Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 64 (2012) 655 – 664

Procedia
Social and Behavioral Sciences

INTERNATIONAL EDUCATIONAL TECHNOLOGY CONFERENCE
IETC2012

# Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong?

Chua Yan Piaw[a]

[a]*University of Malaya, Institute of Educational Leadership, UM City Campus Complex, Kuala Lumpur, 43300 Malaysia*

**Abstract**

The standards for developing computerized-assessment required equivalent test scores to be established for the paper-based test (PBT) and computer-based test (CBT) modes. However, in most studies, the two modes were nearly identical, yet significant differences of test scores were observed. Therefore the validity of replacing PBT with CBT in educational assessment was questioned. This study employed an achievement test, a psychological test and a motivation questionnaire in a Solomon four-group design to examine validity of the CBT and its effects on test performance and motivation. The findings of this study provide evidences for the issue of CBT's validity in educational and psychological assessment.



Keywords: Assessment; computer-based testing; testing effect; performance; motivation

## 1. Introduction

The interest in developing and using computer-based test (CBT) in educational assessment in schools and educational institutions has heightened in recent years. Delivering assessments via computers is becoming more and more prevalent in educational assessment domain as changes are made in assessment methodologies that reflect practical changes in pedagogical methods (Kate Tzu, 2012; Genc, 2012; Hsiao, Tu & Chung, 2012; OECD, 2010). CBT is seen as a catalyst for change, bringing transformation of learning, pedagogy and curricula in educational institutions (Scheuermann & Pereira, 2008).

To establish a valid and reliable CBT, the International Guidelines on Computer-Based Testing (International Test Commission 2004) stated that equivalent test scores should be established for the conventional paper-based testing (PBT) and its computer-based mode. This set of testing standards is supported by the classical true-score test theory – the basis of computer-based and paper-based testing (Allen & Yen 1979). Under this theory, a test taker who takes the same test in the two modes is expected to obtain nearly identical test scores. The standards are also supported by empirical studies (OECD, 2010;

Corresponding Author: Chua Yan Piaw chuayp@um.edu.my

 doi:10.1016/j.sbspro.2012.11.077

Wilson, Genco, & Yager, 1985). For example, OECD (2010) reported that there were no difference in test performance between CBT and PBT among student participants (n = 5,878) from Denmark, Iceland and Korea.

Interestingly, however, in a review of educational and psychological measurement approaches, Bunderson, Inouye & Olsen (1989) reported that 48% of previous studies showed no difference between the two testing modes in test performance, 13% of studies showed the superiority of CBT and 39% of studies showed that PBT was superior. The concept of equivalence was supported by only nearly half of the studies, and the differences were ascertained in achievement tests such as science, language and mathematics tests, and also obviously in psychological tests such as personality and neuropsychological assessment (e.g. Friedrich & Bjornsson, 2008; Choi, Kim & Boo, 2003; DeAngelis, 2000).

A possible explanation for this phenomenon is either CBT has a low validity as an assessment tool for educational and psychological measurements, or there might have been other effect that confounded the effect of testing mode on test performance in these repeated-measures studies. As observed by Yu & Ohlund (2010), a possible confounding variable is testing effect; the effect of taking a pretest on taking a posttest that systematically confounds the treatment effect of CBT on test performance.

## 2. Testing effect in repeated-measures studies

A careful review to the literature discovered that most of testing mode comparability studies has been conducted using pretest-posttest experimental designs without identifying testing effects on test takers. Therefore, the findings might be misinterpreted. For example, in a study, a participant answered the same test four times for two pretests and two posttests, "each subject took the same pretest and posttest on paper and computer" (Al-Amri, 2008; *p.*29). The limitation of this design is testing effect might occur when a participant is tested at least twice on a same test, and the act of taking a pretest might influence the outcome of a posttest (Chua, 2011b; Yu & Ohlund, 2010; Shuttleworth, 2009), and it is a bias for a researcher to confidently conclude that there is a treatment effect although the result is significant. This issue needs further research because the Standards for Educational and Psychological Testing guidelines (APA, 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.

