COMPARISONS BETWEEN COMPUTER-BASED TESTING AND PAPER-PENCIL TESTING: TESTING EFFECT, TEST SCORES, TESTING TIME AND TESTING MOTIVATION

Chua Yan Piaw
Institute of Educational Studies,
University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: chuayp@um.edu.my

ABSTRACT

Computer-based testing is an effective green computing strategy used to reduce paper consumption. Previous studies have been conducted to evaluate the comparability of Computer-based testing (CBT) and paper-pencil testing (PPT). Some studies revealed significant differences between the two testing modes on test score, testing time and testing motivation, while others have reported opposite or inconsistent results. However, most of the studies have been conducted through basic experimental or quasi-experimental designs without identifying testing effects on test takers. In experimental designs, testing effects influence the cause-effect relationship between the treatment (testing modes) and experimental variables. Therefore, the findings might be misinterpreted. This study employed a Solomon four-group experimental design to (1) identify the reliability of the two testing modes, (2) identify and compare the testing effects between the two testing modes, and (3) examine the effects of the two testing modes on test score, testing time and testing motivation. Results indicate that as a whole, testing effects significantly influenced testing time and testing motivation for the PPT mode. The CBT mode was more reliable in terms of internal and external validities and it reduced testing time and increased testing motivation of the participants.

Keywords*: *Computer-based testing, paper-pencil testing, experimental design, testing effects, test score, testing Time, testing Motivation

1.0 INTRODUCTION

Computer-based testing or computer-based assessment is a green computing (green IT) strategy used to reduce paper consumption. Imagine how many tons of paper would be saved within a year if schools, universities and educational institutions were to replace paper-pencil testing (PPT) with computer-based testing (CBT). Reducing paper consumption will indirectly reduce greenhouse gases and energy consumption. Minnesota Pollution Control Agency reported that 40 reams of paper are equal to 1.5 acres of pine forest absorbing carbon for a year and each ream of paper is equal to roughly 12 pounds of carbon dioxide not removed from the atmosphere [1]. Through the paper making process, paper industry represents around 10% of all global greenhouse emissions. Energy consumption by the paper industry is projected at 25.8 billion kWh of electricity and 54.3 billion BTU's of fossil fuels in 2010 [2].

Researches have been conducted to evaluate the comparability of CBT and PPT. Some studies revealed that there was a significant difference between the two testing modes on test scores [3], [4], while other studies reported opposite or inconsistent results [5], [6]. The inconsistency of the results led some scholars to suggest for them to conduct systematic studies to carefully check equivalency and comparability, as well as reliability [7], [8] and validity [5], [9] of CBT and PPT before opting for computer-based testing. Besides that, some studies have been focusing on the association between the two assessment modes and test taker variables in some areas of human behaviours, such as testing motivation [3], [10]. For instance, a research revealed that students enjoyed the CBT more than the PPT and were more motivated to perform another CBT than another PPT [11], testing motivation was reported to be negatively associated with testing time [12], and test takers positively preferred CBT for variables such as focusing attention, enjoyment and self- efficacy [13], [6].

However, a careful examination of the research methods used in these studies found that most of the studies have been conducted using basic experimental or quasi-experimental designs without identifying testing effects on test takers. Therefore, the findings might be misinterpreted. For example, In Al-Amri's study, a participant answered the same test four times, two times for the pretest and two times for the posttest, "each subject in the control group and treatment group took the same test once on paper and once on computer" [5] (*p29*). In another example, Bodmann & Robinson used a two groups (treatment and control) repeated measures (pretest and posttest) design to identify the effect of CBT on testing time, without reporting the testing effects [12]. The limitation of these design is testing effects might occur when a participant is tested at least twice on a same test, and the knowledge and experience of taking a pretest influenced the outcomes of taking a posttest [14], [15].

Therefore, there is a possibility that the change in the posttest score is due to testing effect and not the treatment effect (testing modes), and it is a bias for a researcher to confidently conclude that there is a treatment effect although the result is significant. This is because testing effect jeopardizes internal validity (whether an experimental treatment really makes a difference or not on the experimental variables) and external validity (a pretest might increase or decrease a subject's sensitivity or responsiveness to the experimental variables) of the experimental results [14]. Therefore, experimental designs need cross-validation at various times and conditions before the results can be theoretically interpreted with confident. To overcome the problem of misinterpreting experimental results caused by testing effects, Campbel & Stanley strongly recommended the Solomon four-group experimental design.

This design helps researchers to detect the occurrence of testing effects in an experimental study [15].

2.0 OBJECTIVES OF THE STUDY

Taking into consideration all of the issues discussed above, the study is conducted to (1) identify the reliability of the CBT and PPT, (2) identify and compare the testing effect between CBT and PPT, and (3) to examine the effects of CBT and PPT on test scores, testing time and testing motivation.

