International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 16, Issue 2, 2022

Achievement Tests Administration using Computerized Adaptive Testing (CAT) with Constrain Response Time Item

Handaru Catu Bagus[1*], Burhanuddin Tola[2], Awaluddin Tjalla[3], [1,2,3]Postgraduate, Universitas Negeri Jakarta, Indonesia, Correspondence Email: handcab@gmail.com

Assessment models that ignore the ability of individual variations cause the information received will not be optimal. The computerized adaptive testing (CAT) model can overcome this weakness because the items that appear with the level of difficulty will adjust to the ability of the test taker. The purpose of this study was to analyse the effectiveness, efficiency and accuracy of the CAT model when used as an assessment model in school achievement tests. The research methodology is comparative quantitative. This study uses population data from the answers of students who took part in the UNBK in the province of the Special Capital Region (DKI) Jakarta in 2019 with mathematics as a subject. The results of this study show that the number of items selected by the CAT model is less than the CBT model and the items adjust to the level of ability of the traveller and have a small measurement error value of the reference value and are almost the same as the CBT model. Therefore, the CAT model is more efficient in terms of time because the number of questions is less than the CBT model, effective because it adapts to the abilities of the participants and has the same accuracy as the CBT model.

Keywords : *computer, computerised adaptive testing, computer-based test, evaluation, adaptive model, item response theory*

1. Introduction

1.1 Research background

The Indonesian national examination, known as *Ujian Nasional* (UN), is an achievement test with national standards measured for final grade students at the junior and senior high school levels administrated in all Indonesian schools. In its implementation, the UN uses two test administration models, namely paper-pencil test (PPT) or the conventional model and computer-based test administration or computer-based test (CBT). The two test administrations used an assessment model with a similar test design for each test taker, regardless of age and level of education. This assumption underlies test-takers of the same age and education level who have the same abilities. Even though, in practice, there are very varied abilities for every test taker.

The assessment model in the UN is weak because it neglects the diversity of individual abilities that makes the test information obtained will be not optimal. For example, in an easy-test category given to a high ability test taker, the received test information will be less meaningful because the test taker will answer all items correctly. On the contrary, in a difficult-test design category given to a low ability test taker, there is a high chance that the test-takers will get a meagre score. The conventional assessment model in the UN causes the fairness issue that leads to inaccurate and precise test results.

Article 1 paragraph 1 PERMEN DIKNAS no. 75 of 2009 states that the National Examination, hereinafter referred to as the UN, is an activity of measuring and assessing the competence of students nationally at the primary and secondary education levels. The UN was officially held from 2005 to 2020 and starting in 2021 the UN was abolished. And since 2015 the National Examination is no longer a determinant of graduation at all levels, so that the National Examination is only a mapping of the achievement of the learning process in the classroom in accordance with the curriculum, but the examination is carried out by another independent party or institution, namely the National Education Standardization Agency, hereinafter referred to as BSNP with the hope to be more objective. And since 2015 the UN has implemented a computer-based UN, hereinafter referred to as UNBK, so that since 2015 it has had two modes of implementation of the UN, namely PPT mode and CBT mode.

In the technical Standard Operating Procedures (POS) of the National Examination for the 2016—2017 academic year in chapter IV on National Examination Materials, it is clear that the test package given to students ignores variations in individual abilities, given that the UN test package is for one study group at one school which prepared, duplicated and given to students for both modes using an equivalent test package design. Therefore, it is difficult to obtain on-target information due to the design of an equivalent test package, especially for the variance of the test taker's ability category. To overcome these problems, the authors try

to examine the responses of the participants of the UN test using the CBT model, which is modeled with the adaptive test model.

To overcome the weakness of the educational assessment model that has been going on so far, the adaptive test model can be an alternative to test administration. This is because this model allows the test design to obtain on-target information, because the test with the level of item difficulty is adjusted to the ability of the test taker. Therefore, it is hoped that the use of the test design will produce optimal information because the test model will stop after information on the ability of students can be estimated. The technique of estimating the ability of test takers with an adaptive test model uses a modern test theory approach or Item Response Theory (IRT). The IRT approach is oriented to items related to the ability of the test taker, and is not oriented to the test instrument. (Hambleton, 1991) Therefore, with the IRT approach the performance of a person or group of people in an item can be predicted.

In the process of the adaptive test model in selecting and displaying items that are adjusted to the information on the abilities of students, to make it easier this model is assisted by computer or computerized media so that the results obtained will be faster, more effective and accurate in producing optimal information. Based on the explanation above, it can be understood that Computerized Adaptive Testing (CAT) is an adaptive test in which the presentation of the test is assisted by computer media, including the selection of items to processing test results. Bunderson (1989) noted several advantages of CAT, including: increasing control in displaying items, increasing test security, enriching display capabilities, obtaining the same score in a shorter time, reducing error of measurement, increasing scoring and reporting.

This CAT model research is very important to do as an alternative to the assessment model, especially achievement tests either at school or nationally or at the National Examination, which has been implemented in Indonesia. In addition, the problem of leakage and fraud that has occurred in the implementation of the UN can be minimized.

