The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer-Based Testing

Heri Retnawati
Mathematics and Science Faculty, Yogyakarta State University, Indonesia
heri_retnawati@uny.ac.id

ABSTRACT
This study aimed to compare the accuracy of the test scores as results of Test of English Proficiency (TOEP) based on paper and pencil test (PPT) versus computer-based test (CBT). Using the participants' responses to the PPT documented from 2008-2010 and data of CBT TOEP documented in 2013-2014 on the sets of 1A, 2A, and 3A for the Listening and Reading section, the researcher estimated the reliability estimation results using classical test theory and the value of information function and on the item response theory on PPT are then compared with CBT, which has the greater reliability and the value of information functions is said to be more accurate. The study shows that with the classical test theory approach, the reliability coefficients between the scores of the results of PPT and those of CBT are almost the same, and using the item response theory, it was found that although the value of the information function on PPT and CBT relatively similar in several subtests, there is a tendency for participants with the moderate ability that CBT is more accurate than PPT, and for the low and high ability of participants, PPT tends to be more accurate than CBT.
Keywords: accuracy, reliability, value of information function (VIF, paper and pencil test (PPT), computer based test (CBT)

INTRODUCTION

Nowadays, the development of science and technology is advancing. This has an impact on life, including on education. The presence of technology in education is used to assist and improve the quality of learning (Woolfolk, 2007). More specifically, this technology can be utilized in educational assessment, namely the implementation of the test. The utilization of technology in educational assessment is aimed at the effectiveness and efficiency of the implementation of the test (Chee and Wong, 2003; Towndrow & Vallence, 2004).

At first, the test for assessment which is popular is paper and pencil test. Along with the development of the Internet and intranet networks, the access to information inside and outside the school becomes easier. This tool can also be used for other purposes, for example, for the examination based on computer and the Internet, for example, known as Computer Based Testing (CBT). However, the utilization of computers for CBT has not been optimal yet in Indonesia in various tests.

In 2007-2010, the association of Teachers of English as a Foreign Language in Indonesia (TEFLIN) in collaboration with the Directorate General of Senior High school developed a test to measure English competency, later called as the Test of English Proficiency (TOEP). TOEP was developed based on constructs should be measured in a language test that is often referred to as communicative competence (Bachman, 1990; Bachman & Palmer, 1996). The development of the items of TOEP is based on the taxonomy in language proficiency from Munby (1983) which has identified micro-language proficiency skills which include listening, speaking, reading and writing. From 2006 to 2010, eight sets were developed for PPT to TOEP which have proved equivalent (Retnawati, 2014a). Subsequently in 2012, funded by the Department of Higher Education, the CBT system of TOEP was pioneered in using the sets used in PPT, which was then implemented from 2013 on Listening and Reading subtests.

On the preliminary study, there were some technical problems faced in the implementation of TOEP based CBT. The test participants were not familiar with the implementation of the CBT, and they often did the tests based on PPT. The difficulties in the implementation of CBT included the difficulty to log in, use a headset for listening, use the mouse to answer, and in some areas there was a problem about the availability of electricity and the access to Internet. These constraints led to the testees' doubt of the accuracy of the results of the test, moreover, there were several sets of TOEP being used in the administration of the test. Related to the above, the present study investigated the comparison the accuracy of TOEP scores of the PPT and CBT.

There are many advantages and disadvantages of using computer assessment compared with paper based task (Noyes & Garland, 2008). The advantages of online assements are (1) the richness of the interface, for example, the use of graphics allows a dynamic presentation of the test content, (2) the user population, computer-based

testing via the internet allows a more diverse sample to be located, (3) standardisation of test environment, that is, the test is presented in the same way and in the same format for a specified time, (4) online scoring, this results in faster feedback and (5) greater accuracy, that is, reduction in human error. In the other hand, the disadvantages of using computer in assessment are (1) lack of a controlled environment with responses being made at various times and settings and perhaps not even by the designated individual, double submissions may also be a problem, (2) computer hardware and software, these may be subject to freezing and crashing; in the test setting, time can be wasted when computers have to be restarted or changed, (3) the computer screen, for longer tests, it may be more tiring to work on the computer than on paper, (4) serial presentation, it is difficult to attain equivalence with computer and paper presentation, (5) confidentiality.

