Effects of computer-based educational achievement test on test performance and test takers' motivation

Yan Piaw Chua a,*, Zuraidah Mohd Don b,c

a Institute of Educational Leadership, University of Malaya, Level 2, Block C, City Campus Complex UIM, Jalan Tun Ismail, 50480 Kuala Lumpur, Malaysia
b Faculty of Languages & Linguistics Building, University of Malaya, 50603 Kuala Lumpur, Malaysia
c Asia–Europe Institute, University of Malaya, 50603 Kuala Lumpur, Malaysia

A R T I C L E I N F O

Keywords:
Computer-based testing
Biography
Testing effect
Performance
Test-takers' motivation

A B S T R A C T

There has been an increasing interest in recent years in developing and using computer-based tests in educational assessment. To replace paper-based tests with computer-based ones, the standards for developing computerized-assessment (International Test Commission, 2004) requires equivalent test scores to be established for the new computer-based test and the conventional paper-based test. However, in most test mode comparability studies, the actual test items used have been identical, and yet significant differences have been found in test scores in paper-based and computer-based modes. This has been reported for several subjects, including science, languages and mathematics. The validity of using computer-based tests in educational assessment must therefore be questioned. This study involves a biology test and a biology motivation questionnaire using a Solomon four-group experimental design to examine the validity of the computer-based test and its effects on test performance and the motivation of test-takers. The findings provide supportive evidence for the validity of computer-based test in educational assessment.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

There has been a growing interest in recent years in developing and using computer-based tests in educational assessment. Delivering assessment by computer is becoming increasingly prevalent in the domain of educational assessment as changes are made in assessment methodologies that reflect practical changes in pedagogical methods (OECD, 2010). Computer-based testing or computer-based assessment is seen as a catalyst for change, bringing about a transformation in learning, pedagogy and curricula in educational institutions (Scheuermann & Pereira, 2008).

In order to establish valid computer-based testing, the International Guidelines on Computer-Based and Internet-Delivered Testing (International Test Commission, 2004) state that equivalent test scores should be established for tests using the conventional paper-based mode and the new computer-based mode. This set of testing standards is supported by the classical True-Score Test Theory (Allen & Yen, 1979), which is the basis of both computer-based and paper-based testing. According to this theory, someone who takes the same test in the two modes can be expected to obtain nearly identical test scores. The standards are also supported 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 differences were found in achievement tests such as in science, languages and mathematics (see e.g. Choi, Kim, & Boo, 2003; DeAngelis, 2000; Federico, 1989; Friedrich & Bjornsson, 2008; Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991).

One possible explanation is that computer-based testing is by nature of low validity as an assessment tool for educational and psychological measurements in higher education. Another possibility is that some other factors have distorted the effects of testing mode on test performance in these repeated-measures studies. As observed by Yu and Ohlund, a possible confounding variable is testing effect, which is the effect of taking a pretest on the performance in a posttest. It could be that this systematically distorts the treatment effect of computer-based testing on test performance (Yu & Ohlund, 2010).
2. Issues of validity of computer-based testing in educational assessment: Testing effect in repeated-measures

A careful analysis of research reported in the literature reveals that most comparability studies of computer-based testing and paper-based testing have been carried out using pretest-posttest experimental designs or (repeated-measures designs), but that this has been done without measuring testing effects on test-takers. For this reason, it is quite possible for the findings to be misinterpreted. The limitation 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 Educational and Psychological Testing guidelines (American Psychological Association, 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 comparability study.

A recent study has reported that the computer-based 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-based testing mode. In addition, the testing mode reduced testing time and increased the motivation of the participants (Chua, 2012). However, the study has suggested that the extent to which the findings can be generalized was limited by the psychological test (the Creative–Critical Styles Test) used in the study. It was also suggested the study would probably yield different results if the psychological test were replaced with an achievement test. The reasons for this is that psychological traits such as thinking style are more consistent over time and have less historical and maturity effects than achievement skill (Chua, 2012). However, the claim needs further research before 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 clarified if paper-based tests are to be replaced with computer-based tests is that motivational factors might also have an impact on the effect of computer-based testing on test performance (Wise & DeMars, 2003). Wise 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 boredom). The Test-taker Motivation Model (Pintrich, 1989) specifies 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 addition, the Self-determination Theory (Wenemark, Persson, Brage, Svensson, & Kristenson, 2011) states that increased motivation on the part of test-takers will increase their response rates or their willingness to take the test, and so enhance learning. The motivation of test-takers is therefore an aspect worth investigating in testing mode comparability 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-based testing in educational assessment is that insufficient study has been made of the equivalence of computer-based testing and paper-based testing (Bugbee, 1996).