1.2 Research Question

1) Is the CAT model more time efficient than the CBT model if the CAT is modeled in the Achievement Test?
2) Is the CAT model more effective in the number of items presented compared to the CBT model if the CAT is modeled in the Achievement Test?
3) How is the accuracy of the CAT model when the CAT model is compared with the CBT model?

1.3 Research purposes

1) Analyzing the efficiency of processing time with the CAT model compared to the CBT model when CAT is modeled in the achievement test.
2) Analyzing the effectiveness of the number of questions presented in the CAT model compared to the CBT model if the CAT is modeled in the achievement test.
3) Analyzing the test of the relationship or correlation between the results of the ability of test takers using the CAT model compared to the CBT model. This is to measure the accuracy of the CAT model.

1.4 Research Benefits

In terms of benefits from this research, there are two benefits, namely: 1) as an introduction to new concepts to the academic world in the form of reference materials in educational assessment models; 2) as input for policy makers in improving the education assessment system in Indonesia.

2. Literature Review

Educational assessment or assessment is part of the teaching and learning process. assessment is a series of activities to obtain, analyze, and interpret data about the process and learning outcomes of students which are carried out systematically and continuously, so that they become meaningful information in decision making by the school or decision makers. (http://rbaryans.wordpress.com/2007/07/27/, accessed March 30, 2022)

The teaching and learning process also recognizes the types of assessments including the Assessment of Learning or better known as the learning achievement test. This assessment is given to students after students have studied and generally this assessment is given at the end of each semester. The UN is a summative test of student learning outcomes carried out by students at the final level at each level. Learning outcomes are things that are obtained by someone who carries out the learning process with a predetermined rating scale by measuring the level of learning success which is usually done with the help of exams (Suprijadi, 2010). The National Examination is also known as a summative test with a package construction designed to measure the knowledge or skills of test takers from a material that has been studied or taught by a teacher (Gregory, 2013). Therefore, summative assessment is used to measure the achievement of a student after the teaching process from the teacher, so that summative assessment can be said to be the final achievement test of each teaching process.

Achievement test according to Cronbach (Reynolds, et al. (2010:4)), is included in the maximum performance test group. The maximum performance test is the best ability that can be shown by the test taker as an answer to the item. The construction of this type of test must have many stimuli (in the form of questions) that are clearly structured. Questions and the direction of the desired answer in the measuring instrument, must really be understood

by the test takers before they give a response. Because the test taker's response is related to his cognitive ability, the response chosen by the test taker can be said to be a "true" or "wrong" response and be given a commensurate score.

In the implementation of achievement tests either nationally or for the sake of class or school assessment, assessment of students in the form of tests or examinations of test takers can be carried out in various ways, starting with test administration in the conventional way, namely by using paper (paper-pencil test) to the use of technology, such as Computer Based Test (CBT) and Computerized Adaptive Test (CAT). The following will explain about the forms of test administration that are mostly carried out.

2.1 Computer Based Test

Computer Based Test (CBT) is the first generation of test administration in the use of computers in testing. With CBT the appearance of the questions is assisted by computer equipment. In terms of the test package design, CBT is the same as PPT, in that the form of administration is still conventional because all test takers receive the same set of questions, and ignore the diversity of abilities of the test takers. However, the advantage of CBT compared to PPT is that all activities of the test taker can be recorded by the system so that student work activities become more easily controlled and all information is recorded quickly by a computerized system. Therefore, the activities of the test takers from the beginning of the test to the end of the test are recorded and recorded by the system, including student responses, item response times, response patterns when visiting from one item to another, and several other activities. With the advent of computer-based testing, recording response times has become a routine part of the test mode (van der Linden: 2006). The results of this activity recording will be used to find out how long it takes the test taker to respond to the answers to each item. In the implementation of the UN in 2019, CBT as the test administration used in UNBK.

2.2 Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) is the second generation of using computers for testing. (Bunderson, 1989) One application of the IRT approach is the use of CAT. Adaptive means that the items given are in accordance with the abilities of each test taker, so that each individual will get a different set of items. Leung (2005) said that in the CAT a test taker is given questions that are selected based on his estimated ability (θ). Because each individual gets a set of questions that are tailored to his abilities, the CAT is included in tailored-testing. Thus, the IRT-based CAT usually contains fewer items than the usual or conventional PPT or CBT measurements. (Embretson & Reise, 2000)

From the explanation above, it is clear that IRT-based CAT usually contains fewer items than conventional PPT and CBT measurements. (Embretson & Reise, 2000) Empirical results are also explained by Olsen (in Bunderson, 1989) who notes that on a learning achievement test it only takes 30% to 50% of all items in the test package to achieve the same level of precision as the PPT or CBT. By reducing the number of questions given to

test takers, it will directly reduce the amount of response time needed to administer the test. (Bunderson, 1989)

The concept of the algorithm used by CAT is as follows: if the test taker cannot answer correctly on the items given, the computer will give items that have a lower degree of difficulty. If the test taker cannot answer correctly, the computer will provide items that have a lower degree of difficulty. On the other hand, if the test taker can answer correctly, the next item given is the item with higher difficulty. Figure 1 provides a chart of the CAT process proposed by Wainer (1990).