In its developmental process, TOEP considers the item difficulty by using both the classical test theory and the Rasch model of the modern test theory. Accordingly, the accuracy of CBT and PPT is determined by using two theoritical approaches, the classical test theory and item response theory.

The accuracy in the classical test theory is determined by the value of the standard error of measurement (SEM). SEM is estimated in the following formula

$$\sigma_E = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (1)$$

where σ_x is standard deviation of total score and $\rho_{xx'}$ is the reliability coefficient (Allen & Yen, 1979; Crocker & Algina, 1986). The formula shows that the higher the reliability coefficient, the smaller the SEM, and vice versa. The reliability coefficient can be estimated, by the formula such as Cronbach's alpha (Ebel and Frisbie, 1991; Reynolds, Livingstone, Willson, 2010).

In the modern approach, a well-known formula in the measurement involving the level of difficulty commonly is called the Rasch model (Hambleton, Swaminathan, and Rogers, 1991). The model of the relationship between chance to answer correctly (P), ability scale (θ) and item difficulty to-i (b_i), e natural number, and n items in the test is expressed in the following equation:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad \text{, where i : 1,2,3, …,n} \qquad (2)$$

The b_i parameter is a point on a scale of abilities in order to the probability a testee respond properly is 50%. Suppose a test item has a parameter b_i = 0.4. This means that the minimum ability required to have 50% probability to answer correctly is 0.4 on a ability scale. The greater the value of the parameter bi, the greater the ability needed to answer correctly with a 50% probability. In other words, the greater the value of the parameters b_i, the more difficult the item is.

In the item response theory, there is the value of information function. The information function item is a method to describe the strength of an item on the test, the selection of items, and the comparison of several sets of test. The item information function expresses the strength or contribution of test items in uncovering latent trait measured by those tests. If I is an information function, $P_i(\theta)$ is the probability to answer correctly for participants θ with the ability to answer correctly point I, $Q_i(\theta)$ opportunities θ participants with the ability to answer one item I, mathematically, the item information function satisfies the following equation.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad (3)$$

The test information function is a function of the number of items contructing the test information (Hambleton and Swaminathan, 1985: 94). Associated with this statement, the function of test information will be high if the items of the test have information function which is also high. The test information function can mathematically be expressed as follows.

$$I_i(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (4)$$

The difficulty index of item parameter and ability parameter of participants are estimated. Because these are the result of estimation, the nature of true parameters is probability and it is not free from by measurement error. In the item response theory, the standard error of measurement (SEM) is closely related to the function information. SEM has an inverse quadratic relationship with information function, the greater the value of information

function, the smaller SEM or otherwise (Hambleton, Swaminathan, and Rogers, 1991, 94; Retnawati, 2014b). If the value of the function information is represented by $I_i(\theta)$ and the estimated value of SEM is revealed by SEM (θ), then the relationship between the two, according to Hambleton, Swaminathan, and Rogers (1991: 94) is expressed by

$$\text{SEM } (\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \qquad (5)$$

De Gruijter & Van der Camp (2005: 118) statet that the value of the function information item and also the value of the test information function, depend on the latent ability. The value of item information is invariant, so that the ratio of the value of the two items' information functions is also invariant. The ratio of the information value of the two items is stated as follows:

$$\frac{I_{i_1}(\theta^*)}{I_{i_2}(\theta^*)} = \frac{I_{i_1}(\theta)}{I_{i_2}(\theta)} \qquad (6)$$

for all the transformations of θ * of θ. The invariant properties of the ratio of the value information function is used to determine the relative efficiency of the test. The relative efficiency of the two tests is defined as the ratio of the variance mistakes or, equivalently, the ratio of the value of the information function (McDonald, 1999: 279). This value can be compared when two tests measure the same attributes. The same thing is done by Lord (1980: 83) and also Stocking (1999), but it has a different symbol. Conventionally, the relative efficiency of test A and test B is written as:

$$\text{ER(f,A,B)} = \frac{I_A(\theta)}{I_B(\theta)} \qquad (7)$$

so if the ratio is less than one, then test A is said to be less efficient providing less information, or equivalently have a larger error in measurement compared with test B. The comparison of information value of both tests is used to compare the score of the PPT and CBT on TOEP.