Taking into consideration all of the issues discussed above, this study uses an educational achievement test, and a Solomon four-group experimental design to investigate the validity and effectiveness of computer-based testing by comparing it with paper-based testing. Specifically, it seeks to (1) find out whether testing effects occur in computer-based testing and paper-based testing, and (2) trace the impact of test-takers’ motivation on the effects of testing mode on educational achievement test performance. Based 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-based and paper-based testing. In addition, based on Self-determination Theory (Wenemark et al., 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 Malaysian undergraduate student teachers from a teacher training institute located in Peninsular Malaysia. The participants consisted of 60 males (44.12%) and 76 females (55.88%) with an average age of 21. They were selected randomly from a student teacher population (N = 209) using the Sample Size Determination Table of Krejcie and Morgan at a 95% (p < .05) confidence level (Chua, 2011) (p. 211). The participants were enrolled in a teacher education programme at the Mathematics and Science Department. They have the same educational history and background. They possess the same level of computer skills (basic computer, word processing and internet skills) and received formal computer instruction in their academic curriculum. Their performance scores on a five-point Likert scale that consisted of 18 computer skill items (total score = 90) were in the range of 66–80, with an average mean score of 74.59 (SD = 3.88).

Based on their performance in a biology monthly test and the recommendations of their lecturers, the student teachers with similar abilities were divided into 34 equivalent groups (each with four equivalent participants). The four participants in each group were then assigned to four groups through a simple random sampling procedure, each with a sample size of 34. The mean scores for the four groups were nearly 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 randomly 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 basic 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 basic two-group pretest and posttest design is that it is capable of identifying the occurrence of testing effects in addition 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 differences between the values of M4 and M3 as well as M6 and M5, then there are no testing effects. Therefore, the (M6–M2)–(M4–M1) value will give an estimate of the treatment effect of computer-based testing. However, any difference 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-based testing has an effect on the experimental variables (test performance and test-takers’ motivation) if there is a significant treatment effect (testing mode) because there is a possibility that the changes in the experiment variables are caused by testing effects, and not by the treatment effects.

To eliminate the testing effects in examining the treatment effect of computer-based testing, if there is a testing effect in M4 (paper-based testing posttest), then it will be replaced with M3. This is because the two paper-based testing posttest scores are identical if there is no testing effect in M4. The same applies to the computer-based testing posttest. If testing effect occurs in M6, then it will be replaced 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-based testing treatment effect is detected if a significant interaction effect occurs. The participants in this study consisted of males and females, and previous studies indicated that gender was a significant covariate for the associations between testing modes with biology test performance (Ozkan, 2003; Yong, 2009) and motivation (Adedeji, 2007; Adsul & Kamble, 2008), therefore a Split-Plot Analysis of Covariance test was employed to remove the effect of gender as a potential covariate in determining the associations between testing modes with test performance and motivation.

4.3. Instruments of the study

Two instruments were used to collect data. The biology test was used to collect data for participants’ test performance. The Biology Motivation Questionnaire was used to collect data for participants’ motivation towards the same biology test in paper-based and computer-based testing modes for comparison.

4.3.1. The biology test

The biology test is an educational achievement test consisting of 40 multiple-choice 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, (2) the movement of substances across the plasma membrane, (3) the chemical composition of the cell, (4) nutrition, (5) respiration, (6) dynamic ecosystem and (7) endangered ecosystems. It collected data for the participants’ test performance when they answered the biology test in paper-based and computer-based testing modes. The test-retest reliabilities (Pearson correlation coefficients) at a 2 months interval for the biology test in paper-based and computer-based 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 used to assess six components of students’ motivation to learn biology in college or high school courses. The six components are intrinsic, extrinsic, personal relevance, self-determination, self-efficacy, and anxiety.

