Computer Adaptive Test as The Appropriate Model to Assess Physics Achievement in 21st Century

Edi Istiyono
Educational Research and Evaluation
Graduate School of Yogyakarta State University
Yogyakarta, Indonesia
edi_istiyono@uny.ac.id

***Abstract*—Most paper-based assessment tests require students to answer all of the test items and take much time for feedbacks. Therefore, more practical, efficient, and accurate assessment model needs to be developed. This study aims to describe (1) the superiority of Computer Adaptive Test (CAT) compared to Computer Based Test (CBT) and Paper and Pencil Test (PPT) and (2) the feasibility of CAT to measure achievement in physics. The study was conducted using the following procedures: comparing Item Respond Theory (IRT) with Classical Test Theory (CTT), comparing CAT with CBT and PPT, and conducting surveys to assess CAT performance to measure physics achievement. The result shows that, in case of education in the 21st century, (1) CAT is more superior to assess students' achievement in physics than CBT and PPT; and (2) CAT is more feasible to measure physics achievement. Therefore, CAT is very appropriate to assess physics learning nowadays.**

Keywords*— IRT, Computer Adaptive Test, assessment, Physics' achievement

I. Introduction

Classroom learning is conducted to achieve specific learning objectives. As educator, teacher needs to perform assessment in order to achieve learning objectives. Assessment plays an important role as it allows teacher to monitor and evaluate student's learning process, willingness, and improvement of learning outcomes in a continuous manner [1]. In addition, assessment is also considered to have a powerful means of informing teacher on student's level of thought in relation to learning objectives [2]. Condidering the importance of student assessment in learning, teachers need to develop an effective assessment instruments in order to achieve specific learning objectives.

Effective assessment begins with identifying the right learning objectives in order to understand the key scientific ideas and to be able to utilize them through scientific practices [2]. Upon identifying the learning objectives, the next step is to monitor student's progress in specific objectives. The last step is to assess how much of the objectives have been achieved [3]. Based on the steps, a conclusion can be drawn up, i.e. that assessment means an activity of collecting information on students to assist teacher in decision making [4]. Classroom assessment was conducted by observation, performance or project rating, and paper and pencil test [3]. In addition, assessment may also be in the form of homework, quiz, test, and group activity [5]. The form of classroom assessment in specified according to the condition of the students, teacher, and facilities and infrastructure used to support learning.

Analysis of assessment instrument is necessary in order to obtain a good testing. The analysis of test includes the analysis of the characteristics of measurement instrument used and the analysis of test participants' capabilities. There are two analysis of test, i.e. Classical Test Theory (CTT) and Item Responses Theory (IRT). CTT is developed by combining the concept of error and the concept of correlation [6]. CTT perceives score as the sum of true score and false score. In addition, CTT is relatively useful in describing how measurement error may affect on observed score [7]. CTT model assumes specific conditions, in which when the assumption is logical then the conclusion drawn up from the model is also logical. There are three item parameters to estimate, i.e. level of difficulty, discrimination index, and assumption [8]. Level of difficulty is the proportion of students with correct answers. Item discrimination is the correlation between question item score and the total score, i.e. also known as biserial point correlation.

Modern test theory is also known as item response theory (IRT). This theory uses mathematical models to relate question item characteristics and respondent's ability. The correlation is described by item characteristic curve. Mathematical model in IRT means that the probability of subjects answering items correctly depends on the corresponding subject's ability and item characteristics [9]. IRT approach is utilized to analyze test using the principle of relativity and probability. Relativity is defined as student's relative ability towards question items, while probability is means that student's ability in answering questions is depends on the corresponding student's ability and item characteristics.

In conventional assessment system, test for students is conducted using paper and pencil (PPT) method [10]. Many studies have been conducted toward the effectiveness of both conventional and modern assessment. Some studies show that PPT generates higher score than CBT [11]. This is due to a number of factors, one of which is socio-economic factor [12], [13], [14]. In reality, the use of PPT system is dominant.

The primary goal of education reformation is to provide students with 21st century skills required to respond to global challenges. IT means many new knowledge are invented [15]. The development of 21st century skills should be balanced with the development of students and academicians. It is to allow the achievement of learning objectives and good, correct utilization of science and technology.

