Comparative Study of The Difficulty Index of Items Displayed with The Paper-Based Test and The Computer-Based Testing Paper

IOP Conf. Series: Journal of Physics: Conf. Series **1244** (2019) 012044 doi:10.1088/1742-6596/1244/1/012044

Comparative Study of The Difficulty Index of Items Displayed with The Paper-Based Test and The Computer-Based Testing Paper

Syahrul*, Iwan Suhardi, Lu'mu
Faculty of Engineering, Universitas Negeri Makassar, South Sulawesi, Indonesia

*Corresponding author: syahrulab@yahoo.co.id

Abstract. The focus of this study is to analyze whether there are differences in the difficulty index of items if the same items are displayed with the Paper Based Test (PBT) model and the Computer-Based Testing (CBT) model. The PBT and CBT models have the same paradigm of measuring estimated abilities but have differences in the context and feel aspects. These differences include the number of items in the range of eyesight, use of tools, how to work on items, basic knowledge needs about computer operations, and habit factors. These differences can affect the results of estimating the ability of test participants. This study uses development methods and quantitative methods. Development methods are used to develop package items and CBT software. Two groups of respondents were used with equal ability to work on the same question package. One group uses the PBT model, and the other group uses CBT so that the response choices are obtained in each test model. The results of the respondent's answer choices were then analyzed by the ITEMAN software to get the item difficulty index. The item difficulty index of each model was tested statistically using SPSS software, whether there was a significant average difference between the two test models. From the results of the study and analysis in classical theory, it can be concluded that statistically there are differences in the average difficulty index of the items if the same item is displayed with the PBT and CBT models. It was found that the items displayed with the PBT model had a more difficult tendency than when displayed with the CBT model.

Keywords: index of difficulty, paper-based test, computer-based test

1. Introduction

Implementation of National Examination in Indonesia currently uses 2 (two) testing model that is using paper media and computer media. National Examination using paper known as UNKP, namely Ujian Nasional berbasis Kertas dan Pensil. A national exam by using computer media known as UNBK, namely Ujian Nasional Berbasis Komputer.

The role of Computer Based Testing (CBT) began gradually exactly replace the Paper Based Test (PBT) [1][2]. A concrete example it is on the development of the National Exam in Indonesia. Trends in the use of the computer-based testing model in an educational environment predicted to continue rising to replace the paper-based test model. UNBK participant ratio compared to all participants of the UN from 2015 to 2017 increased more than 33 fold as shown in Table 1.

Table 1. UNBK Participant Statistics from 2015 to 2018

Description	UNBK 2015	UNBK 2016	UNBK 2017	UNBK 2018
The ratio of UNBK participants compared to all of the National Examination participants	2,33 %	12,10 %	48,93 %	78%
	from 7.3 million students	from 7.6 million students	from 7.7 million students	from 8.1 million students

The National Examiner Organizer assume that an item is displayed on a computer monitor has the same index of difficulty when displayed on paper media. With these assumptions, the model testing and UNKP UNBK considered equivalent. The researcher's knowledge, a study of difficulty index analysis of items on a display model PBT and CBT have not been done and is still limited to assumptions.

Analyzing the items difficulty index means reviewing items so that questions can be obtained which are easy, medium, and difficult [3]. The formula for determining the items difficulty index using the classical theory is:

$$I_i = \frac{B_i}{N} \quad (1)$$

where:

I_i = difficulty index for each item

B_i = the number of students who answered correctly for each item

N = the number of students.

The items difficulty index in classical theory has a range of values from 0 to 1. The smaller the index obtained, the more difficult the problem is. Conversely, the higher the index obtained, the easier the problem is. Therefore, the traditional items difficulty index shows the 'easiness' of items.

The initial development of the testing system involves the use of sheets of paper for exam scripts and answer sheets, as well as pencils or pens [4]. Therefore, this examination system is often known as a paper and pencil test (Paper and Pencil Test or Paper Based Test or PBT). The weakness of PBT is the confidentiality of tests is difficult to guarantee because it can be read by people who are not authorized or not responsible [5]. Beside. The use of paper is a separate problem, for example, the process of printing, distribution, logistics, and the required storage space for test devices.

The use of computers for testing facilities is often called Computerized Testing or Computer Assisted Assessment (CAA) or Computer Based Testing (CBT). CBT is a testing method that is held by using a computer as the primary media in conducting exam activities [6]. The principle of CBT is mostly moving the PBT paradigm into a computer screen.