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-related validity (Bryan, 2009). The researcher reported that the BMQ is a reliable, valid, and easily administered instrument for studies of the motivation of college students to learn biology.

Bryan reported that the BMQ scales appeared to have good evidence for content validity as the items were developed and selected by experts. It also has high criterion validity because the items tested are related to the students’ achievement. Moreover, each scale has face validity because deception is not used in the items and statements at the beginning of each questionnaire provide a contextually valid purpose for the scale. Each scale has also been proved useful 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 based 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 teachers who answered the BMQ, the internal consistency reliabilities were at a satisfactory level, ranging from .84 to .92 (Intrinsic = .89, extrinsic = .88, personal relevance = .90, self-determination = .84, self-efficacy = .92, and anxiety = .87).

For the computer-based testing mode, the test was developed in a computer-based system by using a Visual Basics program. When participants respond to the test items, their test scores are presented instantly by the computer program. As for the paper-based mode, the test for each participant was marked manually by the researcher.
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 answered the BMQ to identify their motivation towards the two testing modes (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-based testing mode and the two treatment groups answered the computer-based 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 measuring the motivation level of the participants towards the biology test because the test is identical in the two testing modes. It was used to measure participants’ motivation towards the two testing modes.

A key advantage of the control-treatment repeated-measures experimental design used in this study is that individual differences between participants are removed as a potential confounding variable during the course of the experiment (Psychometrika, 2010). These individual differences include history and maturity effects. History effects refer to external events (e.g., reading books, watching TV programmes or exposure to other sources) that can affect the responses of the research participants, while maturity effects refer to changes in a participant’s behavior due to natural growth or development during the course of the experiment (Chua, 2009; Dane, 1990).

5. Results

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

The data in Table 1 indicates that there were significant testing effects on the biology test scores for the paper-based testing mode [t(66) = -3.73, p = .00; d = -.83] and computer-based testing mode [t(66) = 2.34, p = .01; d = .57].

In addition, for the paper-based testing mode, significant 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-based testing mode, significant 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-determination [t(66) = 4.60, p = .00; d = 1.13], self-efficacy [t(66) = 3.60, p = .00; d = .89] and anxiety [t(66) = 5.40, p = .00; d = 1.33].

The results indicate that significant testing effects occurred in the biology test performance and test-takers’ motivation for both the paper-based and computer-based modes. For test performance, the former had a bigger testing effect (d = -.83) with a negative test effect value while the latter had a positive test effect value (d = .57). It means that taking the pretest had an effect on taking the posttest, in that it reduced the posttest score in the paper-based testing mode while increasing the posttest score of the computer-based testing mode. In general, the paper-based testing mode reduced the posttest motivation score with a medium and negative effect size (d = -.68) while in contrast the computer-based testing mode increased the posttest motivation score with a large effect size (d = 1.82).

Since testing effects occurred in both testing modes, to examine the treatment effects of computer-based testing on test performance and test-takers’ motivation, the testing effects were eliminated in the analysis. To eliminate the testing effects in