## 3. Effects of testing motivation on the relationship between testing modes and test performance

Another issue that needs to be clarified in a PBT and CBT comparability study, as raised by Wise and DeMars (2003) is motivational factors which might also have an impact on test performance. Wise and DeMars pointed out that regardless of how much psychometric care is applied to test development, or how equal the testing modes are, to the extent that test takers are not motivated to respond to the test (e.g. due to low efficacy or boredom), test score validity will be compromised. The test taker motivation model (Pintrich, 1989) specifies that the effort test takers will direct towards a test is a function of how well they feel they will do on the test, how they perceive the test to be, and it related to their affective reactions regarding the test. This is the theoretical model that underlies the relationship among motivation, testing mode and test performance. Besides that, the self-determination theory (Wenemark, Persson, Brage, Svensson & Kristenson, 2011) states that increases test-takers' motivation will increase the willingness to take the test or response rates, and thus it will enhance learning. Therefore, testing motivation is 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).

One of the barriers to the implementation of CBT in educational and psychological measurements in education is insufficient study of the equivalence of CBT and PBT (Bugbee, 1996). To overcome the potential for misinterpreting experimental results caused by testing effects, Yu & Ohlund (2010) strongly recommended the use of the Solomon four-group design. This design helps researchers to detect the occurrence of testing effects in an experimental study. Therefore, this study employed a Solomon four-group experimental design to examine the validity and effectiveness of CBT by comparing it with the PBT. It examined whether testing effects occur in CBT and PBT, and investigated the effects of testing motivation on the relationship between testing modes and test performance.

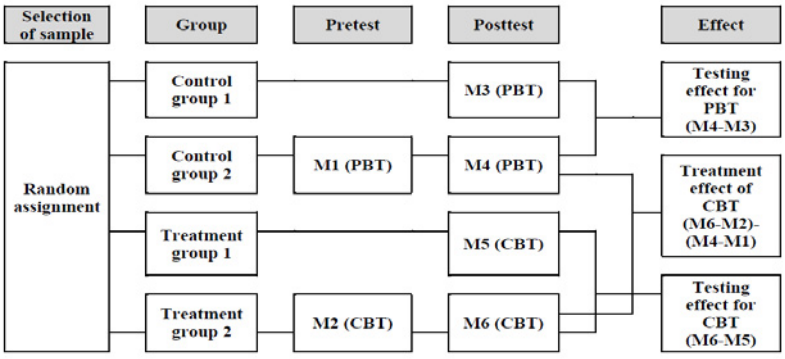
## 4. Method

### *4.1. Research Design*

The Solomon four-group experimental design is "one of the best methods to identify testing effects in experimental designs" (Yu & Ohlund, 2010, p. 9). It consists of two basic categories of research designs: (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 effect besides the treatment effects on experimental variables. It should be pointed out that besides identifying treatment effect, the intention of the design is to help researchers determine if testing effect occurs, that is, to detect whether the change in experimental variable is caused by the change in the treatment effect or testing effect.

The values of M4–M3 and M6–M5 (see Figure 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, there are no testing effects. Therefore, the (M6–M2)–(M4–M1) value will give an estimation of the treatment effect. 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 has an effect on the experimental variables (test performance and testing 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 treatment effect of CBT, if testing effect occurs in M4 (PBT posttest), then it will be replaced with M3. This is because the two PBT posttest scores are identical if testing effect does not occur in M4. The same applies to the CBT posttest. If testing effect occurs in M6, then it will be replaced with M5 in the treatment effect analysis.




Note: M = Measurement

Fig. 1. Design of the Study

To analyse 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 ANOVA analysis is carried out to identify the treatment effects. A CBT treatment effect is detected if a significant interaction effect occurs. Split-Plot ANOVA is one of the most powerful quantitative research methods for testing causal hypotheses (Yu & Ohlund, 2010; Chua, 2009a).

## *4.2. Instruments of the Study*

Three instruments used in this study were the Biology test, the YBRAINS test and the Testing Motivation Questionnaire.