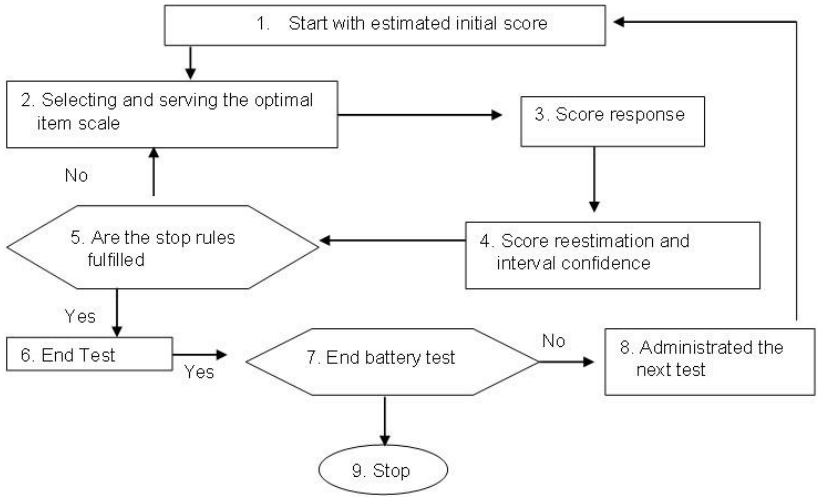


Figure 1. CAT Process Chart

In implementing a test into the IRT-based CAT, there are several things that must be considered. Embretson and Reise (2000) describe five factors that must be considered in CAT. First, Item bank. The purpose of the CAT is to administer a series of questions that can provide maximum information and efficiency for each test taker. To accomplish this, different test takers will receive different sets of items, and their scores on latent ability are estimated based on their responses to the different items. Thus, the capacity to realize CAT depends on an item bank containing a large set of items for which the IRT parameters of each item are known. The purpose of the test is to properly measure the entire range of abilities, so ideally an item bank contains a number of items that have high discriminatory abilities with difficulty parameters spread across the range of abilities. When an item bank meets these criteria, all test takers can be administered the appropriate tests and they can be measured accurately. Second, Administer the first item. If it is assumed that the test taker's ability in a normally distributed population can be started with a difficulty parameter of -0.5 to 0.5. If information is obtained about the ability of the test taker in the ability continuum, that information can be used to select the level of difficulty in the items at the beginning. The average of the population taking the test can be used as an estimate of ability so that it

can be optimal. (Thissen & Mislevy, 1990). Third, scoring. There are three main methods for estimating the position of test takers on the ability continuum, namely ML (maximum likelihood), MAP (maximum a posterori), and EAP (expected a posteriori). Some researchers discourage the use of prior information because it has the potential to influence scores. For example, if only a few items are administered, the estimated proficiency level will be drawn towards the mean value of the initial distribution. For this reason, a step-size procedure is used to score in the early stages of CAT. Fourth, the selection of the next item. The selection of the next item is related to scoring. The strategy that can be used to provide the next item is maximum information and minimum expected posterior standard deviation, which is also called Bayesian estimation. (Thiessen & Mislevy, 1990) Maximum information is done by selecting items at each stage that have a value of b close to the current estimate of . Fifth, Stop the Test. In CAT, each time the test taker's ability is estimated based on the response to the item and the standard error is estimated again, the computer then selects the next item to give. There are two criteria for stopping CAT administration, namely variable length and fixed length. In variable length, CAT administration stops when the standard error measurement has reached the predetermined limit. Thissen and Mislevy (1990) call this criterion the target precision. Determination of standard error, according to Hornke (2000), with a standard error of less than or equal to 0.38, will be commensurate with the reliability coefficient of 0.85. On the other hand, Blais and Raiche (2002) found that if the standard error of measurement is less than or equal to 0.40, the standard error of the individual's ability level is only 0.03 different. The fixed length procedure is a test stop when a certain number of items have been administered. Thissen and Mislevy (1990) call this criterion the maximum number of items. The advantages are that it is easy to perform and the use of the item can be estimated with precision.

This study uses the theoretical framework as mentioned above as the basis for the methodology of the CAT model research process. Therefore, the five factors above become very important as the basis for the CAT model to be applied.

2.3 Item Response Time

Item response time is the amount of time required to complete each item. Each item has a level of difficulty, the more difficult the items given, the longer the time required by the test taker, and conversely, the faster the time, the easier the items. Item response time can be used for various purposes. One that is commonly used is that the average duration needed to take each item can be used to justify the right number of questions being tested on students for one exam session, so that there is no shortage of time (speed test) or too much time. Another benefit of item response time is to increase the accuracy of students' ability estimates, because students who can complete faster are assumed to have higher abilities than those who are slow in progress. In terms of items, there are a number of questions whose questions are very complex, so that the method of solving it becomes very difficult

and requires relatively unlimited time and the limit is the response ability of the test taker. Generally, the achievement test for learning outcomes is a power test, meaning that there is no time limit for test takers to answer all questions. However, with the difficulty level of the item increasing over time, so that with more free time some people will not be able to answer correctly (Chadha, 2009).