Psychometrically there is almost no excess of CBT compared to conventional tests. PBT and CBT use the same number of items for each participant or fix-length test. The approach used in scoring uses classical test theory or CTT. Because using computer media, CBT has advantages compared to PBT, namely (1) improve standardization, (2) improve test security, (3) improve test display capabilities, (4) minimize the error of measurement, and (5) accelerate the scoring and interpretation [5].

Although having a paradigm of measuring the estimated ability of the same test participants, the PBT and CBT models have a striking difference regarding the context and the feeling. Comparison of the context and feeling aspects between the PBT and CBT models faced by examinees is presented in Table 2.

The difference in the aspects of context and atmosphere between PBT and CBT is possible to influence the results of estimating the ability of test participants. Psychometric experts, such as Rudner [7] and Grist [8], argues that the parameters of the items used in PBT may not match their appearance on the computer monitor screen.

Table 2. Comparison of the Context and Feeling Aspects between PBT and CBT Models

Context and Feeling Aspects	PBT Model	CBT Model
Number of items in view	Consists of many items	Usually, there is only 1 (one) item only, even for long items that have to be a scroll.
Test equipment	Paper and pencil	Monitor screen, *CPU*, *keyboard*, *mouse*, and *speaker*
The items form that is capable of being displayed	Text and image	Text, image, audio, and video
The model works on the item	Give a choice of answer questions that are considered correct in pencil	Choose answers that are considered correct with the mouse or keyboard
Aspects of basic knowledge about information technology	Not required	Required
The color of the item at hand	Generally black	Allow all colors
The factor of the test work habits	It is common	Not yet a habit

Since from elementary school, students have been accustomed to working on exam items with paper media. Generally, the new school introduces a computer-based testing model ahead of the UNBK. The difference in context and feeling between the PBT and CBT testing models, as well as habits in the examination process, can affect psychological condition when working on computer-based exam items. The influence of anxiety and anxiety factors before and during the examination process can cause students to be unable to focus on doing the test well when using a computer-based exam model.

Along with the increasing role of the CBT testing model to replace PBT, the level of equity of the difficulty index of the items needs to be further analyzed. From the background of the problem, the researcher tries to examine whether there is a difference in the items difficulty index if the item is displayed with the PBT model and the CBT model.

2. Research Methods

This study uses a combination of development methods and quantitative studies. Development methods are carried out to develop test package test devices and CBT software. The question package consists of 40 items with material taken from the Indonesian X Class High School subjects by the 2013 Curriculum. The character of the package is made as carefully as possible between the PBT and CBT models from the aspect of appearance and technical work on the items. As with the technical work of the PBT test model, CBT software is designed so that respondents can choose the desired item number and can review the response if they want to replace it. How to answer the items on the PBT is selected by giving a circle mark with a pencil on the answer sheet that is considered correct. While the way to answer the item on CBT is chosen by selecting the answer that is considered correct with the mouse device.

Two groups of respondents were used with a total of 100 (one hundred) respondents. These groups are assumed to have equal capabilities. One group uses the PBT model, and the other group uses CBT so that the response choices are obtained in each test model. In the PBT model, the question package that has been developed is printed as many as the number of respondents needed. For the CBT model, the developed question package was included in the database of CBT software developed. The CBT model is developed using a web-based client-server system that can be accessed via a LAN network. Each CBT model respondent uses 1 (set) computer equipment.