(a) *The Biology Test* - The Biology Test is an educational achievement test that consists of 40 multiple-choice items, each item for 2.5 score, with a total test score of 100. The items were developed from seven topics: (1) cell structure and cell organization, (2) movement of substances across the plasma membrane, (3) chemical composition of the cell, (4) nutrition, (5) respiration, (6) dynamic ecosystem and (7) endangered ecosystem. It collected data for the participants' test performance when they answered the Biology Test in PBT and CBT modes for the purpose of comparison. Its test-retest reliabilities (Pearson correlation coefficients) for PBT and CBT versions were .86 and .83.

(b) *The YBRAINS test* - The YBRAINS test (Chua, 2011a) is a psychological test that collected test scores for critical thinking and creative thinking style when participants answered the test in PBT and CBT modes for the purpose of comparison. The PBT YBRAINS test was adapted to a CBT mode in 2009 (see Fig. 2). Both CBT and PBT deliver the same content. The computer-based YBRAINS has won two gold medals at green technology innovation expos, including at the 21st International Invention, Innovation and Technology Exhibition 2010 (ITEX'10, 2010).

The test consisted of 34 items which were used to measure simultaneously the critical and creative thinking styles of a participant. Each item of the test provided the participants with multiple choices – each choice representing a specialised trait of critical thinking or creative thinking style. Each participant was asked to indicate the specific traits that best described his or her own typical behaviour. The responses were then calculated to obtain critical thinking and creative thinking style scores.

For the CBT mode, the test was developed in a computer-based system using Visual Basic. When a participant responded to the test items, his thinking styles (critical and creative styles) would be shown instantly by the computer program. The test scores (critical thinking style and creative thinking style) were recorded in a Microsoft Access database immediately after a participant had completed the CBT test. For the PBT mode, the test score for each participant was calculated manually by the researcher using the same scoring format. Its test-retest reliabilities (Pearson correlation coefficient) for the PBT mode were .71 (critical thinking style) and .78 (creative thinking style), and for CBT versions were .77 (critical thinking style) and .82 (creative thinking style).


Fig. 2. An Example Test Item and the Results of the Test in Graphical Form

(c) *The Testing Motivation Questionnaire* - The third instrument is the adapted version of the Testing Motivation Questionnaire or TMQ (Wigfield, Guthrie & McGough, 1996) (see Appendix A). It measured overall testing motivation and four motivation components (self-efficacy, extrinsic, intrinsic and social motivations) of the participants towards the two testing modes for comparison. The components consist of eleven dimensions of motivation. Challenge and efficacy are categorised under self-efficacy motivation. Curiosity, involvement, importance and work avoidance are categorised under intrinsic motivation. Competition, recognition and grades are listed under extrinsic motivation, and finally social and compliance are the dimensions of social motivation. Although questions have been raised about the factor structure of the motivation dimensions (Watkins & Coffey, 2004), several studies examining its validity and reliability have supported these eleven dimensions (Parault & Williams, 2009; Unrau & Schlackman, 2006; Wigfield & Guthrie, 1997). Based on the motivation dimensions, Wigfield, Guthrie & McGough (1996) developed a 54-item motivation questionnaire to examine a group of students' reading motivation. Since motivation is a universal human behaviour and is identical across disciplines (Guthrie & Wigfield, 1999, *p*. 199), the eleven dimensions were adapted into this study as the dimensions of testing motivation. The TMQ was developed based on a five-point Likert scale to assess participants' motivation towards the two testing modes. The scores ranged from 1 (very different from me) to 5 (a lot like me). The internal consistency reliabilities (Cronbach's alpha) for the eleven motivation dimensions in the PBT and CBT versions were ranged between .72 and 83.

