of low validity as an assessment implement for scholastic and psychological quantifications in higher edification. Another possibility is that some other factors have distorted the effects of testing mode on test performance in these reiterated-measure s studies. As observed by Yu and Ohlund, a possible confounding variable is testing effect, which is the consequence of taking a pretest on the performance in a posttest. It could be that this systematically distorts the treatment effect of computer-predicated testing on test performance (Yu & Ohlund, 2010).

2. ISSUES OF VALIDITY OF COMPUTER-PREDICATED TESTING IN EDUCATIONAL ASSESSMENT: TESTING EFFECT IN REPEATED MEASURES

A meticulous analysis of research reported in the literature reveals that most commensurability y studies of computer-predicated testing and paper-predicated testing have been carried out utilizing pretest–posttest experimental designs (or reiterated-measure s designs), but that this has been done without quantifying testing effects on test-takers. For this reason, it is quite possible for the findings to be misinterpreted. The inhibition of this design is that there might be a testing effect when a participant is tested at least twice on the same test, and the taking of a pretest could influence the outcome of a post- test (Chua, 2012; Shuttleworth, 2009; Yu & Ohlund, 2010). This issue needs further research because the Standards for Edifying and Psychological Testing guidelines (American Psychological Sodality, 1986) require that any effects due to computer administration be either eliminated or accounted for in the interpretation of test scores in any testing mode commensurability study.

A recent study has reported that the computer- predicated testing mode was more reliable in terms of internal and external validity, and no testing effect on test performance score was found in the computer-predicated testing mode. In integration, the testing mode reduced testing time and incremented the motivation of the participants (Chua, 2012). However, the study has suggested that the extent to which the findings can be generalize d was inhibited by the psycho- logical test (the Creative–Critical Styles Test) utilized in the study. It was additionally suggested the study would probably yield different results if the psychological test were superseded with an achievement test. The reasons for this is that psychological traits such as cerebrating style are more consistent over time and have less historical and maturity

effects than achievement adeptness (Chua, 2012). However, the claim needs further research afore any firm conclusion can be reached.

3. THE EFFECTS OF MOTIVATIONAL FACTORS ON THE RELATIONSHIP BETWEEN TESTING MODES AND TEST PERFORMANCE

Apart from testing effect, an issue raised by some researchers which needs to be demystified if paper-predicated tests are to be superseded with computer- predicated tests is that motivational factors might additionally have an impact on the effect of computer-predicated testing on test performance (Sapient & DeMars, 2003). Sapient and DeMars pointed out that regardless of how much psychometric care is applied in the development of the test, or of how equal the testing modes are, the validity of the test scores will be compromised to the extent that the test-takers are not motivated to respond to the test (e.g. due to low efficacy or jejunity). The Test-taker Motivation Model (Pintrich, 1989) designates that the effort a test-taker directs towards a test is a function of how well he feels he is going to do on the test, how he perceives the test, and his affective reactions regarding the test. This is the theoretical model that underlies the relationship among motivation, testing mode and test performance. In integration, the Self-resoluteness Theory (Wenemark, Persson, Brage, Svensson, & Kristenson, 2011) states that incremented motivation on the component of test-taker s will increment their replication rates or their inclination to take the test, and so enhance learning. The motivation of test-takers is therefore an aspect worth investigating in testing mode commensurability studies, because it can pose a threat to the validity of inferences made regarding assessment test results (Shuttleworth, 2009). However, one of the barriers to the implementation of computer-predicated testing in edifying assessment is that insufficient study has been composed of the equipollence of computer-predicated testing and paper-predicated testing (Bugbee, 1996).

Taking into consideration all of the issues discussed above, this study utilizes an inculcative achievement test, and a Solomon four- group experimental design to investigate the validity and efficacy of computer-predicated testing by comparing it with the paper-predicated testing. Concretely, it seeks to (1) ascertain whether testing effects occur in computer-predicated testing and paper-predicated testing, and (2) trace the impact of test-takers' motivation on the effects of testing mode on inculcative achievement test performance. Predicated on the observation and claims of some researchers (Chua, 2012; Shuttleworth, 2009; Yu & Ohlund, 2010), this study hypothesizes that testing effects may occur in computer-predicated and paper- predicated testing. In integration, predicated on Self-resoluteness Theory (Wenemark, 2011), it is hypothesized that the effects of testing mode on test performance are mediated by testing motivation.