When the application of test processing time is limited, and especially when the test has a strict time limit, the placement of items with categories that are easily placed and displayed in the initial number of the test according to Faugher, Melton & Myers, (1968) will result in a higher score than when the items are with the difficult category being displayed in the first number and the easy category items being placed in the last number (Sax & Carr, 1962; Towle & Merrill, 1975). This reinforces the concept that the nature of the power test does not limit the time in which the test is carried out. However, every time the achievement test takes place at school, the time for taking the test must still be limited. According to Nitko, Anthony J. (1996) the estimated response time for the multiple-choice model is between 40 and 60 seconds, while the short answer test model for each item has an estimated response time of 15 to 20 minutes. The response time for items on the UNBK is the aggregate response time for items, which is 120 minutes. So that the aggregate time does not adjust the conditions of the difficulty level of each item.

If the UN uses the CAT implementation mode, the item response time settings can be directly modified according to the needs of the test. So with the CAT implementation mode, in addition to generating the items given to the test taker according to their abilities, with CAT it is also possible for each item that appears the results of the generation can be set the response time of the item. In addition to the item generator that is adjusted to the ability of the taker, the response time of the item, it is also possible to measure the estimation of the test taker's ability each time he responds to the item. each test taker with different abilities takes the same amount of time to complete the test. Thus, slower test takers may not be able to complete all the items on a test with a limited processing time.

2.4 True Theta (The Real Ability)

Every assessment, both in diagnostic tests and achievement tests, always contains measurement errors. Each ability (theta) obtained from the test taker consists of three things, first the observed ability value which is often also referred to as the estimated ability of the test taker (estimated theta), the second is the actual ability value (true score), which is a reference value that is in accordance with the test taker's ability. the actual, and the third measurement error, which is a factor that affects the inconsistency of a measurement so that it affects the score. Because true theta is influenced by the estimated value of the traveler's ability and measurement error, it can be formulated in equation one below in the mathematical equation as follows (Sumarna Surapranata, 2009: 88)

$$T = X + e \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

T : Actual ability (true theta) X : Estimated ability (estimated theta)
e : Error of measurement
So that it can be formulated in equation two below, the error in measurement is the difference between the actual ability and the estimated ability.

$$e = T - X \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

In the fact that it is difficult to measure true theta, it can be said that only God knows the true ability of each test taker, and humans can only measure the estimated ability of the test taker. However, for research needs, especially research using comparative methods or wanting to compare two or more measurement models, true theta can be associated with the value of the ability gain of each test taker from several measurement observations. Or the determination of true theta usually uses the acquisition value of the ability from the measurement results that the participant takes seriously, such as achievement tests at school or selection tests.

3. Research Method

This research methodology uses a quantitative and comparative approach. The research data is the answers or responses of junior high school students who take part in the UNBK or UN with test administration with CBT in 2019 in DKI Jakarta Province. This is because the UN response data owned by DKI Jakarta Province varies so that it is easy to observe and analyze. Meanwhile, the researcher focuses on mathematics, because it is the core competence of the researcher, besides that mathematics is an exact science subject and is in line with technological developments. The sample was selected systematically from a number of junior high school populations who took part in the UNBK in DKI Province.

The method of selection is by sorting the data from the lowest score to the highest score, then randomly selected until the total sample becomes 341 data, which consists of 19 data from percentiles less than 25, 217 data selected from percentiles between 25 to 75 and 105 the remaining data from percentiles more of 75. For the 341 data of the selected test takers, information on the response to the questions and the response time of the items will be obtained. Each item response data is reprocessed by modeling using the CAT model administration so that an estimate of the ability of the test takers will be obtained, and the number of questions that appear for each test taker. Meanwhile, item response time information is used to measure the length of the test associated with the number of items that appear for each test taker with CAT administration modeling.

Meanwhile, because the research methodology is comparative in nature, it is necessary to determine the true theta value as a reference value obtained from the observed values when the test taker takes the pre-UNBK mathematics test and during the UNBK. The results of the

two tests were averaged. And the average results are used by researchers as true theta values or reference values to obtain measurement errors from the results of ability estimates with test administration with CBT and CAT.

With the data mentioned above and by reprocessing by applying test administration modeling with the CAT model, an analysis and discussion of the efficiency of the large number of questions that appear for each test taker will be carried out if the test administration is using the CBT and CAT models, and will be carried out analysis and discussion of the effectiveness and accuracy of the results of the estimation of the ability of the test takers between the CAT and CBT models.

3.1 Operational Definition

Before explaining the research methodology further, the author will explain the definition of effectiveness, efficiency, and accuracy as the limitations of this study. According to Hidayat (1986) effectiveness is a measure that states how far the target (quantity, quality and time) has been achieved. The greater the percentage of targets achieved, the higher the effectiveness. According to Mahmudi (2010: 143-166) effectiveness is the extent to which the units issued are able to achieve the goals set.

Meanwhile, according to SP.Hasibuan (1984) efficiency, quoting H. Emerson's statement, efficiency is the best comparison between input and output. In other words, efficiency is something we do with regards to optimal results by not wasting a lot of time in the process. Meanwhile, the meaning of accuracy is how close the value of the measurement results (estimated theta) to the actual value (true theta) or reference value. The closer to the reference value, the smaller the measurement error, this indicates that the measurement is more accurate.