The comparison of the administration of PPT and CBT has been investigated by many researchers. Al-Amri (2007) explored the comparison of paper and computer-based testing in reading context and the impact of test takers' characteristics. The results are there are no significant differences between paper and computer-based testing in reading context. Jamil, Tariq, & Shami (2012) reported teachers' perceptions of computer-based (CB) vs. paper-based (PB) examinations. The results showed that overall sampled teachers' attitudes were positive towards CB examination systems but in some situations they preferred PB. Comparatively for female participants had highly ranked, highly qualified, less experienced, teachers who have computer training certificate or degree, and teachers who have CB examination experiences were more positive towards CB examinations.

The comparison of the administration of PPT-based and CBT has been studied by many experts. Al-Amri (2007) tapped the comparison of paper and computer-based testing in reading context and the impact of test takers' characteristics. The results are there are no significant differences between paper and computer-based testing in reading context. Jamil, Tariq, and Shami (2012) reported teachers' perceptions of computer-based (CB) v. paper-based (PB) examinations. The results showed that overall the sampled teachers' attitudes were positive towards CB examination systems but in some situations they preferred the New Testament as well. Comparatively female, highly ranked, highly qualified, less experienced, teachers who have computer training certificate or degree, and teachers who have CB examination experiences were more positive towards CB examinations.

Maguire, K.A., Smith, D.A., Brailler, S.A. (2010) examined the difference in test scores for students who engaged in proctored course assessments electronically via computer interface compared to students who took proctored assessments through a paper and pencil format in the classroom. The results indicated that students who completed all assessments electronically scored significantly higher than those students completing all assessments via pencil and paper. No interaction was present between test format and test number, suggesting that none of test format had a more severe learning curve. The findings of this study, taken into conjunction with those of previous studies, suggest that proctored CBT provides an accurate assessment of a student's abilities.

Coniam (2006) describes an English language listening test intended as computer-based testing material for secondary school students in Hong Kong, Test takers generally performed better on the computer-based test than on the paper-based test. Interviews with test takers after taking both tests indicated an even split in terms of

preference, with boys opting for the computer-based test and girls the paper-based test. Choi (2003) verified the comparability of paper-based language test (PBLT) and computer-based language test (CBLT) on the basis of content analyses, correlational analyses, ANOVA, and construct-related validation studies. The content analyses revealed that the sample tests representing 316 *Comparability of two types of language test* PBLT and CBLT were highly comparable in terms of content and linguistic features. The dimensionality check also revealed that the results did not violate the strong assumption of unidimensionality required by IRT, thus ensuring the appropriate application of IRT. The overall results of construct-related validation studies indicate comparability of the subjects' scores across CBLT and PBLT modes. The grammar test showed the strongest comparability, and the reading comprehension test the weakest comparability. The pattern of correlations among subtests, disattenuated correlations, and confirmatory factor analyses support to a certain extent that CBLT and PBLT subtests measure the same constructs.

The results of the existing studies indicate that the test scores of PPT and CBT are comparable and the differences are not significant, neither is the construct validity. From the mean score of the acquisition of PPT and CBT, there is research that concludes that the average scores f CBT results are higher, and also there is positive perception of the administration of CBT. These results are seem contradictory, and need to be strengthened by the results of other studies on the comparison of CBT and PPT.

METHOD

This study was conducted using the quantitative approach, by comparing the reliability and value of the test information function of CBT TOEP and PPT TOEP. The data are in the form of responses of TOEP test takers from all provinces in Indonesia, documented in 2008-2010 for PPT and documented in 2013-2014 for CBT, a sample of 600 test takers for each set of TOEP was established randomly. Three sets of TOEP, set 1A, 2A, and 3A for Listening and Reading section were analyzed.

The accuracy of PPT and CBT TOEP is known by comparing the reliability using the classical test theory approach and comparing the value of information function of both tests directly and through its relative efficiency. The reliabilities are estimated by calculating the reliability using Cronbach's alpha coefficient. On the item reponse theory, the item difficulty is estimated first before the value of the function information is. The estimation of item difficulties on Rasch model are done using the QUEST program (Adams & Khoo,1993). The value of test information function in every sets is estimated based on the difficulty index of the items on PPT and CBT.