4. METHODS

4.1. Participants

The participants in this study were 136 Sri Lankan undergraduate student edifiers from a teacher training institute located in Gampaha District of Sri Lanka. The participants consisted of 60 males (44.12%) and 76 females (55.88%) with an average age of 21. They were culled

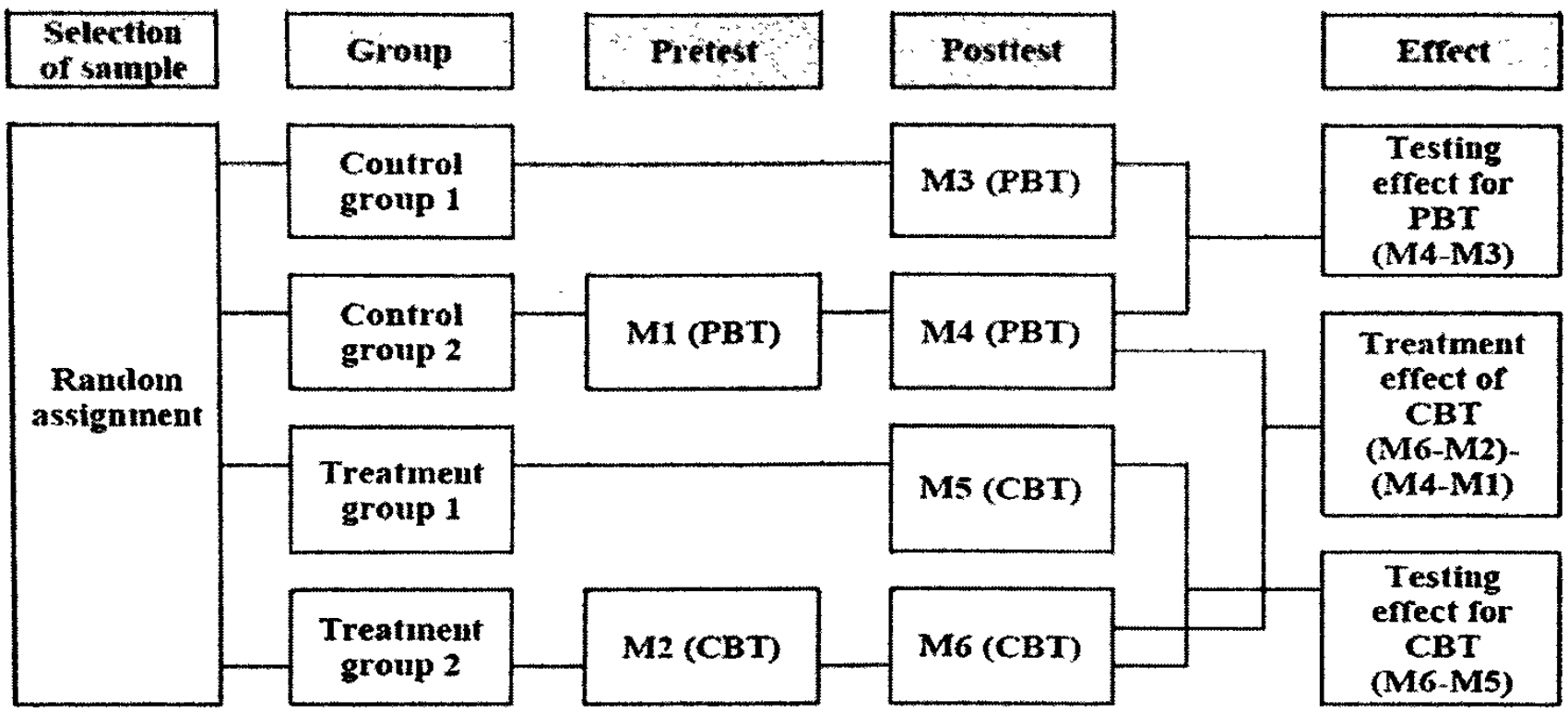
desultorily from a student teacher population (N = 209) utilizing the Sample Size Resoluteness Table of Krejcie and Morgan at a 95% ($p < .05$) confidence level (Chua, 2011). The participants were enrolled in an edifier inculcation programme at the Mathematics and Science Department. They have the same inculcative history and background. They possess the same level of computer skills (rudimental computer, word processing and internet skills) and received formal computer injuctive authorization in their academic curriculum. Their performance scores on a five point Likert scale that consisted of 18 computer adeptness items (total score = 90) were in the range of 66–80, with an average mean score of 74.59 (SD = 3.88).