Technology plays an important role; i. e. in this case is to prepare 21st century workforce [16].One of the products of 21st century's advancement is digital media. It has transformed the operation of all aspects, from books to tablets, and from physical interaction to virtual collaboration. In the 21st century, assessment should take into account the quality standards for content, process, and assessment in order to generate critical and creative human resources capable of responding to all 21st century's challenges and issues [17]. To ensure that the improvement of the idea of 21st century skills, new assessment is necessary—one that can accurately measure richer learning and more complex tasks [18]. As such, an assessment model that suits the learning paradigm and model of the 21st century is required. To respond 21st century challenges, the majority of assessments will be technology-based assessment [19].

Computer-based test (CBT) is the answer for level of security and integrity of test results. CBT is defined as the utilization of information technology for assessment activities. It allows the teacher to write and schedule quizzes and tests through computer system, where the responses are recorded and assessed electronically. CBT is a computer-assisted assessment system aimed at assisting teacher in conducting assessment, including in scoring, test implementation, and the effectiveness and efficiency of such implementation [20]. CBT is first introduced more than 60 years ago [21]. One of CBT assessment programs was conducted in 1991 for certified Novell engineers [22]. In CBT, candidate sits in front of a computer; questions are displayed on the monitor and answers are to be given using keyboard and mouse.

CBT software has a computer core functioning as a server (question supplier and storage), and teacher uploads questions and the correct answers into the system [23]. The most commonly used type of CBT is linear CBT, i.e. fixed length assessment computerized, which presents the same number of items for each test-taker, and the score depends on the number of questions answered correctly. The model and variant also has a number of benefits, such as flexible test administration, automatic score reporting, and utilization of new item format [24]. CBT test items are the same as in paper-based test, except that they are presented in digital format. CBT is also the global industry's brand, which includes different types of assessment, objectives, test design, and types of item adjusted to the accountability of education [22]. The utilization of CBT can reduce malpractices by teacher in assessment [25]. Besides, the duration of test also improves in terms of efficiency. In addition, computers can help save time for briefing [26].

Another assessment system that utilizes technology is Computerized Adaptive Testing (CAT). CAT is a specialized computer-based test. Each test-taker is given a unique test, in which the items are adjusted to match the respective test-taker's ability [27]. CAT is a class of delivery of algorithm from a test aimed at enabling test-takers to achieve higher measurement accuracy and efficient test delivery. Each test comes with individual set of test items selected in sequence and adjusted to the current estimate or the respective test-taker's ability [28]. CAT is based on IRT or decision theory [29], especially for the item selection. As CAT develops, test begins with less difficult questions; students answer these questions and the computer will instantly score the answers. Compared to other test methods, CAT requires shorter time to predict test-taker's ability.

In CAT, test will end once it is considered sufficient. CAT will end when maximum test length is achieved and measurement of ability has been estimated with adequate precision [30]. It makes CAT superior with regards to the effectiveness and efficiency of test in measuring student's ability. Besides, CAT stores a very large number of questions (items), known as question bank (item bank). The question bank has been calibrated using IRT method, allowing CAT to generate valid and reliable test to identify student's achievement [31]. In adaptive test, question item set is adjusted to test-taker's age. In Indonesia, CAT has been developed and researches on it have been conducted [32], [33], [34]. CAT can also be used in large-scale tests. Indonesia has utilized CAT in the 2017 CPNS (Civil Servant Candidates) selection [35]. CAT is the evidence of the improvement of test quality and computer technology.

II. Method

The research step included were: (1) determining the superiority of CAT compared to CBT and PPT by: (a) comparing IRT and CTT test theory, (b) comparing test media among CAT, Computer-Based Test (CBT) and Paper and Pencil Test (PPT). The next step is (2) expert judgments toward the developed CAT to survey teachers and students from 10 senior high schools in Yogyakarta that have been used for CAT applications trial to assess students' physics achievement. The respondents involved were 10 physics teachers and 155 students, while the instruments used were rewiewer sheet and questionnaire. The scoring model was a polytomus of 4 categories with a total score of 100.

III. Results and Discussions

A. *CAT is superior than CBT and PPT*

First, you need to confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

1) IRT is more suitable to employ than CTT

Classical test theory (CTT) has developed widely in psychological testing and education. CTT approach for item analysis is based on correlational data, generally involves maximizing Cronbach's alpha and selecting items accoridung to the factor loading using exploration factor analysis [36]. The comparison between CTT and IRT is shown by Table 1.

