International Journal of English Linguistics; Vol. 8, No. 2; 2018
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Comparative Study on Validity of Paper-Based Test and Computer-Based Test in the Context of Educational and Psychological Assessment among Arab Students

Badia Muntazir Hakim[1]

[1] English Language Institute, King Abdulaziz University, Jeddah, Saudi Arabia

Correspondence: Badia Muntazir Hakim, English Language Institute, King Abdulaziz University, Jeddah, Saudi Arabia. E-mail: badiahakim82@gmail.com

Received: November 17, 2017 Accepted: December 2, 2017 Online Published: December 23, 2017

doi:10.5539/ijel.v8n2p85 URL: http://doi.org/10.5539/ijel.v8n2p85

Abstract

A prerequisite to evolving the concept of computerized assessment method requires a comparable test scores data based on the comparative study scores of both paper-based test (PBT) and computer-based test (CBT) methods. Previous research studies reported that even though both testing types show significant commonalities in many areas of technical and academic concerns, still there are substantial variances found in test scores. Consequently, in various educational assessment systems this significant inconsistency raises serious question on the rationale of substituting PBT with CBT. This is imperative to compare the both testing modes and to provide a concrete base for this replacement. This research study was aimed at implementing an achievement test and a motivation level checking questionnaire in order to test the effectiveness and validity of the CBT and to measure the level of student motivation that directly controls the performance and results. Researcher aims to get some valid data to provide a solid base for the effective use of CBT in academic and placement modes. Results showed that (a) the pretests improved the results by providing experience for the tests themselves (b) participants in CBT group showed better test performance. In fact, CBT is known to be an efficient tool for assessment.

Keywords: CBT (computer-based test), PBT (paper-based test)

1. Introduction

Recently there has been a trend in shifting the mode of assessment from PBT to CBT in the majority of educational institutions and this shift requires institutions to revise and renew their curricula methodology and pedagogy while taking the necessary practical changes for the successful implementation of this new assessment methodology into account (Chen, 2012; Genc, 2012; Hsiao, Tu, & Chung, 2012; OECD, 2010). Since it requires changes not only in the assessment methodology but also in the curricula methodology and pedagogy in educational institutions, substituting PBT with CBT is a major shift which raises serious questions regarding the rationale behind this rather radical change.

In this study, it is proposed that CBT is more beneficial than PBT as a future testing tool for the test-takers. This conclusion was drawn in the light of a comparable test scores data gathered by a four-group experimental study on the comparative study scores of both PBT and CBT methods.

To illustrate the scope of this study, we will start by reviewing the current literature regarding the differences in PBT and CBT testing modes in terms of pedagogical advantages and test performance. Then we will move on to the review of discussions in the present literature related to the testing-effect of repeated-measure plan as a possible shadowing variable which effects the influence of CBT treatment effect on test performance. From that point, the prepositions made about the impact of motivation on testing mode and test performance will be reviewed. The research methodology section will shed a light on the design of the study explaining the methodology for the experimental study and the structure of the assessment motivation questionnaire. Finally, the reader will be presented with the results of the study and the conclusions drawn by those results.

2. Literature Review

Use of CBT as a summative assessment tool carries concrete practical and economic benefits because it provides facility to test an immense number of student cohort with the facility of automated marking of

responses (Charman, 1999; Zakrzewski & Bull, 1998). CBT is a mode of testing that acts as a catalyst for change, provides a base for change in mode of learning, instruction and curricula in educational institutions (Scheuermann & Pereira, 2008). Recently in most of the educational institutions there is a recent trend in shifting the mode of assessment from PBT to CBT. Administering the CBT mode of assessments is becoming predominantly widespread in educational assessment domain because this major variation in assessment methodology leads to practical changes in pedagogy and curricula methodology (Chen, 2012; Genc, 2012; Hsiao, Tu, & Chung, 2012; OECD, 2010). Pedagogical advantages on CBT include: providing a fast and error free feedback; repeatability of tests consisting of randomly-generated test items; unquestionable reliability and fairness; flexibility in allocation of test timing and venue; and, direct responsibility for one's own learning and test taking (Charman, 1999). There is a clear policy statement by the International Guidelines on Computer-Based Testing (International Test Commission, 2006) that in order to administer a valid and reliable CBT it is imperative that corresponding test scores should be established for the conventional paper-based testing (PBT) and its corresponding computer-based method. There has been a strong support base provided to this set of testing standards by the classical true-score test theory—the basis of computer-based and paper-based testing (Allen & Yen, 1979). As per propose theory by Allen & Yen (1979); anyone who takes the same test in the above mentioned two modes (CBT and PBT), it is anticipated that the test taker obtains almost the matching level of test scores. The same idea and theory has also been supported by the empirical studies by OECD (2010); Wilson, Genco, & Yager (1985). In their related study OECD (2010) stated that no major discriminations have been found in the mode of test performance between CBT and PBT. Their findings were based on the data collected from the student participants (n = 5,878) from Denmark, Iceland and Korea.