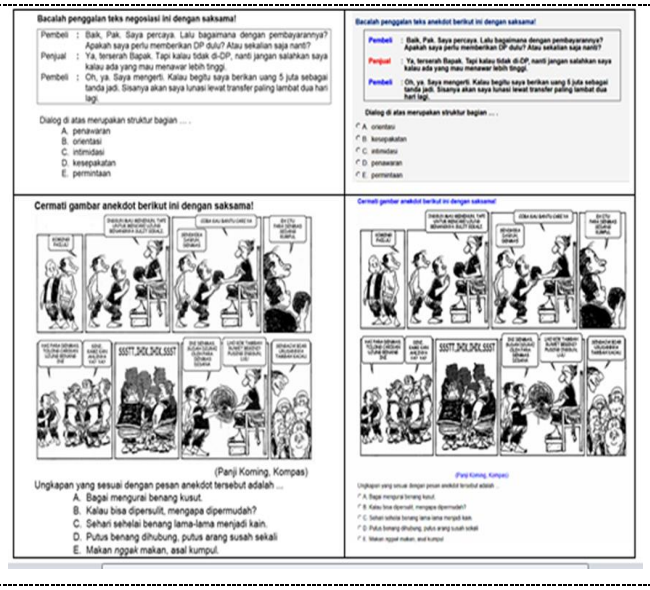

Figure 1. A sample of items on display of PBT and CBT

The results of the answer choices of the respondent groups were then analyzed by the ITEMAN (Item And Analysis Manual) software to get the item difficulty index. The item difficulty index of each model was tested statistically using SPSS software, whether there was a significant average difference between the two test models. In general, the mindset flowchart in this study is presented in Figure 2.

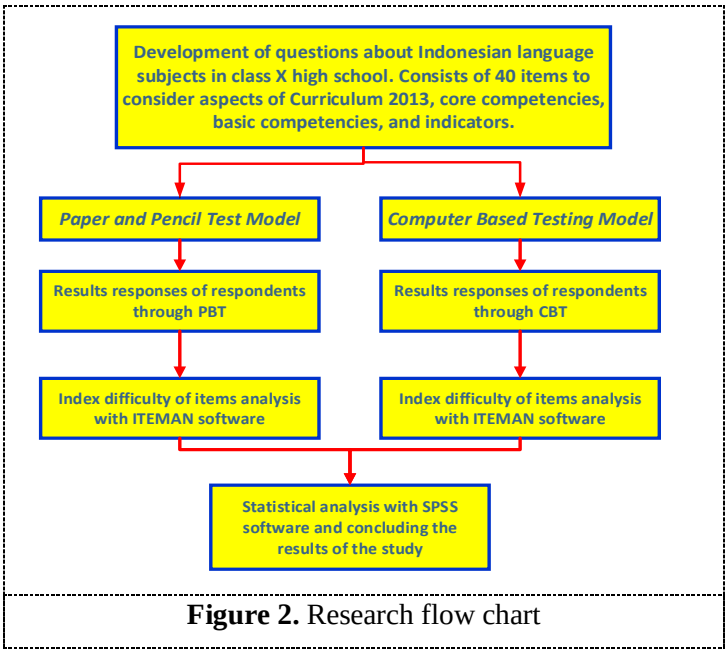

Figure 2. Research flow chart

3. Results and Discussion

The ITEMAN software is used to obtain a comparison of the items index difficulty from the PBT and CBT models. Response answers from the results of testing the PBT and CBT models are arranged in the composition according to the following Figure 3.

Figure 3. Composition analysis of items index difficulty with ITEMAN

The input data format of ITEMAN is done by entering the respondent's response in text form using the Notepad program with the composition as shown in Figure 4.

Figure 4. Data file format

In order for files to be analyzed with ITEMAN software, certain procedures are needed. The placement of data values in each row and column determines the correctness of the analysis results. 4 (four) command lines are needed as control lines are typed starting from the first row of the first column. The following is an explanation of these lines.

3.1. First line

080 0 N 04 states the number of items analyzed is “80 080” item (covering each of the 40 items with the PBT and CBT methods). Character “0” in the fifth column to show an empty answer. The character “N” shows for questions that have not yet been worked. The character “04” shows the character length for the identity of the respondent.

3.2. Second line

The second line contains the answer key, which is the answer key for each item (80 answer keys, each of which is the same 40 key answers for the PBT and CBT methods).

3.3. Third line

The third line states the number of answer choices. The number "5" states the answer choices include, A, B, C, D, and E.

3.4. Fourth line

The fourth line states that the code “Y” for the items analyzed and “N” states items that are not analyzed.

3.5. Fifth line and so on

The fifth line and so on contains the response of the respondent's answer with the provisions of each row showing the answer of one respondent. The output of ITEMAN software is presented in Figure 5.