The results of the estimation of the reliability and the value of the information function on PPT and CBT are then compared directly and by using the graphs. The administration of the test that has a greater value of information function is more accurate. The comparison of the value of the information function between PPT and CBT is also served as the relative efficiency of CBT to PPT, which is illustrated with graphs to be interpreted. If the relative efficiency is greater than 1, then CBT is more accurate than PPT. But on the contrary, if the relative efficiency is less than 1, then the PPT is more accurate than CBT.

RESULT

Using the participant's responses to TOEP set 1A, 2A, 3A both in listening and reading subtests, based on PPT or CBT, the reliabilities' estimation is done. More results are presented in Table 1. The results show that the reliability of TOEP scores tend to be stable at a high category, all of which are not less than 0.90. On set 1A, the reliability score on CBT is lower than the PPT and set 3A, both Listening and Reading-based PPT is slightly lower than in the CBT.

Table 1. Reliabilities of Score on Listening dan Reading Subtests Based on PPT dan CBT

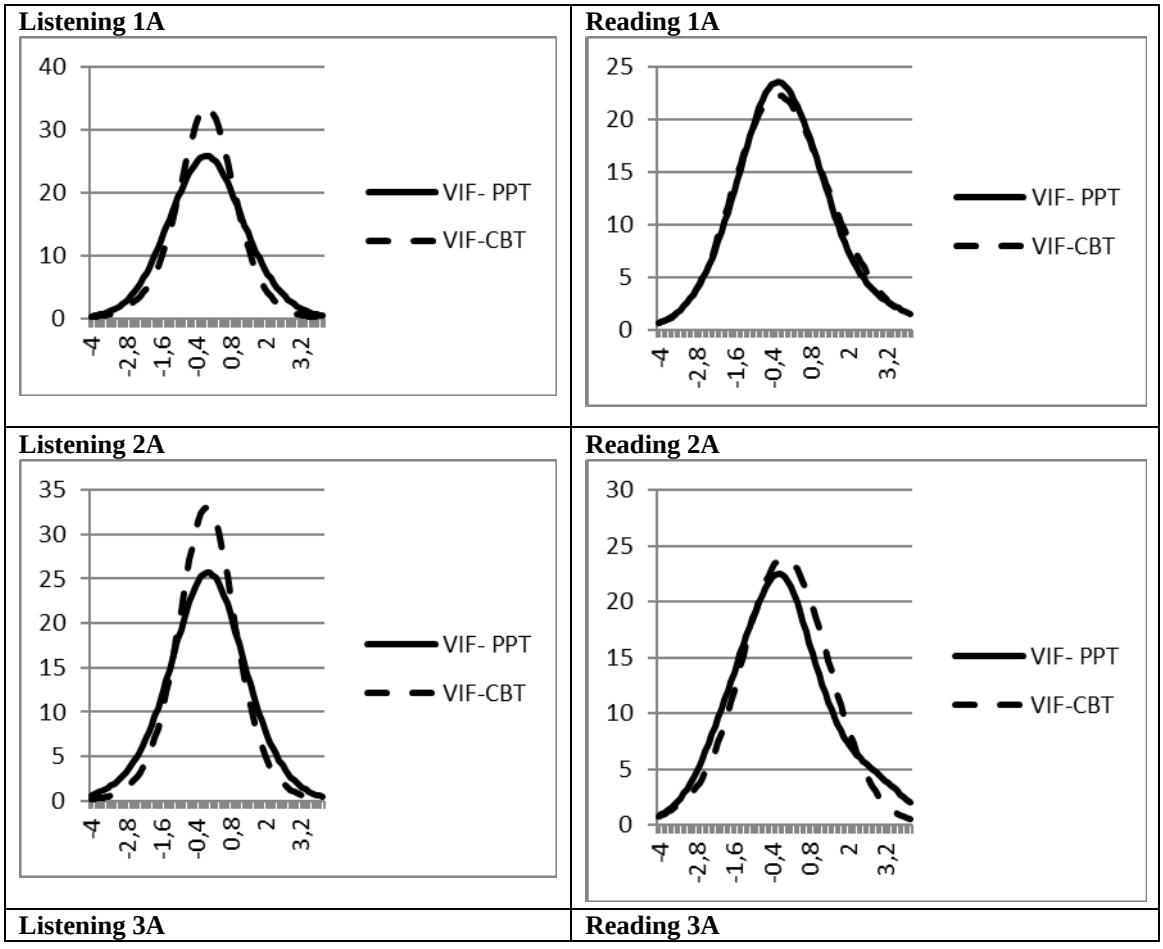
Set	Subset	PPT	CBT
1A	Listening	0.99	0.90
	Reading	0.99	0.99
2A	Listening	0.99	0.99
	Reading	0.99	0.99
3A	Listening	0.98	0.99
	Reading	0.98	0.99