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Control group</th>
<th>Treatment group</th>
<th>Mean (SD)</th>
<th>Mean (SD)</th>
<th>Effect size (Cohen’s d)</th>
<th>Effect size (Cohen’s d)</th>
<th>T-test</th>
<th>Value at df = 66</th>
<th>Mean (SD)</th>
<th>Mean (SD)</th>
<th>t-test</th>
<th>Value at df = 66</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology score</td>
<td>69.97 (7.75)</td>
<td>63.44 (6.59)</td>
<td>-6.53</td>
<td>-3.73*</td>
<td>.59</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>91.06 (7.23)</td>
<td>81.06 (7.23)</td>
<td>-10.00</td>
<td>-2.76*</td>
<td>.39</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>14.41 (2.97)</td>
<td>12.34 (3.45)</td>
<td>-2.07</td>
<td>-.42*.</td>
<td>.42</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrinsic</td>
<td>15.56 (3.62)</td>
<td>14.41 (2.97)</td>
<td>-1.15</td>
<td>-1.71</td>
<td>.71</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extrinsic</td>
<td>15.50 (3.01)</td>
<td>14.41 (2.97)</td>
<td>-1.09</td>
<td>-.42*.</td>
<td>.42</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal</td>
<td>15.25 (3.26)</td>
<td>14.41 (2.97)</td>
<td>-1.04</td>
<td>-1.71</td>
<td>.71</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-regulation</td>
<td>14.71 (2.45)</td>
<td>14.41 (2.97)</td>
<td>-.30</td>
<td>-.10*.</td>
<td>.10</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-worth</td>
<td>15.71 (2.45)</td>
<td>14.41 (2.97)</td>
<td>.16</td>
<td>.16</td>
<td>.16</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>18.59 (4.52)</td>
<td>16.18 (3.63)</td>
<td>-.41</td>
<td>-2.42*</td>
<td>.59</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>15.89 (4.52)</td>
<td>13.34 (3.63)</td>
<td>-2.55</td>
<td>-2.42*</td>
<td>.59</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal</td>
<td>15.00 (4.36)</td>
<td>13.34 (3.63)</td>
<td>-1.66</td>
<td>-2.42*</td>
<td>.59</td>
<td>.80</td>
<td></td>
<td>T &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. The defined effect sizes as ‘‘small when d = .20–.49, ‘‘medium when d = .50–.79, and ‘‘large when d > .80’’ (Cohen, 1988).
examining treatment effect of computer-based testing, M4 (paper-based testing posttest after exposed to pretest) was replaced with M3 (paper-based testing posttest without pretest). This is because the two paper-based testing posttest scores are identical if there is no testing effect in M4. The same applies to the computer-based testing posttest. M6 was replaced with M5 in the treatment effect analysis (see Fig. 1).

5.2. The treatment effects of computer-based testing on test performance and test-takers’ motivation

The results of the Split-Plot ANCOVA analysis (multivariate analysis of variance using the Pillai’s Trace test) before and after eliminating the testing effects (as shown in Table 2) indicate that with testing effect, there was a significant treatment effect of computer-based testing on the biology test scores \([F(1, 66) = 20.35, p < .05]\). However, by removing the testing effect, no significant treatment effect of computer-based testing was found in the biology test scores \([F(1, 66) = .19, p > .05]\). It means that there was no significant treatment effect of computer-based testing on the biology test scores, and the effect of computer-based testing on the biology test scores was actually due to the testing effect.

In addition, the data in Table 2 indicates that significant treatment effects occurred in total test-takers’ motivation after removing testing effects \([F(1, 66) = 9.90, p < .01; d = .60]\) and their three motivation dimensions: intrinsic motivation \([F(1, 66) = 11.84, p < .01; d = .61]\), self-efficacy motivation \([F(1, 66) = 12.84, p < .01; d = .54]\) and anxiety \([F(1, 66) = 4.25, p < .05; d = .56]\) with medium effect sizes (Cohen’s \(d\) values were between .54 and .61). The results indicate that the computer-based testing mode has significantly increased the motivation level of the participants.

To further understand the associations among testing mode, test performance and test-takers’ motivation, an Analysis of Covariance (see 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 indicate that there were no significant treatment effect of testing mode on the biology test performance with \([F(1, 66) = 2.04, p > .05]\) or without \([F(1, 66) = 2.32, p > .05]\) test-takers’ motivation. It means test-takers’ motivation was not a significant mediator for the effect of testing mode on test performance of the achievement test. In other words, with or without 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-based and computer-based testing modes.

6. Discussion

Results of the analyses indicate that there were significant testing effects on the biology test scores for the paper-based and computer-based testing modes. The testing effect for paper-based mode \((d = -.83)\) was negative and also larger than for the computer-based mode \((d = .57)\). In other words, the paper-based mode is associated with more serious testing effect problems than the computer-based mode.