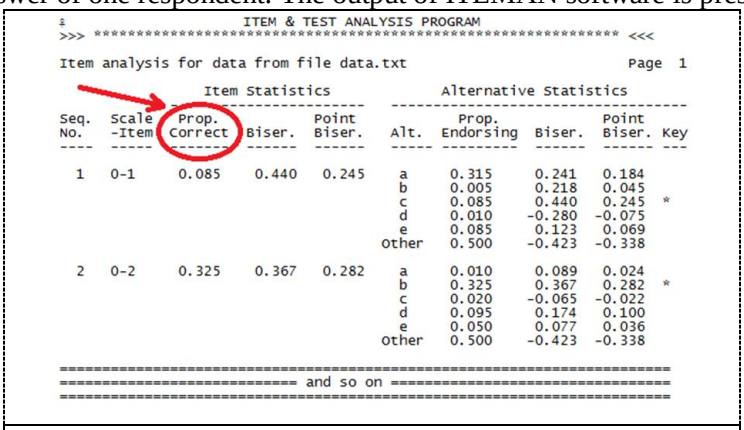

Figure 5. Display of ITEMAN analysis results

The focus of the item difficulty index analysis in this discussion focused on the results in the Prop. Correct which describes the proportion of respondents who answered the test items correctly. Extreme values close to zero or one indicate that the item is too difficult or too easy for test participants. This index is also called the index of item difficulty index in a classical manner. From the ITEMAN program output file, it can be seen the comparison of the items difficulty index is classically between PBT and CBT models.

Table 3. Comparative index of problem points on PBT and CBT models

Item	The Items Difficulty Index	
	PBT	CBT
1	0,085	0,060
2	0,325	0,160
3	0,120	0,215
4	0,440	0,385
5	0,450	0,400
6	0,395	0,340
7	0,280	0,375
8	0,445	0,425
9	0,195	0,245
10	0,415	0,350
11	0,145	0,115
12	0,085	0,165
13	0,460	0,425
14	0,065	0,040
15	0,465	0,460
16	0,305	0,250
17	0,245	0,335
18	0,390	0,275
19	0,400	0,270
20	0,495	0,420
21	0,480	0,405
22	0,430	0,330
23	0,410	0,360

Item	The Items Difficulty Index	
	PBT	CBT
24	0,175	0,125
25	0,335	0,245
26	0,420	0,330
27	0,455	0,425
28	0,020	0,065
29	0,430	0,300
30	0,430	0,275
31	0,330	0,265
32	0,185	0,125
33	0,235	0,175
34	0,125	0,155
35	0,140	0,230
36	0,470	0,375
37	0,390	0,240
38	0,445	0,375
39	0,300	0,230
40	0,345	0,295

Figure 6 presents a comparison of the results of the of items difficulty index between the PBT and CBT testing models in graphical form. From the same package item, 83% of the items in the CBT test model are more difficult than the CBT test model.

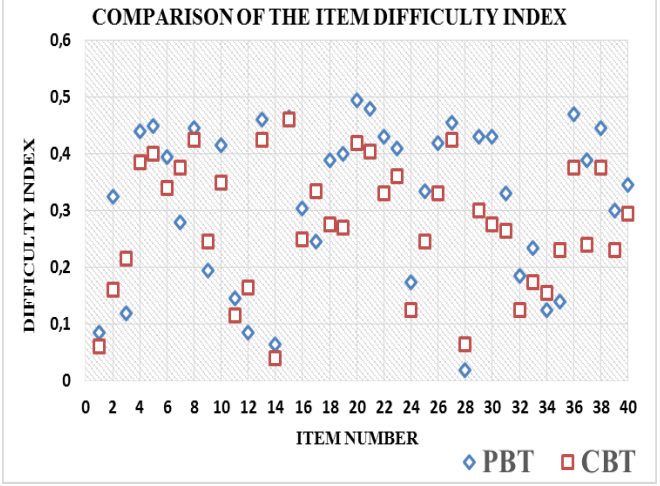

Figure 6. Comparison of the items difficulty index of PBT and CBT models graphically

The difficulty index of the PBT and CBT model results of the ITEMAN program output is then tested with the SPSS program. Statistically tested whether there were significant differences between the difficulty index items with the PBT and CBT models. The procedure used is paired sample T-test. This procedure was chosen because the data tested were the same items and were obtained with the same analysis procedure. The confidence level used is 95%. The results of the SPSS analysis results from the comparison of the items difficulty index about the PBT and CBT methods can be presented in Figure 7.

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PBT	.31888	40	.139851	.022112
	CBT	.27300	40	.113617	.017964

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PBT & CBT	40	.873	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PBT - CBT	.045875	.068807	.010879	.023869	.067881	4.217	39	.000

Figure 7. SPSS output analysis of the items difficulty index

From the results of SPSS output, it can be observed that the average difficulty index of items with the PBT model is 0.3189 with a standard deviation of 0.1399 and the CBT model is 0.2730 with a standard deviation of 0.1136. Correlation results show a value of 0.873 with a significance of 0.000. This means that there is a close relationship between samples or statistically significant correlations.