The related notion of correspondent results both is PBT and CBT was also reinforced by many studies in certain specific subject areas, and the clear discrimination of results was established in achievement tests such as science, language and mathematics, and also the same was very perceptibly ascertained by a chain of psychological tests such as personality and neuropsychological assessment (e.g., Friedrich & Bjornsson, 2008; Choi, Kim, & Boo, 2003; DeAngelis, 2000). In their findings about the review of educational and psychological measurement approaches, Bunderson, Inouye & Olsen (1989) have established that 48% of previous studies revealed negligible difference between the two testing modes (PBT & CBT) in the area of test performance, whereas 13% of studies have reported that CBT test performance is better than PBT and 39% of findings have proved PBT better than CBT.

There could be a straightforward elucidation for this above mentioned peculiar difference of test performance, either the proposed CBT possesses a weak validity as an assessment tool for educational assessments and the related psychological mode, or there might have been various other factors that shadowed the positive impact of CBT mode on test performance as per applied repeated-measures study pattern. In their parallel study, as established by Yu & Ohlund (2010), a possible shadowing variable is testing effect; according to that process of having a pretest preceded by the posttest analytically confuses the treatment effect of CBT on test performance.

2.1 Testing Effect with Repeated-Measure Plan

From the related reviewed studies, this has been clearly and carefully established that frequently conducted comparative account studies are based on the conventional experimental pretest-posttest pattern and these studies have been conducted without considering and classifying the effects of variable repeated-measure test patterns on the test takers. Consequently, this may lead to establishing the findings based on misinterpretations. This can be supported by a theoretical approach established by Al-Amri (2008), according to his study a student took the same test twice as a pretest and twice as a posttest, so the reliability could be shadowed by certain factors. According to this concept of repeated-measure plan, the limitations of this pattern might compromise the testing effect because the same test taker is exposed to each test twice and also the outcome of posttest might be compromised by taking a pretest (Yu & Ohlund, 2010). This literature review finding very strongly confirms that using this pattern to strongly conclude any new idea is a biased approach. In my context, pretest has been considered as a short orientation test before administering the main test (CBT&PBT)

2.2 Impact of Motivation on Testing Mode and Test Performance

Related literature review also brought attention towards another important factor of motivation and its impact on test performance and testing mode. As reported by Wise & DeMars (2005), this needs to be discussed and clarified that motivational factors have a strong impact on the test performance especially while conducting the comparative study of a PBT and CBT mode of testing. They elucidated very clearly that irrespective of accurate application of psychometric care to test development, or even after maintaining the parallel pattern for two

testing modes, if the motivational level of the test takers is low (e.g., due to less preparation or boredom), testing validity will be compromised. The self-determination theory by Wenemark, Persson, Brage, Svensson, & Kristenson (2011) determines the fact that enhanced motivation level of test-takers has a direct impact on the positive response to the testing mode and increases students' interest towards learning process. Consequently, it is proven that increased motivation for testing mode is an important and sensitive point to be investigated in testing mode comparative studies. This aspect might hinder the test validity while inferring the assessment scores from variable testing modes.

Pintrich (1989) presented the test taker motivation model that strongly reflects that the level of test takers' willingness is directly related to certain factors; level of their comfort towards the test taking mode, level of preparation, students' perception about the testing mode, and their reaction towards the level of test complicacy. Noteworthy is the point that the said model is a theoretical approach that relates the motivation, testing mode and test performance.

This has been established by Bugbee (1996) that in any educational institution the strongest hindrance to the successful administration of CBT with its complete psychological measurement and support is lack of sufficient study for the equivalence of CBT and PBT.