Based on the test takers' response to the set tests, the difficulty index is estimated by using the Rasch model. Using the parameters of these items, the value of information function (VIF) are estimated, with the abilities ranging from -4 to +4, both on the Listening and Reading subtests, based PPT and CBT. The estimation results for each subtest are presented in Figure 1.

On Listening sets 1A and 2A, and Reading set 3A, there is a tendency which is almost the same. On the scale of the ability approach to the average (on a scale of 0), the value of information function on the CBT is higher than that on PPT. But on a low and high ability scale, the value of information function on PPT is higher than on CBT. This shows that in the ability scale approaching 0, CBT is more accurate than PPT, and on the low or high ability scale, PPT is more accurate than CBT.

On Reading 1A and 2A, and Listening 3A, the results show different things. On this set, the value of information function on the PPT and CBT is almost the same. This shows that in the three sets, namely Reading 1A, Reading 2A, and Listening 3A, there is the same accuracy scores obtained by TOEP takers between PPT and CBT.

These results are supported by the comparison between value of the information function obtained on PPT and that on CBT. On the Listening subtest 3A, the relative efficiency is relatively stable to the value close to 1, so it can be said the accuracy of the scores on Listening set 3A on PPT and CBT is almost the same. On the Listening sets 1A and 2A, on the ability around 0, the relative efficiency values of more than 1 indicates that CBT is more accurate than PPT. But on the contrary, on a scale approaching abilities approaching -4 and +4, the value of relative efficiency is less than 1. This indicates that the PPT is more accurate than the CBT. More results are presented in Figure 2.



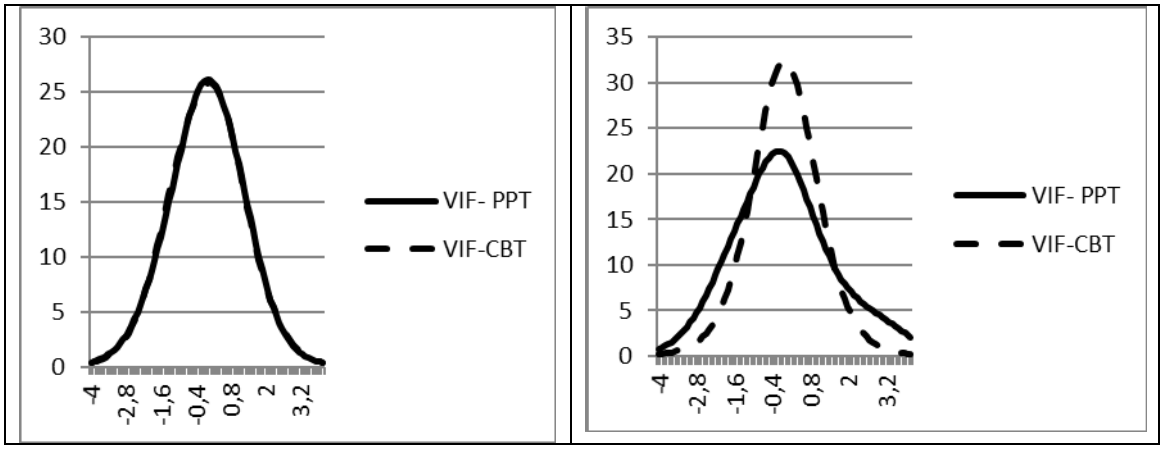


Figure 1. Value of Information Function (VIF) on Subtest Listening and Reading of TOEP based on PPT and CBT

Similar result occurs to Reading subtest. On Reading set 1A, the relative efficiency is around 1, except for high abilities. This shows that the accuracy of PPT and CBT is almost the same, except for the high abilities, in which reading ability is measured more accurately using PPT compared with CBT. On set 2A and 3A, there is a tendency that on the medium ability, the relative efficiency is more than 1, which shows that CBT is more accurate than the PPT. As for the low and high abilities, there is a tendency that PPT is more accurate than CBT. More results are presented in Figure 3.