## *4.3. Participants*

The participants in this study were 140 Malaysian undergraduate student teachers from a teacher training institute located in Peninsular Malaysia. Among the participants, there were 61 males (43.57%) and 79 females (56.43%) with an average age of 21 years. The participants were randomly selected from a student teacher population (N = 219) based on the sample size determination table of Krejcie and Morgan (Chua, 2011b, *p*.211) at a 95% ($p < .05$) confidence level. They were enrolled in a teacher education programme (mathematics and science), and have the same educational history and background. They have the same level of computer applications skill and received formal computer instruction in their academic curriculum. Based on their performances in a biology monthly test and the recommendation of their lecturers, the student teachers with similar abilities were arranged into 35 equivalent groups (each with four equivalent participants). The four participants in each group were then assigned into four groups through a simple random sampling procedure, each with a sample size of 35. The four groups were then randomly assigned to two control and two treatment groups for the experimental study.

## *4.4. Procedures*

At the first phase, control group 2 answered PBT mode of the Biology Test and YBRAINS test and treatment group 2 answered their CBT modes (pretests for test performance). Immediately after the tests, the two groups answered the TMQ questionnaire to identify their motivation towards the two testing modes (pretests for testing motivation). Two week later, at the second phase, the four groups answered the Biology Test and the YBRAINS test. The two control groups answered the PBT modes and the two treatment groups answered the CBT modes (posttests for test performance). Immediately after the tests, the four groups answered the same TMQ questionnaire to identify their motivation towards the two testing modes (posttests for testing motivation).

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

# 5. Results

## 5.1. Testing Effect

The data in Table 1 indicates that there were no significant testing effects on the scores of the achievement and psychological tests for PBT and CBT modes (p>.05).

However, significant testing effect were found in overall testing motivation [t(68)=–8.89, *p*=.00; d=2.16] and its three motivation components, that is, self-efficacy [t(68)=–6.48, *p*=.00; d=1.57], intrinsic motivation [t(68)=–4.81, *p*=.00; d=1.17] and social motivation [t(68)=–4.27, *p*=.01; d=1.04] with large testing effect (Cohen's d>.80). More specifically, significant testing effects occurred in six of the eleven motivation dimensions, namely efficacy, curiosity, involvement, work avoidance, competition and compliance. The negative mean difference values for overall testing motivation and its three components; self-efficacy motivation; intrinsic motivation and social motivation show that the PBT posttest motivation scores were lower than their pretest scores, and it indicates that fatigue testing effects occurred in the PBT. It means that the participants were less motivated to complete the PBT posttest. On the other hand, no significant testing effects were found in the CBT for total testing motivation and all the four testing motivation components.

Table 1. Testing effects for PBT and CBT modes on test performance and testing motivation