TABLE 1. Comparing IRT and CTT

No	IRT	CTT
1	The estimation of ability is not bias toward item characteristics although the sample was not representative yet.	The estimation of ability is bias toward item characteristics and depend on the sample.
2	The item standard error measurement (SEM) is different for each item,	The item standard error measurement (EM) employed to all items on a specific

	alhtough there is also a total SEM.	population. SEM is generally used for the test.
3	The score given is based on the difficylty level, discrimination index and pseudo guessing of the item, so is the ability.	The score given is based on the difficylty level. The score is basically calculting the number of right and wrong answers, therefore, students with the same number of right and wrong answer considered to have the same ability.

Classical test theory is superior in terms of the concept that is easy to understand and use, therefore, CTT is preferable in many cases. However, CTT also has several limitations and differences with IRT (see Table 1). First, the estimate of item difficulty and discrimination depends on specific groups of test-takers who complete the test. CTT is s dependent sample (items are dependent to the samples employed), characteristics of the evaluator/researcher and the test is inseparable, and the test is more test-oriented rather than item-oriented [37]. CTT provides only one standard error measurement (SEM) that applied for all item in the test, therefore it is more test-orinted. Second , the estimate of test-taker's ability depends on the specific test items given. The result of test analyzed using classical test theory depends on the characteristics of students and the characteristics of items, making it unfeasible to correctly measure student's ability. Moreover, generation of size in CTT that taps only a small part of the underlying construction [38]. In addition, CTT is also more sample and dependent item [39], [37], [40]. MacDonald [41] illustrated dependency in CTT, i.e. if a test consists of relatively easy items, the test-taker's statistics (in this case, the observed score) will be relatively high, resulting in an impression that the test-takers possess high level of ability. On the other hand, if the test consists of relatively difficult items, the test-taker's statistics will be relatively low, resulting in an impression that the test-takers possess only low level of ability. When test-takers who complete the test possess high level of ability, the p value of item will also be high, signifying that the corresponding item is easy.

Third, CTT focuses on the information of test level, therefore, CTT only provides the estimation of the reliability of single entirety (Cronbach's alpha). Such weakness of CTT triggered the emergence of new, more suitable theory, i.e. modern test or item response theory (IRT). IRT has primary benefits. First, in sampling error, item parameter does not depend on the sample's level of ability, i.e. the samples are invariant. Second, the score achieved by an individual does not depend on specific item samples responded by the individual. Third, IRT focuses on level of item information, thus IRT is expected to be able to eliminate the limitation of CTT. Moreover, IRT has individual SEM for each item, provides index from informative item contribution, and allows the elimination of excessive or non-discriminative items. Thus, IRT is able provides estimates on respondent's level of ability and difficulty in differentiating items [42].

IRT provides additional information that is feasible to check individual items in a more detailed manner than CTT. IRT information function shows the most useful items in specific construction. Item information function is a combination of the ability to differentiate items and the level of difficulty. Item information function facilitates reliable size to explore across the underlying construction. Low information function may signify that a specific item may be unsuitable. The utilization of item response theory is particularly useful to accurately measure student's level of ability. The fundamental difference between CTT and IRT lies on the scoring invariance, where modern scoring is invariant (unchanging or fixed) against test items and test-takers. The invariance of test item parameters across test-taker groups constitutes IRT's most prominent characteristic [11]. Thus, IRT comes as an alternative approach that is able to analyze a test and interpret student's level of ability against the corresponding test. Based on the explanation, IRT is more suitable to employ than CTT.

2) CAT is more practical and appropriate for assessing achievement in physics than CBT and PPT

In traditional paradigm, test is conducted on paper, i.e. paper-pencil test (PPT). PPT approach, according to Thompson, has a number of issues; the most prominent is inefficiency [43]. Comparison among CAT, CBT and PPT wil be given in Table 2.