For testing, whether there are differences in the difficulty index of items from the PBT and CBT models, the steps taken are:

1. Formulate a null and alternative hypothesis
   H0: The average item difficulty index with the PBT and CBT methods is the same
   Ha: The average item difficulty index with the PBT and CBT methods is not the same
2. Determine the confidence interval used used
   The confidence interval used is 95% or by using alpha 5%.
3. Determine decision-making rules
   The decision-making rule is to accept H0 if t count is smaller than t table and rejects H0 if t count is higher than t table. Based on t table with alpha, 5% 2-tailed test or 2.5%, and the degree of freedom of 39 or (n-1) is obtained by t table value of 1.68488. So the decision taken is to accept H0 if t count is smaller than 1.68488 and rejects H0 if t counts are more significant than 1.68488.
4. Calculate t count or t statistic
   The t calculated value from the output of SPSS software obtained the value of t = 4.217.
5. Decision making and results interpretation
   After calculating t statistics, the last step is to decide on the results of the analysis and interpretation of the results. The average difference between the PBT and CBT methods is 0.045875 with a standard deviation is 0.068807. The results of the t statistic calculation produce value is 4.217, and the significance is 0.000.

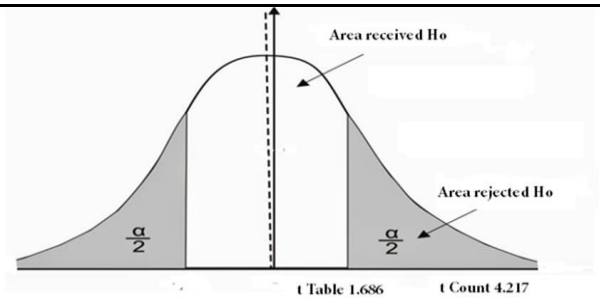

Figure 8. Two tail hypothesis test

With a significance result of 0.000, a decision can be made to reject H0 because the significance level is smaller than alpha (0.025). The results of the calculation of the t count value (4.217) turned out to fall in the rejection area, then H0 was rejected. Therefore Ha is accepted. That is, from the results of the analysis it can be stated that statistically, the items difficulty index of the two methods (PBT and CBT) is different for each item.

The mean difference of 0.045875 shows the difference in the average difficulty index of the items between the PBT and CBT models, namely 0.3189 for PBT and 0.2730 for CBT. Regarding classical measurement theory, there is a smaller difficulty index value means that the item has more rigorous criteria than another method. So, in general, it can be concluded that the same item if done with the CBT model will be felt more difficult by the test participants compared to when done with the PBT model.

The results of the analysis show that in classical theory there are differences in the difficulty index of an item if the item is displayed with the PBT model and the CBT model. The results of this study indicate that the items displayed on the CBT software monitor screen using a mouse and keyboard have different difficulty indexes when presented on sheets of paper using a pencil. The difference is possible because although it has a paradigm of measuring the estimated ability of the same test participants, the PBT and CBT models have a striking difference regarding the context and the feeling.

The possibility of these differences is due to the unfamiliarity of students working on problems with computer-based testing models. In general, not many schools have applied computer-based testing models to classroom learning practices. Generally, the learning process still uses paper media in the testing process both at the time of the daily, mid-test, or end-test. The use of paper media in the learning process has become a habit even since students go to school starting from kindergarten, elementary school, junior high school, and high school.

There has been no computer-based testing model in schools due to many reasons. The main reason is the lack of computer facilities that can be used by all students in the school. The number of computer laboratories generally consists of only a few rooms in each school, so the number of computers is not proportional to the total number of students. With the policy of the computer-based national examination applied by the government, it is generally addressed by holding socialization on the use of computer-based examinations for students at the end of the level a few months before the National Examination. However, such a short time does not necessarily result in the habit of students doing the computer-based test.

Having the essential ability to operate computer equipment is not a guarantee that students are familiar with computer-based testing models. The paradigm is considering that the habit of using paper media testing has been going on for many years, while the socialization of testing using a new computer is carried out within a few months. There may be psychological barriers that influence the results of differences from the PBT and CBT testing models. Unusual work on computer-based exams makes students unable to show their best abilities when doing the exam. The habit factor of students working on items using the PBT model without realizing it has a less supportive effect.