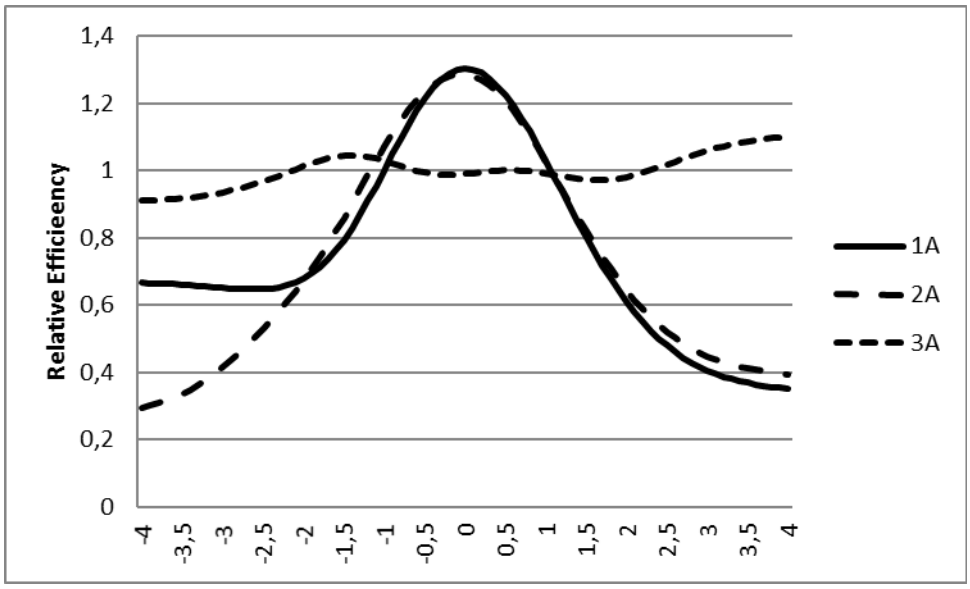


Figure 2. Relative Efficiency on Listening Subtest between CBT to PPT

The results of the analysis based on the classical test theory shows that the reliability scores on TOEP based on PPT and CBT are almost the same. This shows that, the accuracy of the score on PPT and CBT can be compared and the value is close to 1. With the high reliability, fewer measurement errors and higher accuracy of a test set will be obtained.

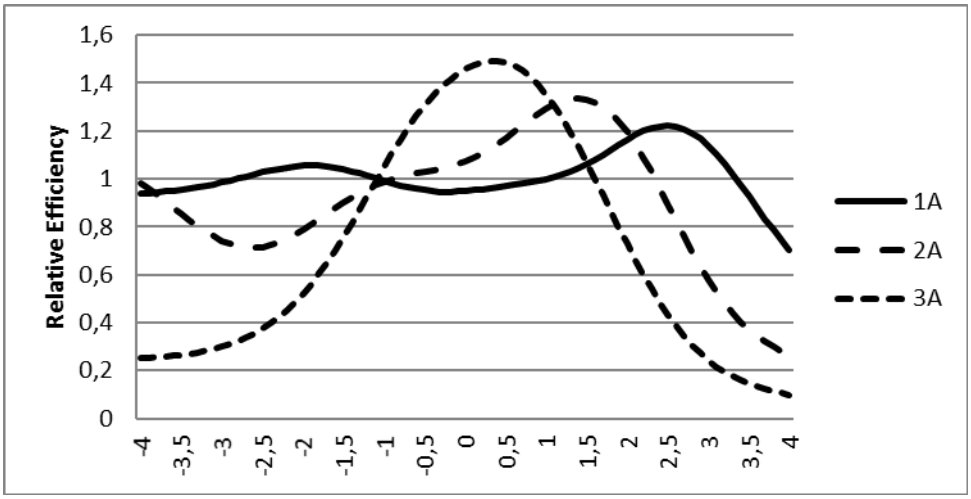


Figure 3. Relative Efficiency on Reading Subtest between CBT to PPT

In the estimation of the reliability and value of information function, the concern is the acquisition of scores, which does not overly affect the accuracy of measurement of the listening and reading abilities of the test takers of TOEP on PPTand CBT. This is in line with the finding of the research by Al-Amri (2007), which states there is no significant difference in scores of PPT and CBT and by Choi (2003), which proved that there is no difference between the construct validity of PPT and CBT. But the results of this study are different from the results of the study by Maguire, KA, Smith, DA, Brailler, SA (2010) which shows that students who completed all assessments electronically scored significantly higher than those students completing all assessments via pencil and paper.