| Subscale | Testing Effect for PBT | | | | | Testing Effect for CBT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Control group 1 | Control group 2 | Mean Dif. | T test | Effect size (d) | Treatment group 1 | Treatment group 2 | Mean dif. | T test | Effect size (d) |
| | Mean (*SD*) | Mean (*SD*) | | t value at df = 68 | | Mean (*SD*) | Mean (*SD*) | | t value at df = 68 | |
| **Performance** | | | | | | | | | | |
| Biology score | 68.2 (12.1) | 64.7 (12.1) | -3.50 | -1.22 | -.29 | 69.55(13.2) | 68.1 (13.2) | -1.43 | -.42 | .10 |
| Critical style | 10.1 (1.5) | 9.9 (2.1) | -.18 | -.52 | .13 | 10.2 (1.5) | 10.3 (1.6) | .11 | .21 | .05 |
| Creative style | 11.2 (1.3.) | 11.1 (1.1) | -.04 | .15 | .04 | 11.2 (2.37) | 12.8 (2.2) | 1.52 | .98 | .23 |
| **Motivation** | 133.5 (11.2) | 115.6 (9.0) | -17.89 | -8.89** | 2.16 | 155.1 (13.1) | 158.8 (13.1) | 3.64 | 1.32 | .32 |
| *1. Self-efficacy* | 20.5 (2.1) | 16.5 (2.5) | -3.97 | -6.48** | 1.57 | 25.6 (3.5) | 26.4 (3.6) | .71 | 1.07 | .26 |
| Challenge | 10.7 (2.3) | 11.6 (2.8) | .87 | .58 | .14 | 13.6 (2.4) | 13.8 (2.5) | .19 | .25 | .06 |
| Efficacy | 9.7 (3.6) | 4.8 (1.5) | -4.84 | -8.46** | 2.05 | 12.0 (1.5) | 12.5 (1.7) | .52 | 1.16 | .28 |
| *2. Intrinsic* | 46.8 (6.4) | 38.0 (6.8) | -8.74 | -4.81** | 1.17 | 63.0 (5.7) | 62.0 (6.0) | -.99 | 1.47 | .36 |
| Curiosity | 13.1 (2.5) | 5.6 (2.3) | -7.50 | -13.34** | 3.23 | 18.5 (2.2) | 19.4 (2.2) | .91 | 1.36 | .33 |
| Importance | 15.2 (4.8) | 15.2 (3.1) | .00 | .24 | .06 | 14.5 (3.5) | 15.2 (3.6) | .72 | .99 | .24 |
| Involvement | 11.7 (2.7) | 7.3 (3.2) | -4.45 | -6.17** | 1.50 | 17.3 (2.6) | 17.4 (2.8) | .13 | .15 | .04 |
| W. avoidance | 6.5 (1.6) | 9.7 (2.3) | 3.21 | 7.01** | 1.70 | 12.6 (3.4) | 9.8 (3.3) | -2.75 | - .83** | .92 |
| *3. Extrinsic* | 37.1 (5.3) | 36.5 (4.2) | -.68 | -.81 | .20 | 35.2 (3.4) | 35.9 (3.5) | .71 | .63 | .15 |
| Competition | 19.48 (1.34) | 18.34 (1.17) | -1.14 | -2.31* | .56 | 17.3 (2.1) | 17.7 (1.1) | .34 | -.91 | .22 |
| Recognition | 9.25 (2.87) | 9.48 (3.17) | .23 | .30 | .07 | 8.9 (3.1) | 9.6 (3.1) | .75 | .86 | .21 |
| Grade | 8.46 (1.09) | 8.69 (1.19) | .23 | -2.43 | .60 | 8.8 (1.3) | 8.5 (1.2) | -.38 | -1.53 | .37 |
| *4. Social* | 29.06 (3.31) | 24.56 (3.62) | -4.50 | -4.27* | 1.04 | 31.2 (3.2) | 34.5 (3.6) | 3.21 | 1.44 | .35 |
| Social | 12.23 (2.34) | 12.87 (2.16) | .64 | .67 | .16 | 16.51 (3.35) | 17.88 (3.51) | 1.37 | 1.36 | .33 |
| Compliance | 16.83 (2.59) | 11.69 (2.27) | -5.14 | -7.40** | 1.79 | 14.78 (2.32) | 16.62 (2.47) | 1.84 | 1.08 | .26 |

Notes: **p*<.05, ***p*<.01. The values of Cohen's d effect size were calculated based on the mean and standard deviation scores. Cohen (1988) defined effect sizes as "small when *d* = .21 to .49," "medium when *d* = .50 - .79," and "large when *d* ≥ .80".

## 5.2. Treatment Effect

The results of the Split-Plot ANOVA analysis (multivariate analysis of variance using the Pillai's Trace test) after eliminating the testing effects (as shown in Table 2) indicate that no significant treatment effects were found in test performance for biology score and the two thinking style sub-scales.

As a whole, significant treatment effect occurred in total testing motivation [F(1, 68)=15.68, p<.01; d=.69]. The CBT significantly increased self-efficacy motivation [F(1, 68)=23.26, p<.01; d=.94], intrinsic motivation [F(1, 68)=27.59, p<.01; d=1.10] and social motivation [F(1, 68)=38.52, p<.01; d=1.22] of the

participants. The data also indicates that treatment effects significantly occurred in five of the eleven test motivation dimensions, they are challenge, efficacy, curiosity, involvement and social, and the treatment effect sizes were medium to large (d values were between .57 to 1.37). It indicates that the CBT mode has significantly increased the motivation level of the participants.