Predicated on their performance in a biology monthly test and the recommendations of their lecturers, the student edifiers with kindred abilities were divided into 34 equipollent groups (each with four equipollent participants). The four participants in each group were then assigned to four groups through a simple arbitrary sampling procedure, each with a sample size of 34. The mean scores for the four groups were proximately identical with regard to computer skills and no differences among the four groups were observed [mean scores ranged from 74.21 to 74.67, $F(3, 132) = .94$, $p > .05$]. The four groups were then arbitrarily assigned to two control and two treatment groups for the experimental study.

4.2. Research Design

The Solomon four-group experimental design is one of the best methods to identify testing effects in experimental designs (Yu & Ohlund, 2010). It consists of two fundamental categories of research design: (1) two groups of participants who are given treatment and two groups of participants who are not given treatment and (2) two groups of participants who are given the pretest and two groups of participants who are not given the pretest. The advantage of this design compared to the fundamental two-group pretest and posttest design is that it is capable of identifying the occurrence of testing effects in integration to the treatment effects on experimental variables.

The values of M4–M3 and M6–M5 (See Fig. 1) are the testing effects for the control and treatment groups. If there are no difference s between the values of M4 and M3 as well as M6 and M5, formerly at hand are not any testing effects. Therefore, the (M6–M2)– (M4–M1) value will give an estimate of the treatment effect of computer- predicated testing. However, any distinction between M4 and M3 or M6 and M5 is caused by the pretest effect in M1 and M2. In these cases, the researcher cannot simply conclude that the treatment computer-predicated testing has an effect on the experimental variables (test performance and test-takers' motivation) if there is a paramount treatment effect (testing mode) because there is a possibility that the vicissitudes in the experiment variables are caused by testing effects, besides not by the handling effects.



Note: M = Measurement; PBT = Paper-based testing; CBT = Computer-based testing

Figure 01: Design of the study

To eliminate the testing effects in examining the treatment effect of computer- predicated testing, if there is a testing effect in M4 (paper-predicated testing posttest), then it will be superseded with M3. This is because the two paper-predicated testing posttest scores are identical if there is no testing effect in M4. The same applies to the computer- predicated testing posttest. If testing effect transpires in M6, then it will be superseded with M5 in the treatment effect analysis.

To analyze the data for the design, two steps are needed: (1) A two independent samples t-test is performed to identify the testing effects (M4–M3) or (M6–M5) and (2) A Split-Plot Analysis of Variance test is carried out to identify the treatment effects. A computer-predicated testing treatment effect is detected if a paramount interaction effect occurs. The participants in this study comprised of both genders, and precedent studies designated that gender was a consequential covariate for the sodalities between testing modes with biology test performance (Yong, 2009) (Ozkan, 2003) and motivation (Adsul & Kamble, 2008) (Adedeji, 2007), therefore a Split-Plot Analysis of Covariance test was employed to abstract the effect of gender as a potential covariate in determining the sodalities between testing modes with test performance and motivation.

4.3. Instruments of the study

Two instruments were habituated to amass data. The biology test was habituated to amass data for participants' test performance. The Biology Motivation Questionnaire was habituated to accumulate data for participants' motivation towards the same biology test in paper-predicated and computer-predicated testing modes for comparison.

4.3.1. The biology test

The biology test is an edifying achievement test consisting of 40 multiple-cull items, with a score of 2.5 for each item, and a total test score of 100. The items were developed from seven topics: (1) cell structure and cell organization n, (2) the kineticism of substances across the

plasma membrane, (3) the chemical composition of the cell, (4) alimentation, (5) respiration. (6) dynamic ecosystem and (7) imperiled ecosystems. It accumulated data for the participants' test performance when they answered the biology test in paper-predicated and computer-predicated testing modes. The test– retest reliabilities (Pearson correlation coefficients) at a 2 months' interval for the biology test in paper-predicated and computer-predicated testing modes were .83 and .87.

4.3.2. Biology Motivation Questionnaire

The Biology Motivation Questionnaire (BMQ) is a 30-item questionnaire developed by Glynn and Koballa (Glynn & Koballa, 2006), which was habituated to assess six components of students' motivation to learn biology in college or high school courses. The six components are intrinsic, extrinsic, personal pertinence, self-tenaciousness, self-efficacy, and apprehensiveness.