In the use of CBT, it is necessary to consider the aspect of computer self-efficacy, that is how confident a student sees himself to be successful in computer-based tests. This computer self-efficacy factor plays an essential role in determining the success of students in the exam. Computer self-efficacy helps reduce students' anxiety levels in taking computer-based exams [9]–[12]. With reduced levels of anxiety, students can focus more on working on questions and can show their best abilities. The hope is that the results of the exam can be maximized.

On the other hand, one way to reduce test anxiety by using a computer is to improve students' computer experience and confidence in computer-based exams [13], [14]. The best way is to optimize the preparation period. Providing opportunities for students to become familiar with the CBT model is essential [15]. Familiarizing students with increasing trials of CBT models before the test day can reduce anxiety factors. Familiarizing students with computer-based exams is very useful for students who are economically disadvantaged and do not have a computer at home to improve their computer operating experience.

4. Conclusions

From the results of studies and analysis in classical theory, it was concluded that statistically there were differences in the difficulty index of an item if the item was displayed with the Paper-Based Test (PBT) model and the Computer-Based Testing (CBT) model. It was found that the items if the items displayed with the PBT model tended to be more difficult than when displayed with the CBT model.

The need for more socialization for students using the CBT model in the learning model and testing in the classes so that students are more accustomed to using the CBT model. The socialization is in line with the Indonesian government's plan to expand the role of the computer-based testing model (UNBK) to replace the paper-based testing model (UNKP) in the National Examination in the future. The need for further study is to analyze the relationship between the index of familiarity and habitability of students on the operation of computers with difficulty index items on the PBT and CBT testing models in order to become the basis for psychometrics to understand the tendency of differences in difficulty index items in the PBT and CBT testing models.

References

[1] A. C. Bugbee Jr, "The equivalence of paper-and-pencil and computer-based testing," *J. Res. Comput. Educ.*, vol. 28, no. 3, pp. 282–299, 1996.
[2] O. Publishing, *PISA computer-based assessment of student skills in science*. OECD, 2010.
[3] M. J. Allen and W. M. Yen, "Introduction to measurement theory. California: Wadsworth." Inc, 1979.
[4] K. S. Shultz, D. J. Whitney, and M. J. Zickar, *Measurement theory in action: Case studies and exercises*. Routledge, 2013.
[5] C. V. Bunderson, D. K. Inouye, and J. B. Olsen, "The four generations of computerized educational measurement," *ETS Res. Rep. Ser.*, vol. 1988, no. 1, p. i-148, 1988.
[6] B. A. Orenyi and M. M. Omotosho, "Computer-based Test Software System: A Review and New Features," *Computer (Long. Beach. Calif).*, vol. 55, no. 15, 2012.
[7] L. M. Rudner, "An online, interactive, computer adaptive testing tutorial," *ERIC Clear. Assess. Eval.*, 1998.
[8] S. Grist, L. Rudner, and L. Wise, "Computer adaptive tests. ERIC clearinghouse on tests, measurement, and evaluation," *Am. Inst. Res. Washington, DC*, 1989.
[9] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Q.*, pp. 189–211, 1995.
[10] J. W. Creswell, "Research design pendekatan kualitatif, kuantitatif, dan mixed," *Yogyakarta: Pustaka Pelajar*, 2010.
[11] S. P. John, "Influence of computer self-efficacy on information technology adoption," *Int. J. Inf. Technol.*, vol. 19, no. 1, pp. 1–13, 2013.
[12] H. K. Sam, A. E. A. Othman, and Z. S. Nordin, "Computer self-efficacy, computer anxiety, and attitudes toward the Internet: A study among undergraduates in UNIMAS.," *J. Educ. Technol. Soc.*, vol. 8, no. 4, 2005.
[13] M. Zeidner and G. Matthews, "Encyclopedia of Psychological Assessment," 2003.
[14] R. M. Liebert and L. W. Morris, "Cognitive and emotional components of test anxiety: A distinction and some initial data," *Psychol. Rep.*, vol. 20, no. 3, pp. 975–978, 1967.
[15] M. Russell, "Testing On Computers," *Educ. Policy Anal. Arch.*, vol. 7, p. 20, 1999.