Table 2. Split-Plot ANOVA analysis results for the effect of CBT on test performance and testing 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, 68) | |
| | Mean (*SD*) | Mean (*SD*) | Mean (*SD*) | Mean (*SD*) | | |
| **Performance** | | | | | | |
| Biology score | 67.5 (12.8) | 68.2 (12.1) | 68.5 (13.1) | 68.9 (12.7) | .26 | -.03 |
| Critical style | 9.8 (1.4) | 10.1 (1.5) | 9.7 (1.3) | 10.3 (1.4) | .12 | .28 |
| Creative style | 10.9 (1.2) | 11.2 (1.3) | 11.4 (1.2) | 11.7 (2.6) | .28 | -.11 |
| **Motivation** | 124.9 (18.5) | 133.5 (11.2) | 131.4 (16.1) | 150.1 (13.6) | 15.68** | .69 |
| *1. Self-efficacy M.* | 17.5 (4.8) | 20.5 (2.1) | 17.1 (4.5) | 24.1 (3.3) | 23.26** | .94 |
| Challenge | 9.7 (2.8) | 10.7 (2.3) | 9.5 (2.9) | 13.8 (2.7) | 28.35** | 1.11 |
| Efficacy | 7.7 (2.1) | 9.7 (3.6) | 7.5 (2.7) | 11.2 (2.7) | 13.27** | .67 |
| *2. Intrinsic M.* | 44.1 (9.1) | 46.8 (6.4) | 51.5 (4.9) | 58.3 (4.4) | 27.59** | 1.10 |
| Curiosity | 12.2 (2.8) | 13.1 (2.5) | 12.8 (2.6) | 16.2 (2.2) | 26.52** | 1.06 |
| Importance | 12.5 (3.3) | 15.2 (4.8) | 13.7 (2.7) | 15.5 (3.6) | 1.74 | .15 |
| Involvement | 11.8 (3.7) | 11.7 (2.7) | 12.7 (2.6) | 14.2 (2.7) | 12.71** | .57 |
| Work avoidance | 7.3 (2.4) | 6.5 (1.6) | 12.2 (3.3) | 12.2 (3.1) | 2.74 | .40 |
| *3. Extrinsic M.* | 33.5 (3.1) | 37.1 (5.3) | 33.9 (3.4) | 34.9 (3.5) | 2.06 | -.55 |
| Competition | 17.3 (4.1) | 19.4 (1.3) | 17.6 (1.4) | 17.2 (1.8) | .69 | -.92 |
| Recognition | 7.5 (2.2) | 9.2 (2.8) | 7.9 (2.2) | 9.1 (3.0) | 1.22 | -.22 |
| Grade | 8.6 (1.5) | 8.4 (1.0) | 8.3 (1.6) | 8.5 (1.3) | .57 | .29 |
| *4. Social M.* | 28.7 (4.3) | 29.0 (3.3) | 28.3 (4.0) | 33.7 (4.2) | 38.52** | 1.22 |
| Social | 12.3 (2.7) | 12.2 (2.3) | 12.4 (3.4) | 17.1 (3.5) | 39.51** | 1.37 |
| Compliance | 16.3 (2.5) | 16.8 (2.5) | 16.2 (3.8) | 16.6 (3.5) | .87 | -.08 |

Notes: *$p$<.05, **$p$<.01

To further understand the association among test performance and testing motivation, a Pearson Product-moment inter-correlation test was conducted (see Table 3). Besides that, since there was a treatment effect of CBT on testing motivation, an Analysis of Covariance (see Table 4) was performed to identify whether testing motivation is a moderator variable for the association between testing mode and test performance.