If the CAT model is associated with efficiency, effectiveness and accuracy as described above, it can be hypothesized earlier that for effectiveness it is understood when the selection of items determined and selected by the system for test takers is in accordance with the ability of the test takers, so that items that are not in accordance with the test takers will not selected by the CAT model administration system. Therefore, this effectiveness will lead to a large number of questions raised by the system. From the data, it has been informed that the administration of the test with CBT will be given 40 items for each test taker, but with the administration of the test using the CAT model, items that are not in accordance with the test taker will not be selected by the system, so the items selected and raised by the system the system of each test taker will be less than the administration of the CBT model.

Furthermore, the relation with efficiency is in line with the effectiveness results above, that the number of items selected and raised by the system for each test taker with the CAT model is less, it will cause an accumulation of item response time when administering the

test using the CAT model will save more time compared to with the CBT model. Meanwhile, accuracy is understood that the error information obtained is between true theta and the estimated theta of the test taker. In other words, test administration using the CBT model will produce an estimate of the ability of the test taker as well as test administration using the CAT model. The difference between true theta and the estimated ability will result in a measurement error. This accuracy can be tested by calculating the MSE (mean square error) of each administrative model. And test administration with the CAT model will result in a smaller MSE than the CBT model, therefore the CAT model is more accurate than the CBT model.

3.2 Test Administration using CAT Model

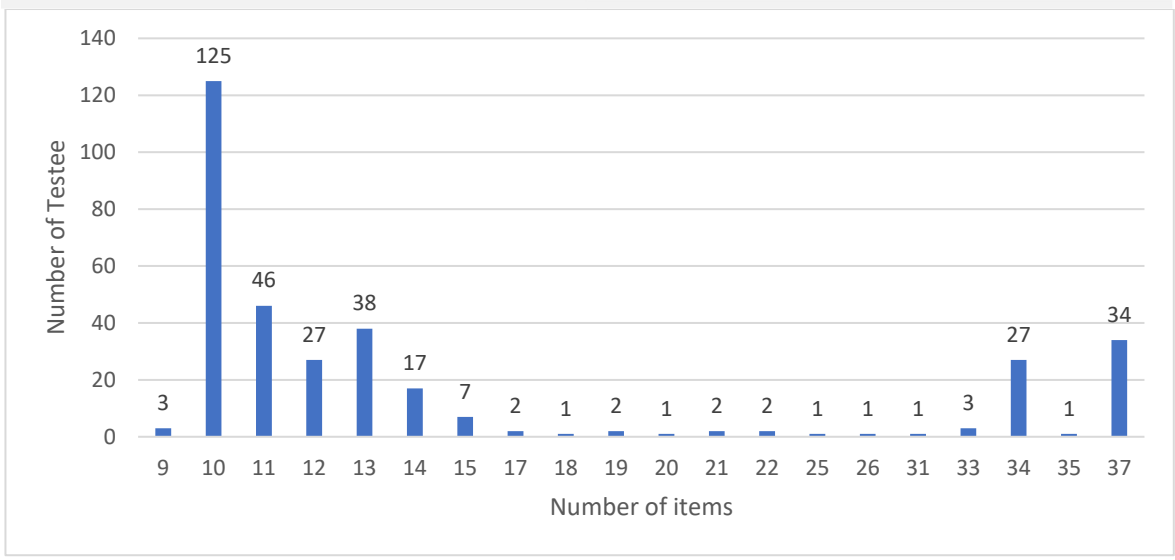
The procedure for administering the CAT model test has been described in the theoretical framework above. Therefore, in this study using the following procedures: 1) Item bank. The question bank used is statistical information on questions at the 2019 UNBK using mathematics subjects with a total of 40 items; 2) The selection of the first question, selected by the system randomly, and selected questions that have a moderate level of difficulty or with a difficulty level value between -0.5 to 0.5 on a logit scale; 3) Calculation of capability estimation, CAT model in capability estimation uses maximum likelihood approach and with step-sizing procedure in each capability determination in the early stages; 4) The selection of the following items, the CAT model in selecting the following items using maximum information is done by selecting items at each stage that have a value of b close to the current estimate of ; 5) stop the test on the test administration with the CAT model using a variable length, and therefore the CAT administration stops when the standard error measurement has reached the limit of less or equal to 0.4.

4. Result And Discussion

From the results of research processing, there are three things that are analyzed and discussed by the researcher, all of which are related to the objectives of this study. The topics discussed are 1) Efficiency analysis of the CAT model compared to the CBT model; 2) Analysis of the effectiveness between the CAT model compared to the CBT model; 3) Analysis of accuracy between results using the CAT model compared to the CBT model.

4.1 Analysis and discussion of the efficiency of the CAT model compared to CBT

As explained above, efficiency with the application of the CAT model is related to the large number of questions raised by the system for each participant. From the processing results can be shown in graph 1 below.



Source: primary data, analysis

Figure. 2. Comparison between the number of questions that appear with the number of test takers in the application of the CAT model.

In figure 1 above, it can be seen that there were 125 test takers who only got 10 questions. And the number of test takers who received 10 questions had the highest number of participants compared to the number of other questions. From graph 1 above, it can also be seen that the number of questions raised by the system was between 9 and 15 items. Therefore, it can be said that the CAT model system has been able to estimate the ability of the test takers with only 15 questions.