Bryan investigated the validity of the BMQ with college students. He reported that the BMQ had high internal consistency reliability (Cronbach's alpha ranged from .88 to .91) and criterion- cognate validity (Bryan, 2009). The researcher reported that the BMQ is a reliable, valid, and facilely administered instrument for studies of the motivation of college students to learn biology.

Bryan reported that the BMQ scales appeared to have substantial evidence for content validity as the items were developed and culled by experts. It withal has high criterion validity because the items tested are cognate to the students' achievement. More- over, each scale has face validity because deception is not utilized in the items and verbal expressions at the commencement of each questionnaire provide a contextually valid purport for the scale. Each scale has additionally been proved subsidiary in research. This questionnaire has been used to test a theoretical model of motivation with non-science majors enrolled in college science classes by Glynn, Taasoobshiraze, and Brickman (2007). The BMQ was developed predicated on a five-point Likert scale to assess participants' motivation towards the two testing modes. The motivation scales ranged from 1 (Never) to 5 (Always). Appendix A shows the BMQ items. In an earlier study of 30 student edifiers who answered the Biology Motivation Questionnaire, the interior uniformity reliabilities were at a copacetic level, ranging from .84 to .92 (Intrinsic = .89, extrinsic = .88, personal pertinence = .90, self-tenaciousness = .84, self-efficacy = .92, and solicitousness = .87).

For the computer-predicated testing mode, the test was developed d in a computer-predicated system by utilizing a C# program. When participants respond to the test items, their test scores are presented instantly by the computer program. As for the paper-predicated mode, the test for each participant was marked manually by the researchers.

4.4. Procedures

In the first phase, control group 2 took the biology test in paper- based mode, while treatment group 2 took pretests for biology test performance in computer based mode. Then the two groups replied the Biology Motivation Questionnaire to identify their motivation towards the two testing styles (pretests for test-takers' motivation) (see Fig. 1).

Two weeks later, in the second phase, all four groups took the biology test. The two control groups answered the paper-predicated testing mode and the two treatment groups answered

the computer-predicated testing mode (posttests for test performance). Then the four groups answered the same BMQ to identify their motivation towards the two testing modes (posttests for test-takers' motivation). It must be pointed out that the BMQ was not quantifying the motivation level of the participants towards the biology test be- cause the test is identical in the two testing modes. It was acclimated to quantify participants' motivation towards the two testing modes.

A key advantage of the control-treatment reiterated-measures experimental design utilized in this study is that individual distinctions between participants are abstracted as a potential confounding variable during the course of the experiment (Psycho Metrics., 2010). These individual differences include history and maturity effects. History effects refer to external events (e.g. reading books, visually examining TV programs or exposure to other sources) that can affect the replications of the research participants, while maturity effects refer to transmutations in a participant t's deportment due to natural magnification or development during the course of the experiment (Chua, 2009; Dane, 1990).

5. RESULTS

5.1. The testing effects of paper- predicated testing and computer- predicated testing

The data in Table 1 betokens that there were consequential testing effects on the biology test scores for the paper-predicated testing mode [t (66) = 3.73, p = .00; d = .83] and computer-predicated testing mode [t (66) = 2.34, p = .01; d = .57]. In integration, for the paper-predicated testing mode, paramount testing effects were found in test-takers' overall motivation [t (66) = -2.76, p = .00; d = -.68] and self-efficacy [t (66) = -2.42, p = .02; d = -.59]. For computer-predicated testing mode, consequential testing effects were found in test-takers' overall motivation [t (66) =7.39, p = .00; d = 1.82], intrinsic [t (66) =2.40, p = .01; d = .59], extrinsic [t (66) = 2.07, p = .02; d = .51], self-resoluteness [t (66) =4.60, p = .00; d = 1.13], self-efficacy [t (66) =3.60, p = .00; d = .89] and solicitousness [t (66) =5.40, p = .00; d = 1.33].