TABLE 2. COMPARISON AMONG CAT, CBT AND PPT

No	CAT	CBT	PPT
1	CAT analysis is using IRT	CBT analysis is using IRT	PPT analysis is using CTT
2	The item on CAT is adjusted to students' level of ability	The item in CBT is not take in to account of students' level of ability	All item in PPT has be finished by the students
3	In CAT, the next item presented is depend on students answer of the previous item.	In CBT, the item appeared randomly	In PPT the sequence of the items is choosen indenpendently by the test taker

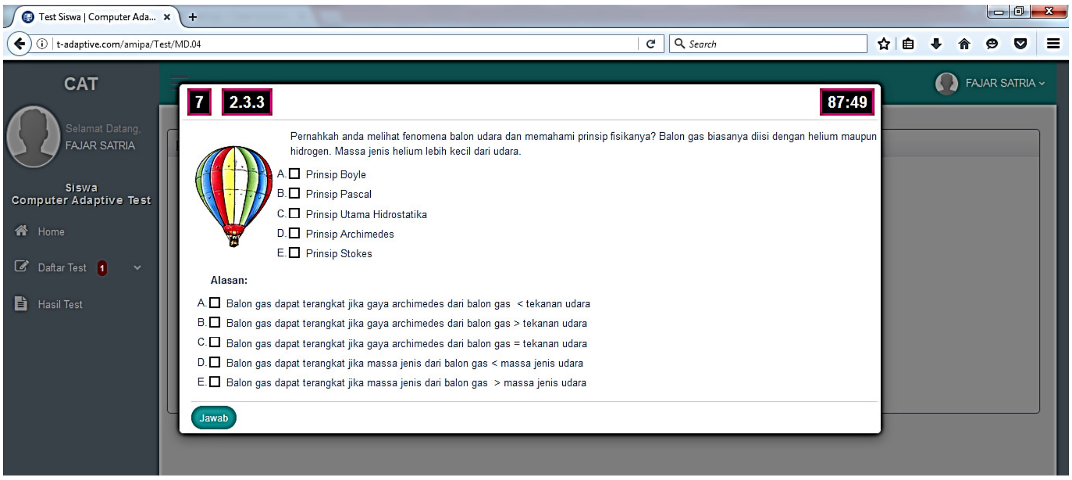
Based on the Table 2, there are some differences among CAT, CBT and PPT. PPT analysis using CTT demonstrated guessing effect issue, i.e. whether the answer is random or not. Classic methods of correcting assumptions do exist, but they are actually more biased than not applying correction at all. Another issue is that the majority of question items in PPT come with moderate level of difficulty. This sacrifices test-takers with high or low level of ability. Test-takers are measured using far lower precision.

CAT is based on IRT theory that provides model-based approach to predict guessing [44]. IRT provides the more powerful procedure to correlate and equate, ensuring a stable scale with proportional score across all test-takers. This will improve test duration effectiveness as well as minimize the error of each test-taker, conditional to the respective test-taker's level of ability [45]. In addition, the utilization of CAT is also beneficial to process large-scale data. Analysis of large-scale data generates better results [46], and CAT makes such analysis easier to perform.

Basically, the question items presented in both CAT, CBT and PPT are the same, but the questions in CBT are randomly presented in non-printed form, while PPT provides a prnted one. In CBT and PPT, each test-taker is to answer the entire question items. Such type of test cannot accurately measure the test-taker's level of ability. On the contrary, CAT is an ideal test setting can accurately measure the test-taker's level of ability, where the test items' level of difficulty is adjusted to the test-taker's level of ability. If the students' anwers the question correctly, the next question will have higher level of difficulty. In contrary, if the students' anwers the question wrongly, the next question will have lower level of difficulty. CAT is a computer-assisted

B. *CAT is more feasible to measure achievement in physics*

In the field of education, computer can be utilized to deliver learning materials and measure achievement [47]. CAT provides only questions with level of difficulty suited to the test-takers' level of ability, allowing questions to be presented shorter by 50% than PPT but with equivalent or better precision of measurement [48]. The higher the test-takers' level of ability, there are more number of difficult questions to complete. CAT displays only questions suited to the test-takers' level of ability, allowing for more efficient test duration.



Test Siswa | Computer Ada...

t-adaptive.com/amipa/Test/MD.04

7 2.3.3 87:49

Pernahkah anda melihat fenomena balon udara dan memahami prinsip fisikanya? Balon gas biasanya diisi dengan helium maupun hidrogen. Massa jenis helium lebih kecil dari udara.