From figure 1 above, it can also be seen that there are many test takers who get the maximum number of questions or the number is the same as when the CBT model was applied. From the graph, there are 34 test takers who get 37 questions and 27 test takers get 34 questions, this is because at the beginning of the questions they can respond with answers that are always right or always wrong, so the method of estimating the ability of test takers with maximum likelihood is not able to estimating the ability of the test taker, so that the determination of the estimation of his ability using the step sizing method, where if the test taker still has a pattern of response responses from the beginning which is always correct, then the estimated ability is calculated by adding a value of 0.5 from the estimated value of the test taker's ability to the previous item. And if the test taker still has a pattern of responses, the answer from the beginning is always wrong, then the estimated ability is calculated by subtracting the value of 0.5 from the estimated value of the test taker's ability on the previous item.

Source: primary data, analysis

Figure 2. Variations in the number of questions with the CAT model test administration

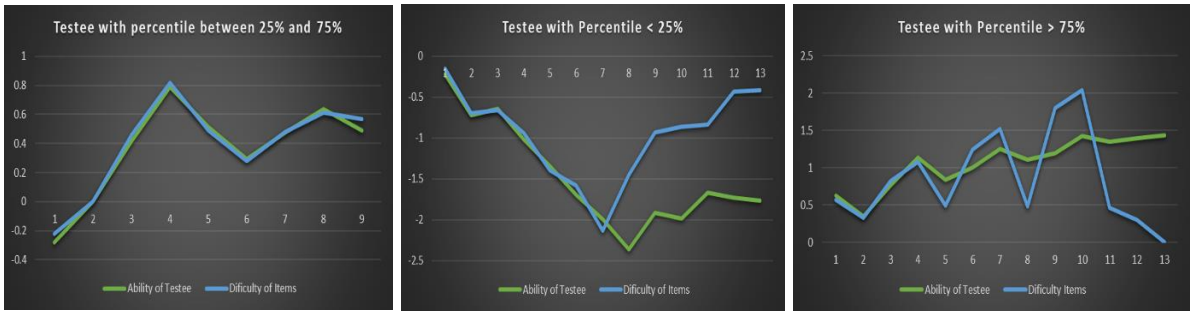
Figure 2 above with test administration using CAT shows that of the 341 selected sample data, there are 77.13% of test takers' data, which is enough to only take 15 questions, but the system is able to estimate the ability of the test takers. Figure 2 above also shows that only 19.35% of test takers have taken more than 31 questions. When compared with the test using the CBT model where each test taker will get the same number of items, namely 40 items, therefore the implementation of test administration with the CAT model is more efficient than the CBT model.

The next discussion is efficiency in terms of the length of time it takes to answer all the questions taken, then the CAT model will be more efficient than the CBT model. This can be proven by the large number of questions taken multiplied by the average response time of the items. From the 341 sample data, the average response time for the items can be calculated, which is 160.141 seconds, so it can be said that the test taker in working on one item takes 160.141 seconds. With the response time for each item, it can be said that most or as many as 77.13% of test takers have time to complete the 15 questions taken, which is 40 minutes. This is compared to the CBT model, where each test taker has to do the same number of questions, which is 40 questions and takes 120 minutes. So it can be concluded that the CAT model is 1/3 times more efficient in processing time than the CBT model.

4.2 Analysis and discussion of the effectivity of the CAT model compared to CBT

In the discussion of effectiveness, as explained in the methodology above, the selection of questions raised is adjusted to the ability of the test taker. Therefore, in the application of administration with the CAT model, the system will not select items that are not in accordance with the ability of the test taker. It is clear that the test model is said to be effective if the items given to the test taker are adjusted to the ability of the test taker. So, because the items are adjusted to appear, not all items are given to the test takers, so that each test taker will get a different question according to the ability of the taker and the number of questions answered by each test taker is not the same.

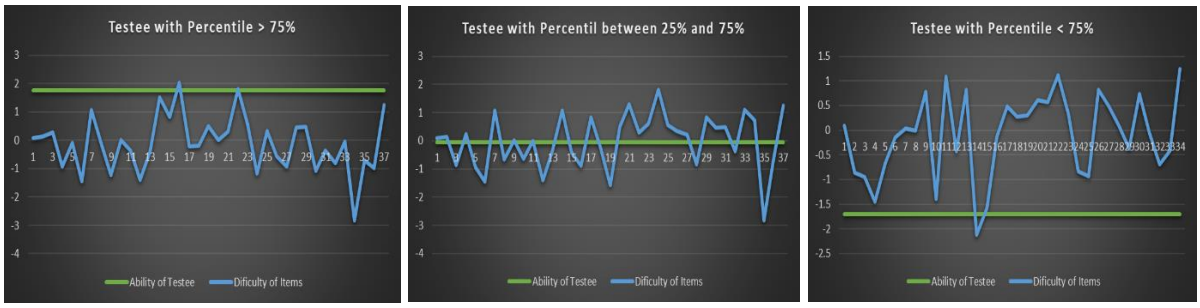
Figure 3 below is an example of the processing results of three participants with the administration of the CAT model test by describing the questions selected by the CAT model system according to the abilities of the participants.