Table 3. Pearson Product-moment Inter-correlation between test performance and testing motivation

| Correlation | Test Performance | | |
|---|---|---|---|
| | Biology Score | Critical Style | Creative Style |
| Testing Motivation | -.20 | -.17 | .13 |

Table 4. Analysis of Covariance for testing motivation towards the effect of CBT on test performance

| Score | Source | Mean Square | F(1, 31) | p |
|---|---|---|---|---|
| Biology test | Testing motivation | 494.99 | 2.32 | .13 |
| | Testing mode | 434.83 | 2.04 | .15 |
| Critical style | Testing motivation | .40 | .24 | .62 |
| | Testing mode | .35 | .21 | .64 |
| Creative style | Testing motivation | .52 | .76 | .38 |
| | Testing mode | .32 | .471 | .49 |

Table 3 indicates that there were no significant correlation between the three test performance scores with and testing motivation. It means answering the test with greater testing motivation would not

necessary help a test taker to achieve a higher test performance score. Furthermore, the data in Table 4 shows that there were no significant main effects of CBT on the three test performance scores and testing motivation was not a significant moderator for the effect of CBT on test performances of the achievement test and psychological test.

## 6. Discussion

Results of the analyses indicate that no significant testing and treatment effects were found for test performance in the two testing modes. In other words, the test scores were consistent over time and across the two testing modes. It shows that a participant who sits for both the CBT and PBT would most probably yield similar pretest and posttest scores. The two CBT tests are valid in terms of test performance and can be used as a replacement for their PBT.

The results also indicate that the achievement test and psychological test have fulfilled the requirements of the international guidelines on computer-based testing (International Test Commission 2004) and consistent with true-score test theory (Allen & Yen, 1979) that parallel tests are required to show nearly equal mean scores. However, it does not support the suggestion of some researchers (e.g. Clariana & Wallace, 2002) that it is not necessary that equivalent measures be produced from CBT and PBT versions; at the same time it suggests that it is the responsibility of instructional designers to craft and design high-quality CBTs that parallel the conventional PBTs, and extensively pilot test them to ensure equality before implementing computer-based testing.

The results of this study also provide an explanation for why some previous studies have revealed a significant difference between the two testing modes in test performance although theoretically no difference should be observed. Testing effects did occur in this testing mode comparability study although none was identified and reported by the researchers of past studies; instead they found significant treatment effects. However, for the researchers to conclude that CBT has an effect on the experimental variables (test performance) is misleading because there is a possibility that the changes in the experiment variables are caused by testing effects, and not by the treatment effects. Thus, the findings of these studies might have been jeopardised by testing effects and misinterpreted.

The findings also show that the CBT mode is more stable and consistent in terms of internal and external validity because no testing effects were found in all of the four testing motivation components. For treatment effect, the results indicate that there was a significant treatment effect on testing motivation. The CBT had increased the participants' self-efficacy, intrinsic and social motivation. It reflects the ability of the CBT to stimulate the participants to answer the CBT posttest with higher concentration.

However, answering the tests with greater testing motivation did not help a test taker to achieve higher scores; no significant treatment effects were found in the two tests. This is another interesting finding, that testing motivation is not a catalyst for the effect of testing mode on test performance. The study rejects the prediction of some previous studies that motivation level of test takers to answer the CBT and PBT might have an impact on test performances (e.g. Wise & DeMars, 2003). It provides evidence that testing motivation is not a moderator of the relationship between testing mode and test performance. It is consistent with the finding of OECD (2010), that the effects of motivational factors on the relationship between testing mode and test performance are insignificant and either very weak or non-existent.

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 CBT, as generated from this study is that it produces more valid test results for repeated measures and increases test-takers' motivation which will, in turn, heighten their willingness to be tested and increases testing participation rate. Based on the results of this study, computer-based testing can be used as a valid replacement for the conventional paper-based testing in educational institutions.

**References**

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 and Linguistics*, *10*, 22–44. Retrieved January 28, 2012 from http://www.essex.ac.uk/linguistics/publications/egspll/volume_10/pdf/EGSPLL10_22-44SAA_web.pdf .

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

Bugbee, A. C. (1996). The equivalent of paper-and-pencil and computer-based testings. *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., K. S. Kim, and J. Boo. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, *20*(3), 295–320.