3.0 METHOD

3.1 Research Design

To attain the objective of this study, this study employs the Solomon four-group experimental design [16]. The advantage of this design compared to the basic two groups pretest and posttest design is that it is capable of identifying the occurrence of testing effect besides the treatment effects on experimental variables. In the Solomon four-group experimental design, there are two basic experimental research designs. They are: (1) Two groups of participants who are given treatment and two groups of participants who are not given treatment, (2) Two groups of participants who are given the pre-test and two groups of participants who are not given the pre-test.

For each treatment condition, there will be a group which is given the pre-test and another group which is not. It is one of the best methods to identify testing effects in experimental designs [14] (*p9*). It should be pointed out that the intention of the design is not to reduce the testing effect but to help researchers determine if the effect occurs, that is to detect whether the change in experimental variable is merely caused by the change in the treatment, or is caused by the testing effect. Fig. 1 shows the Solomon four-group experimental design for this study.

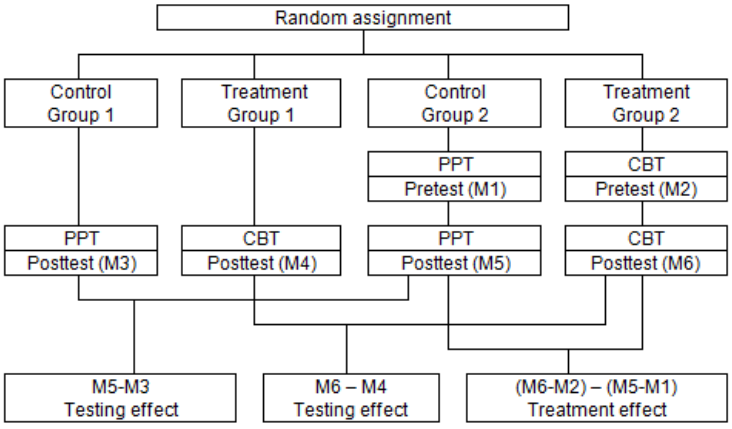

Fig. 1: The Solomon four-group design

The values of M5-M3 and M6 – M4 are the testing effects for the control and treatment groups. If there are no differences between the values of M5 and M3 as well as M6 and M4, there are no testing effects. Therefore, the (M6-M2) – (M5-M1) value will give an estimation of the treatment effect. However, any difference between M5 and M3 or M6 and M4 is caused by the pre-test effect in M1 and M2. In this case, the researcher cannot simply conclude that the treatment has an effect on the experimental variables (such as test score, testing time and testing motivation) if there is a significant treatment effect, it is because there is a possibility that the changes in the variables are caused by testing effects.

To analyse the data for the design, two steps are needed: (1) A t-test is performed to identify the testing effects (M5 – M3) or (M6 – M4), (2) A split-plot ANOVA analysis is carried out to identify the treatment effect. A treatment effect is detected if a significant interaction effect occurs. The split-plot ANOVA is one of the most powerful quantitative research methods for testing causal hypotheses [14], [24].

3.2 Instruments of the Study

Two instruments used in the study to collect data are the Yanpiaw Brain Styles test (YBRAINS) [17] and the Testing Motivation Questionnaire. The reason for employing the two psychological tests in this study, instead of using achievement tests (e.g. mathematics tests) is to minimize historical and maturity effects. This is because psychological traits such as thinking style are more consistent over time and have less historical and maturity effects compared to achievement skill such as mathematics score [18]. Historical effect is events (e.g. reading books,

watching TV programme, or exposed to other sources) which occur uncontrollably and influence the responses of the research participants, while maturity refers to any change which exists or surfaces in a participant during the course of the experiment [18] (*p*80). In short, researcher needs to ensure that the treatment is the only factor that systematically influences the experimental variables [19].

3.2.1 Computer-Based Test and Paper-Pencil Test – The YBRAINS Test

The YBRAINS test is available in two modes: CBT and PPT. The test was adapted to a computer-based testing mode in 2009 [17]. Both the CBT and PPT deliver the same content. In this study, the two treatment groups answered the CBT while the two control groups answered the PPT.