Picture 3a **Picture 3b** **Picture 3c**

Source: primary data, analysis

Figure 3. Results of the effectiveness of the CAT model



Picture 4a **Picture 4b** **Picture 4c**

Source: primary data, analysis

Figure 4. Results of the effectiveness of the CBT model

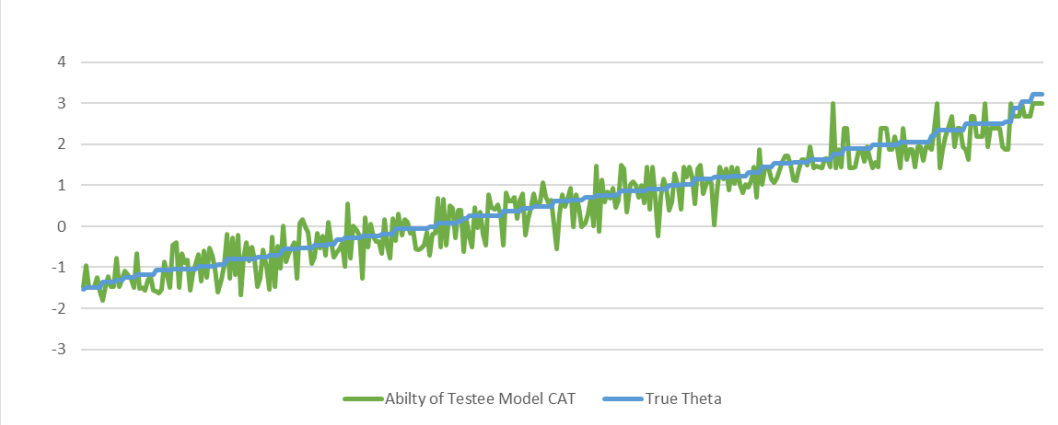
Figure 3 above shows that the items selected by the system are adjusted to the ability of the test taker. Figure 3a shows an example of a test taker with a percentile above 75, from the picture the ability tends to increase because each item is given the test taker responds with the correct answer. However, it is shown that the items that should appear are not available for the test taker, so that from number 5 to number 13, the level of item difficulty is not available with the increase in the estimated ability of the test taker.

Figure 3b shows an example of a test taker with a percentile between 25 - 75, from the picture it is shown that the items selected by the system are in accordance with the ability of the test taker. So it is clear the effectiveness of the selection of items by the CAT model system. Figure 3c shows an example of a test taker with a percentile below 25, from the picture the ability of the test taker decreases and from item number 8 onwards his ability increases, or the response of the test taker is correct. And the level of difficulty of the items given from numbers 8-13 also increases according to the increasing ability of the test taker.

Meanwhile, if we compare the test administration with the CBT model in Figure 4 above, it can be seen that the three pictures (4a,4b and 4c) of the items selected by the system do not match the ability of the test taker. Therefore, based on the analysis above, it is clear that the CAT model is more effective than the CBT model, and it will be more effective when the bank items owned by the CAT model are spread out for each participant's ability scale. If the item bank owned is not spread out for every ability of the taker, the test taker with the CAT model will be selected by the system for items far from the ability of the test taker, this is clearly seen from Figure 3a.

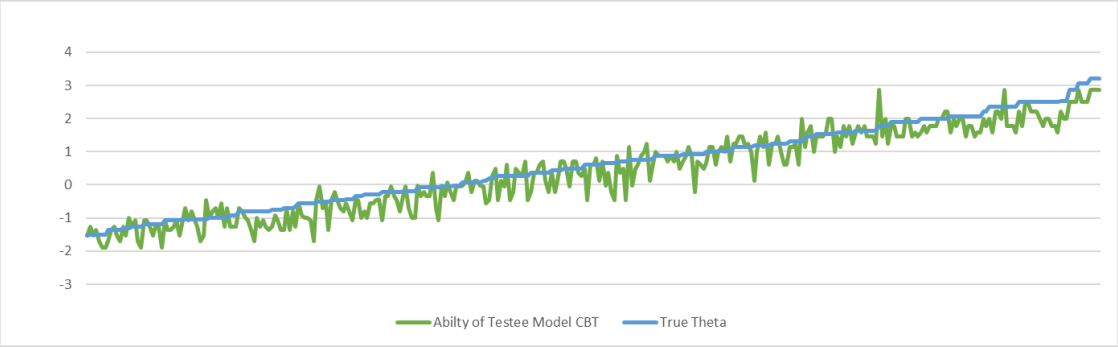
4.3 Analysis and discussion of the accuracy of the CAT model compared to CBT

In the discussion of accuracy as described above, the model is said to be accurate if it has a small measurement error. Therefore, in this analysis, the measurement error value will be compared with the MSE (Mean Square Error) test between the CBT model and the CAT model. To obtain a measurement error value such as equation 2 above, the reference value (true theta) in this study uses a reference value in the form of the average value of each test taker from the implementation of pre-UN and UN. So that the measurement error is obtained by calculating the difference between the reference value and the results of the estimated ability of the test taker.