3. Method

3.1 Research Design

In this study plan, Solomon four-group experimental design was used with some modifications. It consists of two basic categories of participants:

(1) two groups of participants who were given CBT and two groups of participants who were given PBT and (2) two groups of participants who were given the pretest and two groups of participants who were not given the pretest. The groups of participants are illustrated in Figure 1. This design helps to compare the basic two-group pretest and Target treatment test (CBT & PBT) and it also facilitates to identify the testing effect besides the treatment effects on experimental variables. The main aim to use this design is to help the researcher in finding the strong testing effect as compared to the experimental variable and also to measure the treatment effect with the presence or absence of pretest in both cases (CBT & PBT).

The results of A_1-A_2 will be compared for testing effects of PBT with or without pretest and B_1-B_2 for testing effects of CBT with or without pretest (see Figure 1). If there are no or negligible differences between the values of A_1 and A_2 as well as B1 and B2 it means that there are no testing effects. Consequently, the comparison data between $(A_1–B_1)–(A_2–B_2)$ will give an estimation of the treatment effect. However, any difference between the values of $(A_1–B_1)$ & $(A_2–B_2)$ will be considered due to the pretest effects. In such a scenario, in order to correlate the experimental variables (test performance and testing motivation to the target treatment CBT, researcher used a posttest motivational check questionnaire. In such a case, it is important for the researcher to establish that the target treatment has an effect on the experimental variables (test performance and testing motivation) because there is a possibility that the changes in the experiment variables are due to testing effects, and not by the treatment effects.

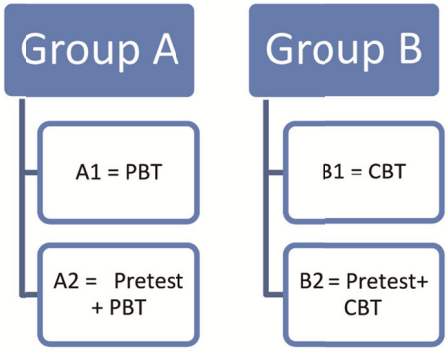

Figure 1. Illustration of the comparison groups

In order to analyze the obtained data for the design, one type of method was used:

For two independent samples z-test is performed to identify the testing effects (A_1- A_2) or (B_1-B_2)

3.2 Research Instruments

There were 4 research instruments used for the current study.

1) An orientation sample test with a brief outline of the main test
2) An achievement test—PBT version
3) The same achievement test—CBT version
4) The Assessment Motivation Questionnaire

3.2.1 Structure of the Achievement Test

A comprehensive achievement test was developed by the author based on the patterns which were officially determined and used by the university where the study was conducted. The achievement test consisted of items targeting the pre-intermediate English language proficiency level, with four sections and a total of 60 questions:

a) Section A—Listening 16 questions
b) Section B—two reading passages with 16 questions
c) Section C—Grammar 16 questions
d) Section D—Vocabulary 10 questions

3.2.2 Structure of the Assessment Motivation Questionnaire

Another important study instrument was a kind of modified version of the Assessment Motivation Questionnaire or AMQ (Wigfield, Guthrie & McGough, 1996). This tool helps in assessing the testing motivation and four motivation components self-efficacy, extrinsic, intrinsic and social motivation. It helps to compare the participants' motivation level towards the two testing methods. This questionnaire has three main sections;

a) Section for Self-efficacy motivation covers points of challenge and efficacy
b) Section for Intrinsic motivation covers points of curiosity, involvement, importance and work avoidance
c) Section for extrinsic motivation includes points for competition, recognition and grades.
d) Section for social motivation covers points for social and compliance

The above-mentioned extensions of motivational approach were modified and used for the current research to measure the motivational level for two separate testing modes.

3.3 Participants

The participants in this study were 200 Saudi female foundation year students from the English Language Institute. These students with an average age between 18-21 years were all from the pre-intermediate English language proficiency level, studying the textbook "English Headway Plus" by Oxford University Press. For the implementation of the study plan, their consent to participate in this study was taken. Mainly they were divided into two groups A and B. Participants in Group A were given PBT in two subgroups (A_1 = PBT and A_2 = PBT with pretest orientation). All the participants in Group B were given CBT (B_1 = CBT and B_2 =CBT with Pretest orientation). T was ensured that all participants in Groups B had the same level of computer applications skill and received formal computer instructions. The four groups were then randomly subjected to the treatment and experimental study.