The YBRAINS test (25 items) was used to collect data concerning brain styles (left brain, right brain of whole brain style) and openness styles (open thinking style, mixed thinking style or closed thinking style). Each item of the test provides participants with multiple choices – each choice representing a specialized function of the left brain, or a parallel function of the right brain. The participant was asked to indicate which of the specific brain functions best described his/her own typical behaviour. The responses were then calculated to obtain a brain style score (divide the total point scored by the number of responses made by the participant). The brain style score was then categorized into three thinking and learning styles based on a 9-point index which was then divided into three sections: *left brain style*: 1.0 – 4.5 points; *whole brain style*: 4.6 - 5.4 points; and *right brain style*: 5.5 – 9.0 points. For the openness levels, the scoring is similar to provide a three sections' openness styles: *open thinking style*: 1.0 – 4.5 points; *mixed thinking style*: 4.6 - 5.4 points; and *closed thinking style*: 5.5 –9.0 points (Fig. 2). The whole brain and mixed thinking styles are set at a narrow range of 4.6 - 5.4 point (a 10% of the 9-point scale) because for a person to be balanced in brain style, the maximum range of difference between the scores of the two hemispheres is estimated as 10% [17].

The three brain styles and thinking styles were then arranged in a two-dimensional matrix, to create a typology diagram (Fig. 2). The typology diagram presents a new model: the Brain Style Model. The YBRAINS test was developed in a computer-based system by using a visual basics programme. When a participant responds to the test items, his brain styles (left, right or whole brain style, and open, mixed or closed thinking style) will be presented instantly by the computer programme. Fig. **3** indicates an example item of the YBARINS test in CBT mode and its test scores in a graphical form. (*The computer-based YBRAINS test has won two gold medals in green technology innovation expos, including the 21st International Invention, Innovation & Technology Exhibition, 2010*)

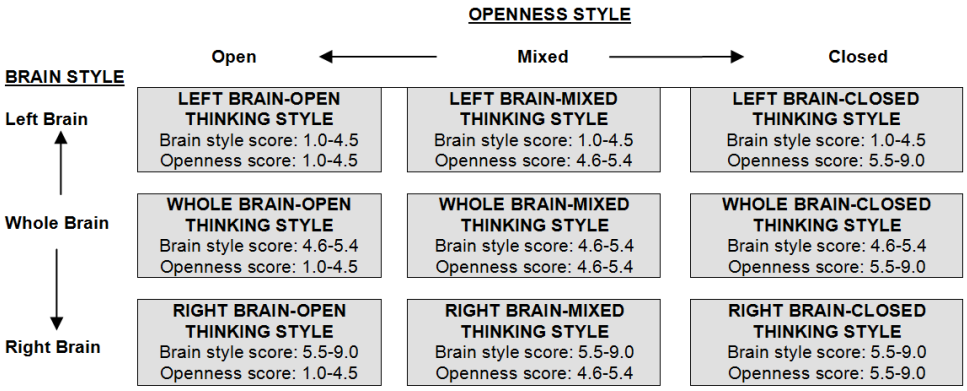

Fig. 2: Typology diagram of the brain style model generated from the YBRAINS test

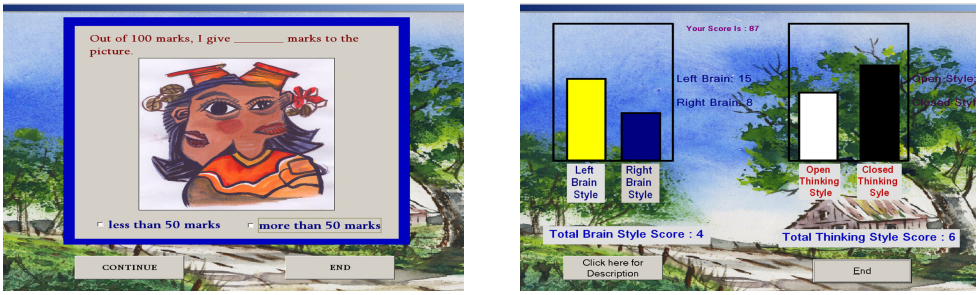

Fig. 3: An example test item and the results of the test in graphical form

Through the CBT test, the test scores (open leadership style, closed leadership style, creative thinking style and critical thinking style) and testing time were recorded in a Microsoft Access Database immediately after a participant had completed the test. As for the PPT mode, testing time was recorded manually immediately after a participant had answered the test. The test score for each participant was calculated manually by the researcher after the testing.

3.2.2 The Testing Motivation Questionnaire

Motivation is a universal human behaviour and is identical across disciplines. Most researchers agreed that motivation is a ''multifaceted set of goals and beliefs that guide behavior'' [20] (p.199). Wigfield and Guthrie set forth eleven dimensions of motivation [21] and Wigfield, Guthrie, and McGough developed a 54-item motivation questionnaire to examine eleven motivation dimensions of a group of students [22]. These eleven dimensions include: challenge, efficacy, curiosity, involvement, enjoyment, work avoidance, social, grades, competition, compliance, and importance. Although questions have been raised about the factor structure of the motivation dimensions [23], some studies examining it have supported these eleven dimensions [21], [24].