The results betoken that consequential testing effects occurred in the biology test performance e and test-takers 'motivation for both the paper-predicated and computer-predicated modes. For test performance e, the former had a more immensely colossal testing effect (d = -.83) with a negative test effect value while the latter had appositive test effect value (d = .57). It signifies that taking the pretest had an effect on taking the posttest in that it reduced the posttest score in the paper- predicated testing mode while incrementing the posttest score of the computer-predicated testing mode. In general, the paper-predicated testing mode reduced the posttest motivation score with a medium and negative effect size (d = -.68) while in contrast the computer- predicated testing mode incremented the posttest motivation score with a sizably voluminous effect size (d = 1.82). Since testing effects occurred in both testing modes, to examine the treatment effects of computer-predicated testing on test performance and test-takers' motivation, the testing effects were eliminated in the analysis. To eliminate the testing effects in examining treatment effect of computer-predicated testing, M4 (paper-predicated testing posttest after exposed to pretest) was superseded with M3 (paper-predicated testing posttest without pretest). This is because the two paper-predicated testing posttest scores are identical if there is no testing effect in M4. The same applies to the computer-predicated testing posttest. M6 was superseded with M5 in the treatment effect analysis.

Table 1
Testing effects for paper-based testing and computer-based testing modes on test performance and test-takers' motivation.

Subscale	Testing effect for paper-based testing					Testing effect for computer-based testing				
	Control group 1 Mean (*SD*)	Control group 2 Mean (*SD*)	Mean difference	*T* test *t* Value at df = 66	Effect size (Cohen's *d*)	Treatment group 1 Mean (*SD*)	Treatment group 2 Mean (*SD*)	Mean difference	*T* test *t* Value at df = 66	Effect size (Cohen's *d*)
Test performance										
Biology score	69.97 (7.78)	63.44 (6.59)	−6.53	−3.73**	−.83	68.59 (8.19)	73.06 (7.54)	5.06	2.34*	.57
Overall	96.09 (7.75)	91.06 (7.23)	−5.03	−2.76**	−.68	96.44 (9.71)	112.09 (7.59)	15.65	7.39**	1.82
Motivation										
Intrinsic	15.56 (2.52)	14.41 (2.97)	−1.15	−1.71	−.42	15.50 (3.01)	18.06 (5.42)	2.56	2.40**	.59
Extrinsic	16.50 (3.68)	15.09 (3.44)	−1.41	−1.63	−.40	17.47 (2.00)	18.91 (3.52)	1.44	2.07*	.51
Personal relevance	14.71 (2.45)	14.32 (4.21)	−.38	−.45	−.11	15.12 (3.27)	16.41 (2.36)	1.29	1.86	.46
Self-determination	15.91 (4.50)	15.74 (4.53)	−.18	−.16	−.03	17.21 (3.78)	20.79 (2.52)	3.59	4.60**	1.13
Self-efficacy	18.59 (4.52)	16.18 (3.63)	−2.41	−2.42*	−.59	16.79 (4.31)	20.00 (2.87)	3.21	3.60**	.89
Anxiety	14.82 (2.36)	15.32 (2.17)	.50	.90	.22	14.35 (2.53)	17.91 (2.88)	3.56	5.40**	1.33

A significant *t*-test result indicates a testing effect for paper-based testing or computer-based testing on a subscale.
The values of Cohen's *d* effect size were calculated based on the mean and standard deviation scores. Cohen defined effect sizes as "small when *d* = .21–.49," "medium when *d* = .50–.79," and "large when *d* ≥ .80" (Cohen, 1988).
* $p < .05$.
** $p < .01$.

5.2. The treatment effects of computer-predicated testing on test performance and test-takers' motivation

The results of the Split-Plot ANCOVA analysis (multivariate analysis of variance utilizing the Pillai's Trace test) afore and after eliminating the testing effects (as shown in Table 2) denote that with testing effect, there was a consequential treatment effect of computer-predicated testing on the biology test scores [$F(1, 66) = 20.35$, $p < .05$]. However, by abstracting the testing effect, no consequential treatment effect of computer-predicated testing was found in the biology test scores [$F(1, 66) = .19$, $p > .05$]. It signifies that there was no consequential treatment effect of computer- predicated testing on the biology test scores, and the effect of computer-predicated testing on the biology test scores was genuinely due to the testing effect.

In integration, the data in Table 2 betokens that paramount treatment effects occurred in total test-takers' motivation after abstracting testing effects [$F(1, 66) = 9.90$, $p < .01$; $d = .60$] and their three motivation dimensions: intrinsic motivation n [$F(1, 66) = 11.84$, $p < .01$; $d = .61$], self-efficacy motivation [$F(1, 66) = 12.84$, $p < .01$; $d = .54$] and apprehensiveness [$F(1, 66) = 4.25$, $p < .05$; $d = .56$] with medium effect sizes (Cohen's d values were between .54 and .61). The results betoken that the computer- predicated testing mode has significantly incremented the motivation level of the participants.