Source: primary data, analysis

Figure 5. Comparison between the reference value and the estimated ability of the test-taker with the CAT model



Source: primary data, analysis

Figure 6. Comparison between the reference value and the estimated ability of the test-taker with the CBT model

From Figure 5 and Figure 6 above, it can be seen that the estimation of the ability of the test taker is close to the reference value, and the distance between the difference and the reference value is called error. If you look at it, there is no significant difference between the reference value and the results of the estimated ability of the test takers using both the CBT model and the CAT model.

From the results of the MSE test, it can be proven that by using the CAT model the MSE value is 0.158411619 while using the CBT model the MSE value is slightly greater, namely 0.160009318. so it can be concluded that although the difference in MSE values between the measurement of the test taker's estimated ability and the administration of the CBT and CAT models is very small, it can be said that the CAT model is more accurate than the CBT model.

5. Conclusions And Suggestions

5.1 Conclusions

From this research resulted in three conclusions. First, the CAT model is more efficient than the CBT model. It can be understood from the results of the study that the number of questions taken by test takers with the CAT model is less than the PPT model, so the time required is 1/3 times more efficient than the CBT model. The CAT model can have the same number of questions as the CBT model if the question bank owned is not spread out in variations of the participant's ability scale. In addition, the consistency of participants' answers will affect the efficiency of the CAT model. Second, the CAT model is more effective than the CBT model. It can be understood from the results of the study that the items selected by the system using the CAT model and taken by the test taker are adjusted to the ability of the taker, so that the information obtained from the estimation of the test taker's ability is more optimal than the CBT model. Third, the MSE test results between the CAT and CBT models yielded the same value, although the CAT model was smaller than the

CBT model. It can be understood from the results of the study that the CAT model has more accuracy than the CBT.

5.2 Suggestions

1) The central government in this case the Ministry of Education and Culture needs to think about an Achievement Test policy for students to use the CAT model as an alternative to the CBT model.
2) Stakeholders Education, especially in schools, can apply the CAT model as an assessment model in schools.
3) For academics, it is necessary to conduct further and in-depth research, considering that the CAT model is a new assessment model and has not been widely applied in Indonesia.

References

Apa yang Harus Dilakukan Guru dalam Mengembangkan Silabus.2007. (http://rbaryans.wordpress.com/2007/07/27/, diakses 28 Juli 2010).

Blais, J. & Raiche, G. 2002. *Some Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Makalah disajikan pada 11th International Objective Measurement Workshop, New Orleans, April 2002.

Bunderson, C.V., D. K. Inouye, and J.B. Olsen 1989. The Four Generations of Computerized Educational Measurement. Dalam Robert L. Linn. *Educational Measurement*. 3rd ed. New York: American Council on Education & Macmillan Publishing Company.

Departemen Pendidikan Nasional.2009. Peraturan Menteri Pendidikan Nasional No. 75 Tahun 2009 Tentang Ujian Nasional SMP/MTs, SMPLB, SMA/MA, SMALB dan SMK Tahun Pelajaran 2009/2010.

Departemen Pendidikan Nasional.2009. Prosedur Operasi Standar (POS) UN SMP, MTs, SMPLB, SMA, MA, SMALB dan SMK Tahun Pelajaran 2009/2010

Embretson, S.E, and S.P.Reise.2000. *Item Response Theory for Psychologist*. New Jersey : Lawrence Erlbaum Associates, Inc.

Emerson, H., (1986). Sistem Birokrasi Pemerintah. CV. Mas Gunung, Agung Jakarta, hal. 16

Hambleton, R.K., H. Swaminathan, and H.J.Rogers .1991. *Fundamental of Item Response Theory*. Volume 2. California: Sage Publications, Inc.

Hasibuan, S.P. 1984. *Manajemen Dasar dan Suatu Pengantar*. Jakarta: Haji Masagung.

Hidayat. 1986. Teori Efektifitas Dalam Kinerja Karyawan. Gajah Mada University Press. Yogyakarta

Hornke, L.F. 2000. Item Response Times in Computerized Adaptive Testing. *Psicológica*. 21, 175-178.

Kerja dan keletihan.2010. (*http://sitinurmala-mala.blogspot.com/2010/03/kerja-dan-keletihan.html*, diakses 28 Juli 2010)

Leung, C., H.Chang, and K.Hau .2005. Computerized Adaptive Testing: A Mixture Item Selection Approach for Constrained Situations. *British Journal of Mathematical & Statistical Psychology*, Nov 2005, 58, Proquest Psychology Journals, pp 239.

Mahmudi, 2010. Manajemen Kinerja Sektor Publik. Penerbit UUP STIM YKPN, Yogakarta

Sumarna Surapranata. (2009). Analisis Validitas, Reliabilitas, dan Interpretasi Hasil Tes. Bandung: PT. Remaja Rosda Karya.

Thissen, D., & R. J. Mislevy .1990. Testing Algorithms. Dalam H. Wainer, N.J. Dorans, R. Flugher, & B.F. Green, *Computerized Adaptive Testing: a Primer*. New Jersey: Lawrance Erlbaum Associates, Publishers.

Wainer, H. 1990. Introduction and History. Dalam H. Wainer, N.J. Dorans, R. Flugher, & B.F. Green. *Computerized Adaptive Testing: a Primer*. New Jersey: Lawrance Erlbaum Associates, Publishers.