The research finding from observing the comparative value of the information function, shows that though the value of function information on CBT and PPT is relatively similar in several subtests, there is a tendency that for test takers with moderate ability, CBT is more accurate than PPT, but for test takers in the low and high ability, PPT tends more accurate than CBT.

On the implementation of CBT, there are many obstacles that could hinder the test takers to do the tests. These constraints include test takers' unfamiliarity with the implementation of the CBT, the difficulty to log in, the difficulty using a headset for listening, using the mouse to answer, essentially related to the ability of the test takers using the information technology. Besides the obstacles, in some areas the availability of electricity and the slow internet network is a constraint in the implementation of CBT. The constraints in the administration of CBT in this study are in line with the opinion Noyes & Garland (2008).

CONCLUSIONS

The study shows that with the classical test theory approach, the reliability coefficients between the resulting scores of PPT and CBT almost the same, and using the item response theory, the researcher was found that although the value of the information function of PPT and CBT is relatively similar in several subtests, there is a tendency for testees with the moderate ability that, CBT is more accurate than PPT, and for those with the low and high ability, PPT tends to be more accurate than CBT.

REFERENCIES

Adams, R. J., & Khoo, S. T. (1993). *Quest: The interactive test analysis system*. Hawthorn: Australian Council for Educational Research.

Al-Amri, S. (2007). Computer-based vs. paper-based testing: does the test administration mode matter?. *Proceedings of the BAAL Conference 2007*.

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

Bachman, L. F. & Palmer, A. S. 1996. *Language testing in practice.* Oxford: Oxford.

Bachman, L. F. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Chee, T.S., & Wong, A.F.L. (2003). *Teaching and learning with technology.* Singapore: Prentice Hall.

Choi, I.C. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing* 20(3) 295-320.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening. *ReCALL* 18(2):193-211.

Croker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehard and Winston Inc.

De Gruijter, D.M. & van der Kamp, L.J.T. (2005). *Statistical test theory for education and psychology*. Retrieved from http://www.tu-dresden/de/erzwiae/ewmm/lehre/ in June 10, 2006.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc.

Jamil, M., Tariq, R.H., & Shami, P.A. (2012). Computer-based vs paper-based examinations: Perceptions of university teachers. *The Turkish Online Journal Technology*. October 2012, Volume 11 Issue 4.

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Maguire, K.A., Smith, D.A., Brailler, S.A. (2010).Computer-based testing: a comparison of computer-based and paper-and pencil assessment. *Academy of Educational Leadership Journal*, Volume 14, Number 4.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Elrbaum.

Munby, J. (1981). *Communicative syllabus design: a sociolinguistic model for defining the content of purpose-specific language programmes*. Cambridge: Cambridge University Press.

Noyes, J.M. & Garland, K.J. (2008). Computer-vs. paper-based task: Are they equivalent? *Ergonomic*. Vol. 51. No.9 September 2008, 1352-1375.

Retnawati, H. (2014a). The Equating of the test of English proficiency (TOEP). Paper. *Proceeding ICEEPS* Tokyo Japan 2014.

Retnawati, H. (2014b). Teori respons butir dan penerapannya (untuk peneliti, praktisi, pengukuran, dan pengujian, mahasiswa pascasarjana. Yogyakarta: Parama.

Reynolds, C., Livingstone, R.B., Willson, V. (2010). *Measurement and education*. Upper Sadle River, New Jersey: Pearson.

Stocking, M.L. (1999). Item response theory. In Masters, G.V. dan Keeves, J.P. (Eds). *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.

Towndrow, P.A., & Vallence, M. (2004).*Using IT in the language classroom:Aguide for teachers and students in Asia*(3rd ed.). Singapore: Longman Pearson Education South Asia Pte. Ltd.University Press.

Woolfolk, A. (2007). *Educational psychology* (*10th ed*.).New York: Pearson Education, Inc.