Chua, Y. P. (2004). *Creative and critical thinking styles*. Serdang, Malaysia: Universiti Putra Malaysia Press.

Chua, Y. P. (2008). *Research methods and statistics book 3: Data analysis for nominal and ordinal scales*. Shah Alam, Malaysia: McGraw-Hill Education.

Chua, Y. P. (2009a). Writing a series of best-selling research reference books. *Journal of Scholarly Publishing*, *40*(4), 408–419. doi: 10.3138/jsp.40.4.408.

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

Chua, Y. P. (2011a). Establishing a brain styles test: The YBRAINS test. *Procedia Social and Behavioral Sciences*, *15*, 4019–4027. Retrieved August 16, 2011 from http://www.sciencedirect.com/science/article/pii/S1877042811009530.

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

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.

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

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

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

Friedrich, S., & Bjornsson, J. (2008). *The transition to computer-based testing – New approaches to skills assessment and implications for large-scale testing*. http://crell.jrc.it/RP/reporttransition.pdf (accessed May 23, 2011).

Genc, H. (2012). An evaluation study of a call application: with belt or without belt. *TOJET: The Turkish Online Journal of Educational Technology*, 11(2). Retrieved July 2, 2011 from http://www.tojet.net/articles/v11i2/1125.pdf

Guthrie, J. T., & Wigfield, A. (1999). How motivation fits into a science of testing. *Scientific Studies of Testing*, *3*, 199–205.

Hsiao, H. C., Tu, Y. L., Chung, H. N. (2012). Perceived social supports, computer self-efficacy, and computer use among high school students. *TOJET: The Turkish Online Journal of Educational Technology*, 11(2).

International Test Commission. (2004). *International Guidelines on Computer-Based and Internet-Delivered Testing*. Retrieved January 21, 2011 from http://www.intestcom.org/itc_projects.htm.

ITEX '10. (2010). Results of the International Invention, Innovation and Technology Exhibition 2010, May 14–16, 2010. Retrieved January 2, 2012 from http://www.ippp.um.edu.my/images/ippp/doc/itex%202010.pdf.

Kate Tzu, C. C. (2012). Elementary EFL teachers' computer phobia and computer self-efficacy in Taiwan. *TOJET: The Turkish Online Journal of Educational Technology*, 11(2). Retrieved June 18, 2012 from http://www.tojet.net/articles/v11i2/11210.pdf

Morgan, C., and O'Reilly, M. (2001). Innovations in online assessment. In F. Lockwood and A. Gooley (Eds.), *Innovation in Open and Distance Learning: Successful Development of Online and Web-based Learning* (pp. 179–188). London: Kogan Page.

OECD. (2010). *PISA Computer-based assessment of student skills in science*. http://*www.oecd.org/publishing/corrigenda* (accessed December 21, 2011).

Parault, J. S., & Williams, H. M. (2009). *Testing motivation, testing amount, and text comprehension in deaf and hearing adults*. Oxford, UK: Oxford University Press. doi:10.1093/deafed/enp031

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*. Retrieved January 12, 2012 from http://www.psychmet.com/id16.html.

Shuttleworth, M. (2009). Repeated measures design. *Experiment Resources*. http://www.experiment-resources.com/repeated-measures-design.html (accessed January 25, 2012).

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 Report,* 23306 EN.

Unrau, N., & Schlackman, J. (2006). Motivation and its relationship with reading achievement in an urban middle school. *Journal of Educational Research*, *100*, 81–101.

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

Watkins, M. W., & Coffey, D. Y. (2004). Testing motivation: Multidimensional and indeterminate. *Journal of Educational Psychology*, *96*, 110–118.

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.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81.

Wigfield, A., & Guthrie. J. T. (1997). Relations of children's motivation for testing to the amount and breath of their testing. *Journal of Educational Psychology*, *89*, 420–432.

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 Education Assessment Conference, Seattle.

Yu, C. H., & Ohlund, B. (2010). *Threats to validity of research design*. Retrieved January 12, 2012 from http://www.creative-wisdom.com/teaching/WBI/threat.shtml.