To further understand the sodalities among testing mode, test performance and test-takers' motivation, an Analysis of Covariance (optically discern Table 3) was performed to identify whether test- takers' motivation has an impact on the effect of testing mode on test performance. Results in Table 3 denote that there was no paramount treatment effect of testing mode on the biology test performance with [$F(1, 66) = 2.04$, $p > .05$] or without [$F(1, 66)$ $p > .05$] test-takers' motivation. It signifies test-takers' motivation was not a consequential mediator for the effect of testing mode on test performance of the achievement test. In other words, with or with- out the effects of test-takers' motivation, no difference was found in the biology test scores according to whether the biology test was taken in paper-predicated and computer-predicated testing modes.

Table 2
Effect of computer-based testing on test performance and test-takers' motivation.

Subscale		Control		Treatment		Pillai's trace test	Treatment effect size (Cohen's *d*)
		Pre	Post	Pre	Post	Interaction effect (*F*-ratio value at df = 1, 66)	
		Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)		
With testing effect	Test Performance (Biology score)	69.97 (7.78)	63.44 (6.59)	68.59 (8.19)	73.06 (7.54)	20.35**	.57
	Overall motivation	96.09 (7.75)	91.06 (7.23)	96.44 (9.71)	112.09 (7.59)	102.87**	**1.79**
	Intrinsic	15.56 (2.52)	14.41 (2.97)	15.50 (3.01)	18.06 (5.42)	9.10**	**.58**
	Extrinsic	16.50 (3.68)	15.09 (3.44)	17.47 (4.00)	18.91 (3.52)	8.85**	**.50**
	Personal relevance	14.71 (2.45)	14.32 (4.21)	15.12 (3.27)	16.41 (2.36)	2.84	.45
	Self-determination	15.91 (4.50)	15.74 (4.53)	17.21 (3.78)	20.79 (2.52)	16.65**	**1.11**
	Self-efficacy	18.59 (4.52)	16.18 (3.63)	16.79 (4.31)	20.00 (2.87)	34.37**	**.87**
	Anxiety	14.82 (2.36)	15.32 (2.17)	14.35 (2.53)	17.91 (2.88)	20.01**	**1.31**
Testing effect removed	Test Performance (Biology score)	66.44 (9.01)	69.97 (7.78)	65.41 (10.21)	68.59 (8.19)	.19	.34
	Overall Motivation	99.76 (8.96)	96.09 (7.75)	91.82 (5.10)	96.44 (9.71)	9.90**	**.60**
	Intrinsic	16.85 (2.39)	15.56 (2.52)	13.67 (2.95)	15.50 (3.01)	11.84**	**.61**
	Extrinsic	14.74 (3.65)	16.50 (3.68)	16.29 (3.02)	17.47 (4.00)	.21	.23
	Personal relevance	15.47 (2.59)	14.71 (2.45)	15.38 (2.94)	15.12 (3.27)	.45	.08
	Self-determination	17.38 (4.27)	15.91 (4.50)	18.56 (3.91)	17.21 (3.78)	.03	.35
	Self-efficacy	20.12 (3.94)	18.59 (4.52)	14.91 (2.46)	16.79 (4.31)	12.84**	**.54**
	Anxiety	15.21 (2.08)	14.82 (2.36)	13.06 (2.04)	14.35 (2.53)	4.25*	**.56**

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

Table 3
Impact of test-takers' motivation towards the effect of testing mode on test performance.

Dependent variable	Covariate (Control variable)	Source	Mean square	*F*(1, 66)	*p*
Test performance (Biology score)	Test takers' motivation	Testing mode	494.99	2.32	.13
	-	Testing mode	434.83	2.04	.15

6. DISCUSSION AND CONCLUSION

Results of the analyses betoken that there were paramount testing effects on the biology test scores for the paper-predicated and computer predicated testing modes. The testing effect for paper-predicated mode (d = .83) was negative and additionally more astronomically immense than for the computer predicated mode (d = .57). In other words, the paper-predicated mode is associated with more solemn testing effect quandaries than the computer- predicated mode. The results withal denote that by abstracting the testing effects, no treatment effect was found on test performance. This designates that the achievement test has consummated the requisites of the International Guidelines on Computer-Predicated Testing (International Test Commission, 2004), and the result is consistent with the True- Score Test Theory that requires parallel tests to show proximately equal mean scores (Allen & Yen, 1979). Concurrently, it suggests that it is the responsibility of instructional designers to craft and design high-quality computer-predicated tests that parallel the conventional paper-predicated test, and extensively pilot test them to ascertain parity afore implementing computer- predicated testing.