The results also indicate that by removing the testing effects, no treatment effect was found on test performance. This means that the achievement test has fulfilled the requirements of the International Guidelines on Computer-Based Testing (International Test Commission, 2004), and the result is consistent with the True-Score Test Theory that requires parallel tests to show nearly equal mean scores (Allen & Yen, 1979). At the same time it suggests that it is the responsibility of instructional designers to craft and design

### Table 2

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Control Pre Mean (SD)</th>
<th>Control Post Mean (SD)</th>
<th>Treatment Pre Mean (SD)</th>
<th>Treatment Post Mean (SD)</th>
<th>Pillai’s trace test interaction effect (F-ratio value at df = 1, 66)</th>
<th>Treatment effect size (Cohen’s d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With testing effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Performance (Biology score)</td>
<td>69.97 (7.78)</td>
<td>63.44 (6.59)</td>
<td>68.59 (8.19)</td>
<td>73.06 (7.54)</td>
<td>20.35**</td>
<td>.57</td>
</tr>
<tr>
<td>Overall motivation</td>
<td>96.09 (7.75)</td>
<td>91.06 (7.23)</td>
<td>96.44 (9.71)</td>
<td>112.09 (7.59)</td>
<td>102.87**</td>
<td>1.79</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>15.56 (2.52)</td>
<td>14.41 (2.97)</td>
<td>15.50 (3.01)</td>
<td>18.06 (5.42)</td>
<td>9.10**</td>
<td>.58</td>
</tr>
<tr>
<td>Extrinsic</td>
<td>16.50 (3.68)</td>
<td>15.09 (3.44)</td>
<td>17.47 (4.00)</td>
<td>18.91 (3.52)</td>
<td>8.85</td>
<td>.50</td>
</tr>
<tr>
<td>Personal relevance</td>
<td>14.71 (2.45)</td>
<td>14.32 (4.21)</td>
<td>15.12 (3.27)</td>
<td>16.41 (2.36)</td>
<td>2.84</td>
<td>.45</td>
</tr>
<tr>
<td>Self-determination</td>
<td>15.91 (4.50)</td>
<td>15.74 (4.53)</td>
<td>17.21 (3.78)</td>
<td>20.79 (2.52)</td>
<td>16.65**</td>
<td>1.11</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>18.59 (4.52)</td>
<td>16.18 (3.63)</td>
<td>16.79 (4.31)</td>
<td>20.00 (2.87)</td>
<td>34.37**</td>
<td>.87</td>
</tr>
<tr>
<td>Anxiety</td>
<td>14.82 (2.36)</td>
<td>15.32 (2.17)</td>
<td>14.35 (2.53)</td>
<td>17.91 (2.88)</td>
<td>20.01**</td>
<td>1.31</td>
</tr>
<tr>
<td><strong>Testing effect removed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Performance (Biology score)</td>
<td>66.44 (9.01)</td>
<td>69.97 (7.78)</td>
<td>65.41 (10.21)</td>
<td>68.59 (8.19)</td>
<td>.19</td>
<td>.34</td>
</tr>
</tbody>
</table>

\(p < .05\)

\(p < .01\)

### Table 3

<table>
<thead>
<tr>
<th>Dependent variable (Biology score)</th>
<th>Covariate (Control variable)</th>
<th>Source</th>
<th>Mean square</th>
<th>(F(1, 66))</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test performance (Biology score)</td>
<td>Test takers’ motivation</td>
<td>Testing mode</td>
<td>494.99</td>
<td>2.32</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing mode</td>
<td>434.83</td>
<td>2.04</td>
<td>.15</td>
</tr>
</tbody>
</table>

...
high-quality computer-based tests that parallel the conventional paper-based test, and extensively pilot test them to ensure equality before implementing computer-based testing.

A critical issue of the study is about the test type and measures of testing scores in use. As acknowledged in the limitations 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-based and paper-based modes are expected to be comparable. For examples, psychological test scores have been reported as equivalent across the two testing modes in tests of personality (Davis, 1999; Fox & Schwartz, 2002), progressive behavior (Williams & McCord, 2006), sensitive behavior (Booth-Kewley, Larson, & Miyoshi, 2007), self-esteem (Vispoel, Boo, & Bleiler, 2001), morality (Cronk & West, 2002), mood (Fouladi, McCarthy, & Moller, 2002; Tseng & et al., 1998) and depression (Ogles & et al., 1998). On the other hand, achievement test may be influenced by context of test, for example, motivation and willingness of the participants to achieve higher scores in the tests. Nevertheless, the study has shown that the willingness of the participants to achieve higher scores did not produce different results between the two testing modes. Test-takers’ motivation was not a significant 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-based and computer-based testing modes on psychological test performance after removing testing effects (Chua, 2012), that testing mode has no significant effect on either psychological test or achievement test scores.