A. □ Prinsip Boyle
B. □ Prinsip Pascal
C. □ Prinsip Utama Hidrostatika
D. □ Prinsip Archimedes
E. □ Prinsip Stokes

Alasan:

A. □ Balon gas dapat terangkat jika gaya archimedes dari balon gas < tekanan udara
B. □ Balon gas dapat terangkat jika gaya archimedes dari balon gas > tekanan udara
C. □ Balon gas dapat terangkat jika gaya archimedes dari balon gas = tekanan udara
D. □ Balon gas dapat terangkat jika massa jenis dari balon gas < massa jenis udara
E. □ Balon gas dapat terangkat jika massa jenis dari balon gas > massa jenis udara

Jawab

Fig 1. The Example of Item Preview

According to the experts on media and assessment, the preview of CAT obtain an averange score that is appropriate. In addtion, the experts stated that the effectivity aspect of the implementation on CAT obtain an averange score that is excellent and feasible to be used.

TABLE 3. TEACHER AND STUDENT'S RESPONSES TO THE FEASIBILITY OF CAT IN ASSESSING STUDENT'S ACHIEVEMENT IN PHYSICS

No	Response	Percentage of IRT-Based CAT Feasibility (%)
1	Student	78.33
2	Teacher	86.81

It can be seen from table 3, students and teachers perception on CAT feasibility in measuring student's achievement in Physics is 78.33% and 86.81%, respectively, meaning that CAT is highly feasible. It signifies that the utilization of IRT-Based CAT to measure achievement in physics is highly feasible and applicable to measure student's level of ability. In addition, the characteristics of CAT can also result in positive effect towards student's performance. In CAT test, student's score is presented immediately upon completion of the question items. CAT allows immediate viewing of score as the computer immediately calculates the score and estimates the respective student's level of ability upon answering the question items given [49]. Direct feedback in the form of test score results is in positive effect towards student's performance, regardless of their level of ability [50]. CAT provides challenges suited to individual student. Students with lower level of ability do not feel discouraged when receiving questions, while students with higher level of ability can enjoy difficult questions according to their level of ability [38]. IRT-Based CAT test-takers are measured using the same level of precision although they may potentially receive different items. It is why psychometric perspective perceives CAT as highly fair. Adaptive CAT, in which questions are given in line with test-takers' level of ability, generates accurate test results. CAT is used as an updated medium of assessment in line with the development of item response theory.As such, IRT-Based CAT is more feasible to measure achievement in physics. Based on the above description, it is reasonable to say that IRT-Based CAT is very appropriate to implement for assessing physics learning nowadays.

IV. CONCLUSIONS

Based on the analysis, several conclusions can be drawn up, i.e.:

1. CAT is superior for assessing achievement in physics compared to CBT and PPT according to the facts that: (a) IRT is more suitable to be employed compared to CTT and (b) CAT is more practical and approriate for assessing physics achievement compared to CBT and PPT.
2. CAT is more feasible to measure physics achievement since the score of performance assessment from teachers and students were respectively 78.33% and 86.81%. Besides, the results of experts judgement was included in a good category. Therefore, CAT is very

appropriate to be implemented for assessing physics learning nowadays.

References

[1] Regulation of Minister of Education and Culture of the Republic of Indonesia Number 23 of 2016 on Standards for Education Assessment.

[2] Kloser, Matthew., H. Borko., J. F. Martinez, B. Stecher, and R. Luskin, “Evidence of middle school science assessment practice from classroom-based portfolios,” Science Education Vol. 101 (2), 2016, https://doi.org/10.1002/sce.21256.

[3] M. D. Miller, R. L. Linn., and N. E. Grnlund, “Measurement and Assessment in Teaching,” United States of America: Pearson Education, 2009.

[4] L. W. Anderson, “Assessment Enhancing the Quality of Teacher Decision Making, ”New Jersey: Taylor & Francais e-Library, 2008.

[5] H. M. Bush, J. Daddysman., and R. Charnigo, “Improving outcomes with bloom’s taxonomy: from statistics education to research partnerships,” Biometric & Biostatistics5 (4), 2014. http://dx.doi.org/10.472/2155-6180.1000e130

[6] N. J. Salkind “Encyclopedia of Measurement and Statistics Volume 1,” California: SAGE Publication, Inc, 2007.

[7] M. J. Allen and W. M. Yen, “Introduction to Measurement Theory,” United States: Waveland Press, Inc, 1979.

[8] D. Mardapi, “Analisis Butir dengan Teori Tes Klasik dan Teori Respons Butir, ”Jurnal Kependidikan vol. 28, 1998, pp.15-34.

[9] H. Retnawati, “Teori Respons Butir dan Peneraannya,” Yogyakarta: Nuha Medika, 2014.