Parault and Williams reported that intrinsic motivation was associated with comprehension, and comprehension increased intrinsic motivation of a person [25], therefore, in this study, comprehension was also included as one of the testing motivation dimensions, given a total of twelve dimensions. Given the fact that these twelve dimensions are grounded in previous theoretical works in the field of motivation, this study examined all the twelve dimensions of testing motivation. Of these, six of the dimensions, i.e. challenge, efficacy, curiosity, involvement, enjoyment and comprehension were categorized under intrinsic motivation, while the other six dimensions, which are social, grades, work avoidance, competition, compliance, and importance, were included as extrinsic motivation [25]. The contents of the Testing Motivation Questionnaire (TMQ) and number of items for each testing dimension under intrinsic and extrinsic testing motivation are listed below.

Intrinsic motivation
Challenge: Based on the idea that a test taker can get satisfaction from mastering complex ideas presented in testing (5 items).
Efficacy: A person's belief in his capacity to organize and execute actions. Testing efficacy is the belief one has in his ability to be successful at testing (4 items).
Curiosity: The desire to learn about the contents of the test of personal interest (6 items).
Involvement: The willingness to participate in the test when one answers test items (6 items).
*Comprehension***:** The act or grasping the meaning, nature, or understanding of the test contents indicates the motivation level (4 items).
Joyfulness: The feeling and expressing deep happiness and enthusiasm for taking a test (5 items).

Extrinsic motivation
Competition: The desire to outperform others in the test (6 items).
Compliance: Testing due to an external goal or requirement (5 items).
Importance: One's sense that testing is of central importance to him (7 items).
Social: Sharing the knowledge of testing is frequently a social activity and it influences a person's desire to understand the test (6 items).
Work avoidance: Related to participants' dislike for aspects of the testing task and avoidance of the task (4 items).
Grades: Testing to receive a positive evaluation (3 items).

3.3 Participants of the Study

The participants in this study were 120 Malaysian student teachers from a teacher training institute which is located at the center zone of peninsular Malaysia. Among the participants, there were 41 males (34.17%) and 79 females (65.83%) with an average age of 19.1 years old. The participants were randomly selected from a student teachers' population (N=548) based on the sample size determination table of Krejcie & Morgan' [18] (p.211) at a 95% ($p<.05$) confidence level. They were enrolled in a teacher education programme. Participants have the same educational history and background. They have the same level of computer application skill and received formal computer instruction in their academic curriculum. To attain the requirements of the Solomon four groups experimental design, the participants were randomly assigned into four groups through a systematic random sampling procedure, each with a sample size of 30.

3.4 Procedures

The four participant groups were randomly assigned into two control and two treatment groups. At the beginning of the semester, the control group 2 answered a PPT mode of the YBRAINS test and the treatment group 2 answered the CBT mode of the test (see Fig. 1). Immediately after the tests, the two groups answered the TMQ (pretest). At the end of the semester (the posttest was administrated 14 weeks after the pretest to minimize memory effects), the four groups answered the posttest of the YBRAINS test. The two control groups answered the PPT mode while the two treatment groups answered the CBT mode. Then the four groups answered the TMQ (Posttest).

Hawthorn effect appears in an experimental study when the participants noticed that they are observed and try to provide abnormal or untruthful answers. It could appear when they noticed that unequal treatments are given to them (some participants are given treatment and some are not) [18] (p103]. Therefore, to reduce the Hawthorn effect, the four groups answered the tests and questionnaires in four separate rooms (Two treatment groups in two computer labs while two control groups in to classrooms).

4.0 RESULTS

4.1 Reliability of the Tests

The YBRAINS test and the TMQ are built on different measurement scales. Therefore, test-retest reliability was used to identify the reliability of the YBRAINS test while internal consistency reliability was employed to identify the reliability of the TMQ. Table 1 indicates the test-retest reliability coefficients for the CBT and PPT modes of the YBARINS. The significant Pearson product moment coefficients (CBT: .74 to .81; PPT: .65 to .76) show that both brain style and thinking style subscales' scores of the CBT and PPT are reliable.

Table 1: Test-retest reliability (Pearson product moment coefficients) of CBT and PPT

Test Score	Computer-Based Testing	Paper-Pencil Testing
Brain style	.77**	.67*
Left brain style	.81**	.65*
Right brain style	.75**	.69*
Thinking style	.76**	.74*
Creative style	.74*	.73*
Critical style	.79**	.76*

Note: *p< .05, **p< .01

To examine internal consistency (Cronbach's alpha) of each subscale of the TMQ, the responses of the treatment and control groups were examined. Table 2 indicates that for the treatment group, the coefficients were between .72 and 92, while for the control group, the alpha values were between .75 and .86. Based on the results, the twelve testing motivation dimensions were reliable for the treatment and control groups.