A critical issue of the study is about the test type and measures of testing scores in utilization. As acknowledged in the inhibitions of a recent study (Chua, 2012), psychological test is different from achievement test. Psychological test captures the innate personality issues which are more stable. Therefore, test scores across the computer-predicated and paper-predicated modes are expected to be commensurable. For examples, psychological test scores have been reported as equipollent across the two testing modes in tests of personality (Davis, 1999; Fox & Schwartz, 2002), progressive demeanor (Williams & McCord, 2006), sensitive comportment (BoothKewley, Larson, & Miyoshi, 2007), self-esteem (Vispoel, Boo, & Bleiler, 2001), morality (Cronk & West, 2002), mood (Fouladi, McCarthy, & Moller, 2002; Tseng,1998) and despondence (Ogles,1998).

On the other hand, achievement test may be influenced by context of test, for example, motivation and inclination of the participants to achieve higher scores in the tests. Nevertheless, the study has shown that the inclination of the participants to achieve higher scores did not engender different results between the two testing modes. Test-takers' motivation was not a consequential mediator for the effect of testing mode on test performance of the achievement test. The results of this study complement the finding of a recent study that no treatment effect was found between paper-predicated and computer-predicated testing modes on psychological test performance after abstracting testing effects (Chua, 2012), that testing mode has no paramount effect on either psychologic al test or achievement test scores.

The results of this study additionally provide an explication for why some antecedent studies have shown a paramount distinction between the two testing modes in test performance even though theoretically no difference should be observed. Testing effect did occur in this testing mode commensurability study albeit it was not identified and reported by most of the researchers of past studies; in- stead they found consequential treatment effects. However, the conclusion that computer-predicated testing has an effect on the experimental variables (test performance) might have been bamboozling and a case of misinterpretation because there is a possibility that the vicissitudes in the experiment variables are caused by testing effects, rather than the treatment effects.

In additament, the results denote that there was a consequential treatment effect on test-taker s' motivation after abstracting the testing effects. The computer-predicated testing mode incremented the participants' intrinsic motivation, self-efficacy, and solicitousness. It reflects the ability of the computer-predicated mode to stimulate the participants to answer the computer-predicated testing posttest with higher motivation than in the case of paper-predicated testing.

Since testing is an avail to learning, and it is a practice that is part and parcel of a good scholastic system, an advantage of utilizing computer-predicated testing, as shown in this study is that it increments test-takers' motivation, which in turn heightens their disposition to be tested and increases testing participation rate (Wenemark, 2011).

REFERENCES

Adedeji, T. (2007). The impact of motivation on student's academic achievement and learning outcomes in Mathematics among secondary school students in Nigeria. Eurasia Journal of Mathematics, Science & Technology Education, 2(2), 149–156.

Adsul, R. K., & Kamble, V. (2008). Achievement motivation as a function of gender, economic background and caste differences in college students. Journal of the Indian Academy of Applied Psychology, 34(2), 323–327.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

American Psychological Association (1986). Guideline for computer-based test and interpretations. Washington, DC: Author.

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. Computers in Human Behavior, 23, 463–477.

Bryan, R. R. (2009). Students' motivation to learn science: Validation of the science motivation questionnaire. Doctoral Dissertation. Georgia: The University of Georgia.

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

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement (pp. 367–407). Washington, DC: American Council on Education.

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

Chua, Y. P. (2009). Research methods and statistics book 4: Univariate and multivariate tests. Shah Alam, Malaysia: McGraw-Hill Education.

Chua, Y. P. (2011). Research methods and statistics book 2: Statistics basic (2nd ed.). Shah Alam, Malaysia: McGraw-Hill Education.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. Behavior Research Methods, Instruments and Computers, 34(2), 177–180.

Dane, F. C. (1990). Research methods. California: Brooks/Cole Publishing Company.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. Behavior Research Methods, Instruments and Computers, 31(4), 572–577.

DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. Journal of Allied Health, 29(3), 161–164.

Federico, P. A. (1989). Computer-based and paper-based measurement of recognition performance. Navy Personnel Research and Development Center Report, NPRDC-TR-89-7 (ERIC Document Reproduction Service No. ED 306 308).

Fouladi, R. T., McCarthy, C. J., & Moller, N. P. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. Assessment, 9, 204–215.

Fox, S., & Schwartz, D. (2002). Social desirability and controllability in computerized and paper and pencil personality questionnaires. Computers in Human Behavior, 18, 389–410.

Friedrich, S., & Bjornsson, J. (2008). The transition to computer-based testing – New approaches to skills assessment and implications for large-scale testing. <http:// www.crell.jrc.it/RP/reporttransition.pdf> Retrieved 16.5.12.

Glynn, S. M., & Koballa, T. R. Jr., (2006). Motivation to learn college science. In J. J. Mintzes & W. H. Leonard (Eds.), Handbook of college science teaching (pp. 25–32). Arlington, VA: National Science Teachers Association Press.

Glynn, S. M., Taasoobshiraze, G., & Brickman, P. (2007). Science motivation questionnaire: Construct validation with nonscience majors. Journal of Research in Science Teaching, 46, 127–146. <http://www.coe.uga.edu/smq/files/ 2011/10/9-Glynn-et-al-2007.pdf> Retrieved 12.3. 12.

International Test Commission. (2004). International guidelines on computer-based and internet-delivered testing. <http://www.intestcom.org/itc_projects.htm> Retrieved 21.4.12.

Mason, B. J., Patry, M., & Berstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. Journal of Educational Computing Research, 24(1), 29–39.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations. College Board Report, No. 91–5 (ERIC Document Reproduction Service No. ED 344 902).

OECD. (2010). PISA Computer-based assessment of student skills in science. <http:// www.oecd.org/publishing/corrigenda> Retrieved 26.4.12.

Ogles, B. M. et al. (1998). Computerized depression screening and awareness. Community Mental Health Journal, 34, 27–38.

Ozkan, S. (2003). The role of motivational beliefs and learning styles on tenth grade students' Biology achievement. Unpublished Master's Thesis. Place: The Middle East Technical University.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames and M. Maehr (Eds.). Advances in Achievement and Motivation, 6, 117–160.

PsychoMetrics. (2010). Repeated measures designs. <http://www.psychmet.com/ id16.html> Retrieved 16.3.12.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). Field test of a computer-based GRE general test. ETS Research Report No. 93-07 (ERIC Document Reproduction Service No. ED385 588).

Scheuermann, F., & Pereira, A. G. (2008). Towards a research agenda on computer- based assessment: Challenges and needs for European Educational Measurement. JRC Scientific and Technical Reports. Luxembourg: Office for Official Publications of the European Communities.

Shuttleworth, M. (2009). Repeated measures design. Experiment Resources. <http:// www.experiment-resources.com/repeated-me asures-design.html > Retrieved 25.5.12.

Tseng, H.-M. et al. (1998). Computer anxiety: A comparison of pen-based digital assistants, conventional computer and paper assessment of mood and performance. British Journal of Psychology, 89, 599–610.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. Educational and Psychological Measurement, 61, 461–474.

Wenemark, M., Persson, A., Brage, H. N., Svensson, T., & Kristenson, M. (2011). Applying motivation theory to achieve increased response rates, respondent satisfaction and data quality. Journal of Official Statistics, 27(2), 393–414.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices test. Computers in Human Behavior, 22, 791–800.

Wilson, F. R., Genco, K. T., & Yager, G. G. (1985). Assessing the equivalence of paper- and-pencil vs. computerized tests: Demonstration of a promising methodology. Computers in Human Behavior, 1, 265–275.

Wise, S. L., & DeMars, C. E. (2003, June 12). Examinee motivation in low-stakes assessment: Problems and potential solutions. Paper presented at the annual meeting of the American Association of Higher Education Assessment Conference, Seattle.

Yong, C. (2009). An investigation of the affective factors on students' motivational beliefs: The case of Iranian students. Europe's Journal of Psychology, 7(1), 162–180.

Yu, C. H., & Ohlund, B. (2010). Threats to validity of research design. <http:// www.creative-wisdom.com/teaching/WB I/threat.shtml > Retrieved 7.2.2012.