[10] S. Kirschner, A. Borowski, H. E. Fischer, J. Gess-Newsome, and C. von Aufschnaiter, ”Developing and evaluating a paper-and-pencil test to assess components of physics teachers’ pedagogical content knowledge,” International Journal of Science Education, 38(8), 2016, pp.1343–1372. https://doi.org/10.1080/09500693.2016.1190479

[11] Y. P. Chua, and Z. M. Don, ”Effects of computer-based educational achievement test on test performance and test takers’ motivation” Computers in Human Behavior, vol. 29(5), 2013, pp.1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

[12] H. Jeong, ”A comparative study of scores on computer-based tests and paper-based tests” Behaviour and Information Technology, 33(4), 2014, pp.410–422. https://doi.org/10.1080/0144929X.2012.710647

[13] J. J. Shaftel, R. Belton-Kocher, D. D. Glasnapp, J. J. Poggio, E. Belton-kocher, D. D. Glasnapp, and J. J. Poggio, ”The impact of language characteristics in mathematics items on the performance of English language learners and students with disabilities,” Educational Assessment, 11(2), 2006, pp.105–126. https://doi.org/10.1207/s15326977ea1102

[14] B. Offir, Y. Lev, and I. Barth, “Using Interaction Content Analysis Instruments to Assess Distance Learning,” Computers in the Schools, 18(2), 2002, pp.27–41. https://doi.org/10.1300/J025v18n02

[15] R. W. Bybee and B. Fuchs, “Preparing the 21st Century Workforce: A New Reform in Science and Technology Education,” Journal of Research in Science Teaching, 2006.

[16] Common Sense, “Digital Literacy and Citizenship in the 21st Century,” Media White Paper, 2009.

[17] S. Tishkovskaya and G. A. Lancaster, “Statistical Education in the 21st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform,” Journal of Statistics Education, 2017.

[18] A. Rotherham and D. Willingham, “21 Century: To work, the 21st century skills movement will require keen attention to curriculum, teacher quality, and assessment,” Educational Leadership, 2009.

[19] E. Silva, “Measuring skills for the 21st century,” Washington, DC: Education Sector, 2008.

[20] Muniarti, “Computer based test (CBT) sebagai alternatif instrumen evaluasi pembelajaran,” Surakarta: UNS, 2017.

[21] T. O. Oluwatosin and D. D. Samson, “Computer-Based Test: Security and Result Integrity,” International Journal of Computer and Information Technology, 2012.

[22] M. R. Luecht and S. G. Sireci, “A Review of Models for Computer-Based Testing,” The College Board, 2011.

[23] M. Ajinaja, “The design and implementation of a Computer Basewd Testing Using Component-Based Software Engineering,” International Journal Of Computer Science and Technology. Vol. 8 Issue 1, 2017, ISSN: 0976-8491.

[24] Ogunlade and Olafare, “Students’ Charactheristic As Predictors Of Their Perceptions The Effectiveness Of Computer-Based Test In Nigerian Universities,” Pakistan Journal of Education, 28(2), 2011.

[25] Abubakar and Adebayo, “Using computer based test method for the conduct of examination in Nigeria: Prospect, Challenges, and strategies,” Mediteranium Journal Of Social Sciences, 5(2), 2014, pp.47-46. https://doi.org/10,5901/mjss.2014.v5n2p47

[26] P. Black and D. Wiliam, “Inside the black box: Raising standards through classroom assessment,” Phi Delta Kappan Online . Retrieved March 20, 2018 from: http://pdkintl.org/kappan/kbla9810.htm

[27] Linden and Glas, “Computerized Adaptive Testing: Theory and Practice, Boston: Kluwer, 2000, pp.163-182.

[28] H. Jiao, G. Macready, J. Liu, and Y. Cho, “A Mixture Rasch Model – Based Computerized Adaptive Test for Latent Class Identification,” 2012. https://doi.org/10.1177/0146621612450068.

[29] Economides & Roupas, “Overexposure and Underexposure of Item In Computerized Adaptive Testing,” Measurement and Research Departement Report, 2007.

[30] J. M. Linacre, “Computer-Adaptive Testing: A Methodology Whose Time Has Come,” in S. Chae, U. Kang, E. Jeon & J. M. Linacre (Eds), Development of Computerized Middle School Achievement Test (in Korean) Seoul, South Korea: Komesa Press, 2000.