Table 2: Internal consistency reliability (Cronbach's alpha coefficient) of the testing motivation dimensions

Testing Motivation	Treatment Group	Control Group	Testing Motivation	Treatment Group	Control Group
Intrinsic Motivation	.84	.82	Extrinsic Motivation	.81	.75
Challenge	.82	.78	Competition	.77	.75
Efficacy	.76	.82	Compliance	.75	.81
Curiosity	.92	.78	Importance	.83	.79
Involvement	.81	.73	Social	.69	.68
Comprehension	.82	.86	Work avoidance	.81	.75
Joyfulness	.80	.82	Grade	.72	.75

4.2 Testing Effect

Testing effect for the PPT mode (control group) occurs if there is a significant difference between M5 and M3, while testing effect for the CBT mode (treatment group) occurs if there is a significant difference between M6 and M4. The data in Table 3 indicates that there was a testing effect on testing time [t(58)=2.21, p< .05] for the PPT mode. However, no testing effect was found for the CBT mode [t(58)=-.42, p> .05]. For test scores (Table 3), no testing effect was found for all of the brain style and thinking style subscales (p >.05). It indicates that for both the PPT and CBT modes, testing effect did not exist in test scores. It means that in this study, answering the pre test has no effect on the scores of the post test on all of the YBRAINS subscales.

As for the testing motivation, testing effects occurred in five of the twelve dimensions for the PPT mode. The dimensions were challenge [t(58)= -3.75, p=.00], curious [t(58)= -2.93, p=.01], involvement [t(58) =-3.05, p=.00], competition [t(58) =-2.07, p=.04], and work avoidance [t(58)= -2.49, p=.02]. As a whole, there was a testing effect in testing motivation for the PPT mode. However, no testing effect was found for intrinsic and extrinsic testing motivation and all of the testing motivation dimensions for the CBT mode, except for work avoidance. The results indicate that the CBT was free from testing effect except for the work avoidance dimension.

4.3 Treatment Effect

Treatment effect occurs if there is a significant interaction effect in the comparison between the treatment group 2 and the control group 2 (between-subject comparison) on their pretest and posttest scores (within-subject comparison). A treatment effect is detected if a significant interaction effect occurs. The interaction effect could be identified by a Split-Plot ANOVA analysis.

The results of Split-Plot ANOVA analysis (Multivariate Pillai's Trace) in Table 4 indicate that interaction effects occurred in testing time [F(1, 58) = 4.67, p<.05]. The mean scores indicate that CBT has effectively reduced the testing time of taking the posttest (CBT testing time mean score: pretest =14.37, posttest=13.30). On the other hand, no interaction effects were found in the brain style and thinking style subscales' scores.

Table 3: Analysis of testing effects for PPT and CBT modes on testing time, test scores and testing motivation

Variable	Testing Effect for PPT		Testing Effect for CBT	
	Mean difference (M5-M3)	**T Test**	**Mean difference (M6-M4)**	**T Test**
Testing time	1.17	t(58)=-2.21, *p*=.03*	-.17	t(58)=-.42, *p*=.68
Test score				
Brain style	.05	t(58)=.31, *p*=..75	.01	t(58)=-.05, *p*=.96
Left brain style	.00	t(58)=.00, *p*=1.00	.10	t(58)=-.19, *p*=.17
Right brain style	.63	t(58)=-.59, *p*=.56	.33	t(58)=-.56, *p*=.57
Thinking style	-.03	t(58)=-.08, *p*=.94	-.12	t(58)=-.54, *p*=.59
Creative style	-.33	t(58)=.53, *p*=..60	-.47	t(58)=.46, *p*=..65
Critical style	.53	t(58)=-.56, *p*=..58	.17	t(58)=-.14, *p*=..89
Testing motivation	-10.80	t(58)=-3.56, *p*=.00**	1.80	t(58)=-1.84, *p*=.28
Intrinsic Motivation	-7.80	t(58)=-3.90, *p*=.00**	.50	t(58)=-22, *p*=.83
Challenge	-2.57	t(58)=-3.75, *p*=.00**	.30	t(58)=-.46, *p*=.64
Efficacy	.37	t(58)= .79, *p*=.43	.10	t(58)=-.19, *p*=.84
Curiosity	-1.77	t(58)=-2.93, *p*=.01*	-.14	t(58)=.22, *p*=.82
Involvement	-1.83	t(58) =-3.05, *p*=.00**	.20	t(58)=-.25, *p*=.79
Comprehension	.67	t(58)=1.09, *p*=.27	.40	t(58)=-.45, *p*=.65
Joyfulness	.14	t(58)=1.50, *p*=.60	.37	t(58)=1.28, *p*=.20
Extrinsic Motivation	-3.00	t(58)=-1.79, *p*=.08	1.30	t(58)=-1.61, *p*=.11
Competition	-.87	t(58) -2.07=, *p*=.04*	.33	t(58)= -.47, *p*=.63
Compliance	-.37	t(58)=-.89, *p*=.38	.24	t(58)=-.52, *p*=.60
Importance	.20	t(58)=.28, *p*=.778	.20	t(58)=-.23, *p*=.81
Social reasons	-.43	t(58)= -.51, *p*=.610	.27	t(58)=-.30, *p*=.76
Work avoidance	1.47	t(58)=2.49, *p*=.02*	1.93	t(58)=-2.07, *p*=.04*
Grade	.40	t(58)=1.53, *p*=.13	.33	t(58)=-1.01, *p*=.06