3.4 Procedure

Initially 50 participants in each group A and B were randomly sorted into two subgroups A_1, A_2, B_1, B_2 with 25 participants in each.

The participants in Group A_1 and B_1 (N_1=50) were subjected to direct PBT and CBT without any pretest orientation. The remaining participants in A_2 and B_2 (N_2=50) were initially subjected to a pretest orientation and then with a gap of 24 hours they were given the actual PBT and CBT. It was ensured that participants in both main groups A and B would not be in contact with each other. In order to identify their motivation towards the two testing methods the Assessment Motivation Questionnaire (AMQ) was administered to all participants immediately after the administration of PBT and CBT.

4. Results

The present study used data from a research conducted in the English Language Institute of the King Abdul Aziz University, Saudi Arabia.

The research team collected data from the test results of 200 students. The students were randomly selected to

make the testing fair.

The participants of this study were divided into 4 groups of 50 students including 2 groups with pretest for PBT and CBT and 2 groups without pretest for PBT and CBT.

The descriptive information shows the means, standard deviations, minimum and maximum values for all of the four variables.

Table 1. Descriptive statistics of the variables

Sample	N	Mean	SD	Min	Max
PBT	50	35.74	7.9	27.84	43.64
PBT + Pretest	50	39.5	7.2	32.3	46.7
CBT	50	48.58	6.23	42.35	54.81
CBT + Pretest	50	52.44	4.7	47.74	57.14

The values of mean and standard deviation for PBT are 35.74 +_ 7.9 which gives us a minimum value of 27.84 and a maximum value of 46.64. The values of mean and standard deviation for PBT+ pretest are 39.5 +_ 7.2 which gives us a minimum value of 32.3 and a maximum value of 46.7. As per the results of Z statistics (Z =1.645, Sig = 0.005) both group of students (PBT and PBT + Pretest) are showing statistically different results. Since our calculated value does not fall in the critical region so we reject our null hypothesis and accept our alternative hypothesis and conclude that there is a difference between PBT and PBT+ Pretest. The statistical conclusion claims that there is a significant difference between PBT and PBT+ Pretest and they are not equivalent. Their results are not identical.

The values of mean and standard deviation for CBT are 48.58 +_ 6.23 which gives us a minimum value of 42.35 and a maximum value of 54.81. The values of mean and standard deviation for CBT+ pretest are 52.44 +_ 4.7 which gives us a minimum value of 47.74 and a maximum value of 57.14. As per the results of Z statistics (Z =1.645, Sig = 0.000) both group of students (CBT and CBT + pretest) are showing statistically different results. Since our calculated value does not fall in the critical region so we reject our null hypothesis and accept our alternative hypothesis and conclude that there is a difference between CBT and CBT + Pretest. The statistical conclusion claims that there is a significant difference between CBT and CBT + Pretest and is not equivalent. Their results are not identical.

For motivational questionnaire, almost 70% students have shown more self-efficacy for the CBT with a pretest. Around 80% students are intrinsically motivated with the CBT with a pretest. The rate of extrinsic motivation for CBT with a pretest is around 85%. Around 66% students are socially motivated by the CBT with a pretest.

5. General Discussion and Conclusion

In this research study, Solomon four-group experimental design was used with some modifications in order to compare the test performance of participants in CBT and PBT. Our results show that there is a significant difference between the results of PBT(A1) (mean = 35.74, SD = 7.9) and PBT + pretest (A2) (mean = 39.5, S.D = 7.2). Also results of CBT show that there is a significant difference between the results of CBT(B1) (mean = 48.58, SD = 6.23) and CBT + pretest(B2) (mean = 52.44, S.D = 4.7).

Firstly, a comparison was made between the test performances showed in PBT test mode without a pretest and in PBT test mode with a pretest. After comparing the mean and standard deviation (SD) values of A1 and A2 for paper based tests we analyzed that students showed slightly better performance with pretest. Since most students come from a PBT background, this might account for the slight difference. Also, as Yu & Ohlund (2010) suggests testing effect can be a possible shadowing variable here, since having a pretest preceded by the posttest analytically confuses the treatment effect on test performance. Having the possible shadowing variables on the difference observed in mind, we accept our alternate hypothesis that results of PBT and PBT + pretest are different as well as PBT + pretest gives improved performance of students.