[31] S. Kirschner, A. Borowski, H. E. Fischer, J. Gess-Newsome, and C. von Aufschnaiter, “Developing and evaluating a paper-and-pencil test to assess components of physics teachers’ pedagogical content knowledge,” International Journal of Science Education, 38(8), 2016, pp.1343–1372. https://doi.org/10.1080/09500693.2016.1190479

[32] Haryanto, “Pengembangan Computerized Adaptive Testing (CAT) dengan Algoritma Logika Fuzzy,” Jurnal Penelitian dan Evaluasi Pendidikan, 15(1), 2011, pp.47-70.

[33] Winarno, “Pengembangan Computerized Adaptive Testing (CAT) Menggunakan Metode Pohon Segitiga Keputusan,” Jurnal Penelitian dan Evaluasi Pendidikan, 16(2), 2012, pp.574-592.

[34] E. Istiyono , D. Mardapi, and Suparno, “Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PhyHOTS) Peserta Didik SMA,” Jurnal Penelitian dan Evaluasi Pendidikan, 2014.

[35] “Apakah CAT CPNS Kemenkumah 2017, Ini Penjelasan Lengkapnya,” Tribun Jogja, 5 September 2017.

[36] J. Singh, “Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example,” Journal of Business Research 57, 2004, pp.184-208.

[37] R. K. Hambleton, Swaminathan and J. Rogers, “Fundamentals of Item Response Theory,” California: SAGE Publication, Inc, 1991.

[38] R. B. Fletcher and J. A. Hatti, “An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model,” Psychology of Sport and Exercise, 5, 2004, pp. 423-446.

[39] X. Fan, “Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Response Person Statistics,” Educational and Psychological Measurement, 58 (3), 1998, pp.357-381

[40] F. M. Lord, ”Aplications of Item Response Theory to Practical Testing Problems,” New Jersey: Lawrence Erlbaum Associates, Publishers, 1980.

[41] P. Macdonald, and S. V. Paunonen, “A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory Versus Classical Test Theory,” Rducational and Physcological Measurement, 62(6), 2002, pp.921-943. https://doi.org/.10.1177/0013164402238082

[42] B. Pollard, D. Dixon, P. Dieppe, and M. Johnston, “Measuring The ICF Components of Impairment, Activity Limitaion And Participation Restriction: An Item Analysis Using Classical Test Theory and Item Response Theory,” Health and Quality of Lilfe Outcomes, 2009. https://doi.org/10.1186/1477-7525-7-41

[43] T. Davey,“Practical considerations in computer- based testing. ETS Research & Development Divison,” 2011. Available, retrieved on: 22 March 2018. https://www.ets.org/Media/Research/pdf/CBT-2011.pdf .

[44] A. Bartlett, "Preparing pre service teachers to implement performance assessment and technology through electronic portofolios," Action in Teacher Education, 24:1, 2002, pp.90-97. http://dxdoi.10.1080/01626620.2002.10463270

[45] R. K. Hambleton, F. Robin, and D. Xing, "Item Response Models for the Analysis of Educational and Psychological Test Data," in H. E. Tinsley, and S. D. Brown, Handbook of applied multivariate statistics and mathematical modeling, San Diego, CA: Academic Press, 2000, pp.553-581.

[46] Pavlo, Paulson, Rasin, Abadi, Je Witt, Madden, and Stonebraker, "A Comparison of Approaches to Large-Scale Data Analysis," Proceedings of The 2009 ACM SIGMOD International Conferences on Management of Data, 2009, pp. 165-178. https://10.1145/1559845.1559865

[47] D. Karahoca, A. Karahoca, Erdogdu, Uzunboylu, and Gungor, "Computer Based Testing for E-learning: Evaluation of Question Classification for Computer Adaptive Testing," 2009.

[48] J. Liu, "Comparing multi-dimensional and uni-dimensional computer adaptive strategies in psychological and health assessment," Dissertation of University of Pittsburgh, 2007.

[49] A. Santoso, "Pengembangan Computerized Adaptive Testing untuk Mengukur Hasil Belajar Mahasiswa Universitas Terbuka," Jurnal Penelitian dan Evaluasi Pendidikan, 14(1), 2010, pp.62-83.

[50] G. Ling, Y. Attali, B. Finn, and E. A. Stone, "Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test?" Applied Psychological Measurement, 41(7), 2017, pp.495–511. https://doi.org/10.1177/0146621617707556