Note: *p< .05, **p< .01; A significant t-test result indicates a testing effect of CBT or PPT on a subscale

Table 4: Split-plot ANOVA analysis for the effect of CBT on testing time, test scores and testing motivation

Variable	Control		Treatment		Pillai's Trace Test
	Pre	Post	Pre	Post	Interaction effect
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	(F-ratio value at df=1,58)
Testing time	13.40 (1.65)	13.23 (1.41)	14.37 (2.19)	13.30 (1.49)	4.67*
Test score					
Brain style	5.27 (1.30)	5.25 (1.14)	4.94 (.65)	4.99 (.57)	.68
Left brain style	11.03 (2.19)	10.93 (1.76)	9.57 (3.88)	9.57 (3.27)	.06
Right brain style	10.77 (3.77	11.40 (4.47)	10.70 (2.12)	11.03 (2.46)	.38
Thinking style	4.76 (.75)	4.63 (.92)	4.89 (1.28)	4.86 (1.44)	.06
Creative style	11.00 (3.16)	11.33 (2.89)	10.07 (3.16)	10.53 (4.60)	09
Critical style	10.23 (3.84)	9.70 (3.49)	9.97 (4.33)	9.80 (4.62)	.10
Testing motivation	133.67 (19.18)	132.03 (14.24)	145.10 (15.50)	173.43 (12.53)	61.76**
Intrinsic motivation	59.23 (12.34)	57.07 (9.14)	65.23 (9.97)	86.97 (8.25)	82.87**
Challenge	9.50 (3.03)	8.83 (2.84)	11.17 (2.52)	14.10 (2.55)	24.38**
Efficacy	7.26 (1.70)	7.66 (2.05)	8.20 (1.73)	12.96 (1.65)	90.10**
Curiosity	11.00 (2.76)	11.66 (3.02)	11.54 (2.49)	19.67 (2.41)	84.59**
Involvement	11.40 (3.28)	12.00 (2.22)	13.40 (3.41)	17.80 (2.94)	17.92**
Joyfulness	11.53 (1.53)	10.93 (1.60)	12.00 (1.46)	12.37 (.56)	4.17*
Comprehension	7.03 (2.20)	7.46 (2.28)	8.08 (3.04)	10.07 (3.16)	5.40*
Extrinsic motivation	74.73 (8.37)	74.97 (7.28)	79.87 (7.50)	86.47 (7.61)	10.63**
Competition	16.60 (2.58)	17.03 (2.24)	16.67 (2.48)	16.17 (2.79)	.4.00
Compliance	17.33 (1.14)	17.60 (1.52)	17.80 (1.09)	17.73 (1.59)	.53
Importance	12.20 (3.78)	13.87 (2.70)	14.70 (3.42)	15.03 (3.36)	1.92
Social	10.60 (3.41)	11.80 (2.94)	13.00 (2.76)	18.20 (3.33)	37.12**
Work avoidance	7.20 (2.44)	7.06 (2.27)	8.80 (3.04)	10.53 (3.60)	4.41*
Grade	8.30 (1.48)	8.32 (1.59)	8.34 (1.82)	8.48 (1.07)	.23

Note: *p< .05, **p< .01

The Multivariate Test's results in Table 4 indicate that as a whole, interaction effect occurred in testing motivation [$F(1, 58) = 61.76$, $p < .01$]. Besides that, the CBT effectively increased intrinsic motivation [$F(1, 58) = 82.87$, $p < .01$] and extrinsic motivation [$F(1, 58) = 10.63$, $p < .01$] of the test takers. The data also indicates that interaction effects occurred in eight of the twelve testing motivation dimensions. The dimensions were challenge, curiosity, efficacy, involvement, joyfulness, comprehension, social and work avoidance. The data in Table 4 clearly shows that for all the eight testing motivation dimensions, the posttest scores outperformed the pretest scores. It indicates that the CBT testing mode has significantly increased testing motivation of the participants, except the four extrinsic testing motivation dimensions, i.e. competition, compliance, importance and grade.