The results of this study also provide an explanation for why some previous studies have shown a significant difference between the two testing modes in test performance even though theoretically no difference should be observed. Testing effect did occur in this testing mode comparability study although it was not identified and reported by most of the researchers of past studies; instead they found significant treatment effects. However, the conclusion that computer-based testing has an effect on the experimental variables (test performance) might have been misleading and a case of misinterpretation because there is a possibility that the changes in the experiment variables are caused by testing effects, and not by the treatment effects.

In addition, the results indicate that there was a significant treatment effect on test-takers’ motivation after removing the testing effects. The computer-based testing mode increased the participants’ intrinsic motivation, self-efficacy, and anxiety. It reflects the ability of the computer-based mode to stimulate the participants to answer the computer-based testing posttest with higher motivation than in the case of paper-based testing.

Since testing is an aid to learning, and it is a practice that is part and parcel of a good educational system, an advantage of using computer-based testing, as shown in this study is that it increases test-takers’ motivation, which in turn heightens their willingness to be tested and increases testing participation rate (Wenemark et al., 2011).

7. Limitations

There are two limitations that need to be acknowledged and addressed regarding the present study. First, the limitation of the sample by virtue of the homogeneity of ethnicity, nationality and occupation should be noted. The generalizability of the present study is limited by the Malaysian student teachers. Therefore the study does not intent to generalize the findings to other ethnic groups, nations and occupations. More studies should be conducted on heterogeneous samples before broad generalizations can be made. Second, the generalizability of the present study is limited by the small sample size. A larger sample size would have been more reliable. However, despite the small sample, the significance of this study is that it provides evidence of testing effects which should be removed in analyzing and interpreting the effects of computer-testing on test performance and motivation.

Appendix A.

Dimensions and items of the BMQ (Adapted with permission: ©2006 Shawn M. Glynn and Thomas R. Koballa, Jr.).

Intrinsic – Internal factors which stimulate a response. It includes the satisfaction one gets from the task itself.

– I enjoy answering the Biology test (1).
– Answering the Biology test is more important to me than the grade I receive (16).
– I find answering the Biology test is interesting (22).
– I like the Biology test because it is challenging (27).
– Answering the Biology test gives me a sense of accomplishment (30).

Extrinsic – External factors which stimulate a response. It includes rewards such as grades, job, recognition and competition.

– I like to do better than the other students on the Biology test (3).
– Earning a good Biology grade by answering the test is important to me (7).
– I want to perform better in the Biology test because I need recognition from my classmates (10).
– I think about how the Biology test can help me get a good grade (15).
– I think about how my Biology mark can help my future career (17).

Personal relevance – Goal orientation as the reasons students to engage in a particular achievement situation that affect their valuing of the task.

– The Biology test I answer relates to my personal goals (2).
– I think about how the Biology test will be helpful to me (11).
– I think about how I will use the Biology test to achieve my goal (19).
– I like answering the Biology test because it is relevant to my personality (23).
– The Biology test I answer has practical learning value for me (25).

Self-determination – The belief that one can master a situation and produce positive learning outcomes.

– Answering the Biology test is not difficult (5).
– I put enough effort into answering the Biology test (8).
– I use strategies that ensure I answer the Biology test well (9).
– I can learn better from taking the Biology test (20).
– The Biology test is what I have expected (21).

Self-efficacy – A person’s belief in his or her ability to succeed in a particular situation in learning.

– I expect to do as well as or better than other students in any Biology test (12).
– I am confident I can answer most of the questions in the Biology test (21).
– I believe I have master the knowledge and skills in the Biology test (24).
– I am confident I did well in the Biology test (28).
– I believe I can earn a good grade in the Biology test (29).

Anxiety – Students' emotional reactions to the task. It includes tension, apprehension and nervousness.

– I am nervous about how I will do on the Biology test (4).
– I become anxious when it is time to take a Biology test (6).
– I worry about failing the Biology test (13).
– I am concerned that the other students are better in the Biology test (14).
– I hate taking the Biology test (18).

References