The second comparison was made between the test performances showed in CBT test without a pretest and in CBT test mode with a pretest. After comparing the mean and standard deviation values of B1 and B2 for computer based tests, we analyzed that students showed slight better performance with pretest because computer based pretest gives students an idea regarding how to do CBT, so the maximum and minimum are better than the results of B1 group. Here again, we keep Yu & Ohlund's (2010) suggestion of testing effect as a shadowing variable influencing test performance, and we accept our alternate hypothesis that results of CBT and CBT +

pretest is different as well as CBT + pretest gives improved performance of students.

When the mean and standard deviation of A1 was compared with the mean and standard deviation of B1, better results were found for B1(CBT). Whereas, when the mean and standard deviation of A2 was compared with the mean and standard deviation of B2, once again better results were seen in B2 (CBT + pretest).

Motivational questionnaire results also show that students are intrinsically and extrinsically more motivated by the CBT with a pretest.

Allen & Yen (1979) claims that anyone who takes the same test in CBT and PBT modes should obtain almost the same level of test scores. However, in contradiction with Allen & Yen's (1979) suggestion, an overall comparison of all groups including A1, A2, B1, & B2 shows the results of CBT to be significantly better than PBT. From this study, it could be concluded that CBT is more beneficial for the respondents as a future testing tool. Bunderson, Inouye, & Olsen's (1989) suggest that 48% of previous studies revealed negligible difference between the two testing modes in question, which means in 52% of the previous studies there a significant difference was found, and this study also finds its place in that 52% majority.

It can also be concluded from this study that the pretests can improve the results by providing experience for the tests themselves. While interpreting and implementing the results of the study it should be noted that our current generation is computer oriented and they feel convenient and comfortable in CBT hence the results of CBT (B1) and CBT + pretest (B2) are much better than PBT (A1) and PBT + pretest (A2).

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 & Linguistics, 10*, 22-44.

Allen, M., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, California: Brooks.

Bugbee Jr, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*(3), 282-299. https://doi.org/10.1080/08886504.1996.10782166

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series, 1988*(1). https://doi.org/10.1002/j.2330-8516.1988.tb00291.x

Charman, D. (1999). Issues and impacts of using computer-based assessments (CBAs) for formative assessment. In S. Brown, J. Bull, & P. Race (Eds.), *Computer-Assisted Assessment in Higher Education* (pp. 85-94).

Chen, K. T. C. (2012). Elementary EFL teachers' computer phobia and computer self-efficacy in Taiwan. *TOJET: The Turkish Online Journal of Educational Technology, 11*(2).

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295-320. https://doi.org/10.1191/0265532203lt258oa

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

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

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 assessment-new approaches to skills assessment and implications for large-scale testing.

Genc, H. (2012). An Evaluation Study of a CALL Application: With BELT or without BELT. *Turkish Online Journal of Educational Technology-TOJET, 11*(2), 44-54.

Hsiao, H. C., Tu, Y. L., & Chung, H. N. (2012). *Perceived social supports, TOJET: The Turkish Online Journal of Educational Technology, 11*(2).

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*(2), 143-171. https://doi.org/10.1207/s15327574ijt0602_4

OECD. (2010). PISA Computer-based assessment of student skills in science. https://doi.org/10.1787/9789264082038-en

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. *Advances in Motivation and Achievement, 6*, 117-160.

Scheuermann, F., & Pereira, A. G. (2008). Towards a research agenda on Computer-based Assessment.

Challenges and needs for European Educational Measurement. Luxembourg.

Wenemark, M., Persson, A., Noorlind Brage, H., Svensson, T., & Kristenson, M. (2011). Applying motivation theory to achieve increased respondent satisfaction, response rate and data quality in a self-administered survey. *Journal of Official Statistics, 27*(2), 393-414.

Wigfield, A., Guthrie, J. T., & McGough, K. (1996). A questionnaire measure of children's motivations for reading. (Instructional Resource No. 22), 1996 Athens, GA National Reading Research Center, Universities of Georgia and Maryland, College Park.

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*(3), 265-275. https://doi.org/10.1016/0747-5632(85)90017-2

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

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

Zakrzewski, S., & Bull, J. (1998). The mass implementation and evaluation of computer-based assessments. *Assessment & Evaluation in Higher Education, 23*(2), 141-152. https://doi.org/10.1080/0260293980230203