5.0 DISCUSSION

Results of the analyses indicate that both CBT and PPT of the YBRAINS are reliable for experimental research and no testing effects were found for test scores of the two testing modes. In other words, the YBRAINS score is stable and consistent over time. It shows that a participant who sits for the CBT and PPT would most probably yield a similar posttest score. However, testing effects occurred in testing time and testing motivation for the PPT mode. The finding is consistent with previous research results that the PPT mode has internal and external validity problems [14], [15]. For the PPT mode, taking a pretest significantly increased the testing time for the posttest (mean difference =1.17 minutes, see Table 3), while on the other hand, it significantly reduced the intrinsic motivation (mean difference value =-7.80) of the participants for taking the posttest. Besides that, the results show that the testing effect of the PPT mode reduced testing motivation in terms of challenge, curiosity, involvement and competition (negative mean difference values). In contrast, the testing effect increased work avoidance of the test taker (a positive mean difference value). This finding supports evidence of a previous study, that the dimensions of motivation were closely related, and there was a great deal of overlap between them, and that participants were not motivated by just one dimension but a combination of them [26].

Compared to the PPT mode, the findings showed that the CBT mode was more stable and consistent in terms of internal and external validity because no testing effects were found in intrinsic and extrinsic testing motivation, except for the work avoidance dimension. Work avoidance was defined as participants' dislike for aspects of the testing task and avoidance of the task [24]. The finding suggests that in establishing the CBT, the test items must be clearly stated and avoidance of complexity, that is, complex items should be excluded in the test, as this would help reduce work avoidance in test takers.

Another advantage of the CBT mode generated from the results is treatment effect occurred in testing time. CBT has effectively reduced the testing time. This finding supports the study of Bocij and Creasley. They reported that students performed faster in CBTs because they did not have to spend time writing down their responses [27]. Besides that, as a whole, there was a treatment effects on testing motivation. The results showed that the CBT has effectively increased intrinsic and extrinsic motivation of the test takers in challenge, curiosity, self-efficacy, involvement, joyfulness, comprehension and social dimensions. However, answering the test in a shorter time with higher testing motivation level did not help a test taker to achieve a higher score; no treatment effects were found in the brain style and thinking style subscales' scores. This is because for psychological tests, a test score reflects a test taker's behavior, and the scores are theoretically expected to be similar without the influence of testing modes.

In summary, the CBT mode is more reliable in terms of internal and external validity and no testing effect on test score was found in CBT mode. The CBT mode reduced testing time and increased testing motivation of the participants. An advantage of the CBT is increasing testing motivation would increase response rates [27].

The finding of this study is important because one of the four strategic objectives stated in the Malaysia' government's green ICT initiative towards sustainable environment (under the ICT Strategic Transformation 2010-2020 master plan) is to reduce paper consumption [28], *p*2. Hence, emphasis on the use of CBT in research and survey should be addressed to attain the government's green ICT initiative.

6.0 LIMITATIONS

The generalisability of the present study is limited by several factors; the study used a sample of undergraduate student teachers who are computer-literate. Secondly, the findings are limited by the psychological tests used in this study. The study would probably yield different results if the study used achievement tests. Thirdly, the sample size for each group is small. A larger sample size would have been more reliable. However, despite the small sub-sample sizes, the fact is that the study is the first of its kind examining testing effects with a Solomon four group design and reporting the testing effects in an experimental design. It is hoped that future researches will contribute by examining the testing effects and the association between CBT and other test takers variables.

REFERENCES

[1] Minnesota Pollution Control Agency, http://156.98.19.245/paper/, 2011.

[2] J. DeRosa, The Green PDF: Reducing Greenhouse Gas Emissions One Ream at a Time. http://www.scribd.com/doc/60779195/The-Green-PDF-Revolution, 2007.

[3] S. Friedrich & J. Björnsson, *The Transition to Computer-Based Assessment - New Approaches to Skills Assessment and Implications for Large-scale Testing*. *http://crell.jrc.it/RP/reporttransition.pdf, 2008*

[4] I, Choi, et al., "Comparability of a Paper-Based Language Test and a Computer-Based Language Test". *Language Testing,* Vol. 20 No 3, 2003, pp. 295-320.

[5] S. Al-Amri, "Computer-Based Testing vs. Paper-Based Testing: A Comprehensive Approach to Examining the Comparability of Testing Modes". *Essex Graduate Student Papers in Language & Linguistics*, Vol. 10, 2008, pp. 22-44.

[6] J. Boo, "*Computerized Versus Paper-and-Pencil Assessment of Educational Development: Score comparability and Examinee Preferences"*. Unpublished PhD Dissertation, University of Iowa, 1997.

[7] H. Wang & C. David Shin, "Comparability of Computerized Adaptive and Paper-Pencil Tests", *Test, Measurement and Research Service Bulletin*, Vol. 13, March 2010, pp. 1-7.

[8] N. Johnson & S. Green, "*On-Line Assessment: The Impact of Mode on Students Performance".* In the British Educational Research Association Annual Conference, Manchester, UK, 2004.

[9] J. C. Alderson, "*Technology in Testing: The Present and the Future*". *System*, Vol. 28 No 4, 2000, pp. 593-603.

[10] Y. Sawaki, "*Comparability of Conventional and Computerized Tests of Reading in a Second Language". Language Learning & Technology*, Vol. 5 No 2, 2001, pp. 38-59.

[11] J. H. Haahr and M. E. Hansen, "Adult Skills Assessment in Europe: Feasibility Study. Policy and Business Analysis, Final Report", November 2006.

[12] S. M. Bodmann & D. H. Robinson, "Speed and performance Differences among Computer-Based and Paper-Pencil Tests". *Journal of Educational Computing Research, Vol. 31,* 2004, pp. 51-60.

[13] C. Morgan & M. O'Reilly, Innovations in Online Assessment. In F. Lockwood & A. Gooley (Eds.), *Innovation in "Open and Distance Learning: Successful Development of Online and Web-Based Learning"* London: Kogan Page Limited. 2001, pp. 179-188.

[14] C. H. Yu & B. Ohlund "*Threats to validity of research design*". http://www.creative-wisdom.com/teaching/WBI/threat.shtml, 2010.

[15] D. Campbell, & J. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Rand-McNally, 1963.

[16] J. L. Solomon, "An Extension of Control Group Design". *Psychological Bulletin*, Vol. 46, 1949, pp. 137-150.

[17] Y. P. Chua, "Establishing a Brain Styles Test: The YBRAINS Test". *Procedia Social and Behavioral Sciences* Vol. 15, 2011, 4019-4027. http://www.sciencedirect.com/science/article/pii/S1877042811009530

[18] Y. P. Chua, Research Methods and Statistics (2nd edition). McGraw-Hill Education, 2011.

[19] Y. P. Chua, Research Methods and Statistics. McGraw-Hill Education, 2008.

[20] J. T. Guthrie & A. "Wigfield How Motivation Fits into a Science of Reading". *Scientific Studies of Reading*, Vol. 3, 1999, pp. 199–205.

[21] A. Wigfield & J. T. Guthrie, "Relations of Children's Motivation for Reading to the Amount and Breath of Their Reading". *Journal of Educational Psychology*, Vol. 89, 1997, pp. 420–432.

[22] A. Wigfield, J. T. Guthrie & K. McGough, "A Questionnaire Measure of Children's Motivations for Reading". *ERIC Document Reproduction Service* No. ED394137, 1996.

[23] M. W. Watkins & D. Y. Coffey, "Reading Motivation: Multidimensional and Indeterminate". *Journal of Educational Psychology,* Vol. 96, 2004, pp.110–118.

[24] L. Baker & A. Wigfield, "Dimensions of Children's Motivation for Reading and Their Relations to Reading Activity and Reading Achievement". *Reading Research Quarterly*, Vol. 34, 1999, pp. 452–477.

[25] J. S. Parault & H., M. Williams, "Reading Motivation, Reading Amount, and Text Comprehension in Deaf and Hearing Adults". *Oxford University Press*. doi:10.1093/deafed/enp031, 2009.

[26] A. Wigfield, “Reading Motivation: A Domain-Specific Approach to Motivation”. *Educational Psychologist*, Vol. 32, 1997, pp. 59–68.

[27] C. Bocij & A. Greasley, “Can computer-based testing achieve quality and efficiency in assessment?” *International Journal of Educational Technology, 1,* http://www.ao.uiuc.edu/ijet/v1n1/bocij/index.html, 1999.

[28] MAMPU. Malaysia’ government’s green ICT initiative: Towards sustainable environment, http://www.greenit-pc.jp/activity/asia/file/malaysia2.pdf, 2011.

BIOGRAPHY

Dr Chua Yan Piaw is an associate professor at the University of Malaya. He has written 35 books, including academic books for secondary schools and higher education. The subjects of his writing include research methodology and statistics, creative and critical thinking skills, calligraphy, art education, science and chemistry. He has published papers in ISI journals, including the Journal of Scholarly Publishing, International of Journal of Market Research and Education and Urban Society.