Journal of Education and e-Learning Research
Vol. 12, No. 2, 124-144, 2025
ISSN(E) 2410-9991 / ISSN(P) 2518-0169
DOI: 10.20448/jeelr.v12i2.6677

The challenges of implementing computerized adaptive testing in Indonesia

Okky Riswandha Imawan[1,2]✉
Heri Retnawati[1]
Haryanto[1]
Raoda Ismail[1,2]


(✉ Corresponding Author)

[1]*Educational Research and Evaluation, Universitas Negeri Yogyakarta, Special Region of Yogyakarta, Indonesia.*
[2]*Mathematics Education, Universitas Cenderawasih, Papua, Indonesia.*
[1,2]*Email: okkyriswandha.2021@student.uny.ac.id*
[1]*Email: heri_retnawati@uny.ac.id*
[1]*Email: haryanto@uny.ac.id*
[1,2]*Email: raodaismail26@gmail.com*

Abstract

This study explores the challenges of implementing computerized adaptive testing (CAT) for mathematics assessment among prospective elementary school teachers in Indonesia. It aims to describe (1) assessment practices of mathematics lecturers and (2) challenges in adopting CAT. Using a qualitative phenomenological approach, data were collected from 17 mathematics lecturers across Indonesia through questionnaires followed by in-depth interviews with four lecturers unfamiliar with or struggling to use CAT. Findings indicate that paper-and-pencil tests and computer- based testing (CBT) dominate university assessments. Key barriers to CAT adoption include limited lecturer knowledge and skills, technical constraints, inadequate infrastructure, and incompatibility with essay-based assessments. Despite these challenges, lecturers express interest in a more accessible and cost-effective CAT. The study concludes that lecturers acknowledge CAT's potential to enhance fairness and accuracy in testing while they face difficulties in its implementation. A targeted CAT development program aligned with lecturers' needs with institutional support is essential for successful integration. CAT has the potential to improve assessment effectiveness and efficiency in Indonesian universities aligning with technological advancements and 21st-century education trends.

Keywords: Challenge, Computerized adaptive testing, Difficulty, Evaluation, Mathematics, Prospective elementary school teachers, Technology.

Citation | Imawan, O. R., Retnawati, H., Haryanto, & Ismail, R. (2025). The challenges of implementing computerized adaptive testing in Indonesia. *Journal of Education and E-Learning Research, 12*(2), 124–144. 10.20448/jeelr.v12i2.6677
History:
Received: 31 December 2024
Revised: 21 March 2025
Accepted: 14 April 2025
Published: 12 May 2025

Publisher: Asian Online Journal Publishing Group

Funding: This research is supported by the Ministry of Education, Culture, Research, and Technology, Center for Higher Education Funding, Indonesia (Grant number: 0668/J5.2.3/BPI.06/10/2021) and the Indonesia Endowment Funds for Education, Indonesia (Grant number: 0668/J5.2.3/BPI.06/10/2021).
Institutional Review Board Statement: The Ethical Committee of the Yogyakarta State University, Indonesia has granted approval for this study on 20 October 2024 (Ref. No. T/60.1/UN34.9/PT.01.04/2024).
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

Contents

Contribution of this paper to the literature

This study explores the challenges of implementing computerized adaptive testing (CAT) in Indonesia focusing on issues such as infrastructure, digital literacy, and regional disparities. It underscores the importance of comprehensive strategies to address these barriers, providing valuable insights for policymakers, teachers, and lecturers to enhance assessment practices within Indonesia's evolving educational system.

1. Introduction

Education is a fundamental human need as people are naturally inclined to seek knowledge that enables them to address challenges and enhance their quality of life. Consequently, educational institutions are consistently sought after by prospective students, including future teachers (Castillo-Paredes, 2021; Ebenbeck & Gebhardt, 2022). A critical component of education for evaluating learning outcomes and for selecting future teachers is assessment (Dasopang, 2017). Assessment activities engage students and instructors in processes that begin with creating instruments for measurement. Each step of this process is interlinked, meaning that the quality of one stage impacts subsequent stages and, ultimately, the overall quality of the assessment.

Globally, assessment practices have evolved as experts develop new theories to enhance assessment quality. This evolution has seen assessment instruments progress from classical test theory (CTT) to item response theory (IRT) with the aim of creating more accurate and reliable tools (Fleming, Wilson, & Ahlgrim-Delzell, 2018). While many studies favor IRT for use in computerized adaptive testing (CAT), it requires stringent sample sizes. In some cases, analyses from both IRT and CTT yielded similar difficulty index results across different test packages (Kartowagiran, Mardapi, Purnama, & Kriswantoro, 2019). Analysis using CTT can still be done but it is less suitable for use in CAT. However, it works well for paper-and-pencil tests and CBT. The data were analyzed using classical test theory with difficulty levels, discriminating power, and distractor functionality. The test validity was assessed using Aiken's formula, and reliability was estimated using Cronbach's alpha. For example, the study found that, of the 20 initial multiple-choice items, 15 were valid and reliable and had good item characteristics with an average difficulty level rated as moderate at 0.28, a good discriminating power of 0.31, a good reliability coefficient of 0.79, and all distractors were functioning well (Hamdi, Suganda, & Hayati, 2018).

In developing assessment instruments, utilizing contemporary theories like IRT is crucial for maintaining relevance and effectiveness. IRT introduces dimensions of abilities or traits, tailoring test items to individual test-takers' characteristics. IRT operates on three core assumptions which are as follows: unidimensionality, local independence, and parameter invariance (Wilson & Narasuman, 2020). To support CAT, a robust item bank is essential. IRT can be used to evaluate item quality, for example, through Rasch analysis. When inadequate items are identified, they are professionally revised, recalibrated, and included in a high-quality item bank, strengthening the psychometric and theoretical foundation of CAT tools (Yim, Lye, & Koh, 2024). The question bank is tested before being used in CAT. A set of questions is usually grouped into several test packages during the trial. There is a study that describes five methods for developing parallel test items of the multiple-choice type in mathematics. The findings of the study show that teachers use five methods in developing test items, namely: (1) randomizing the question numbers; (2) randomizing the order of the answer options; (3) writing questions with the same context but different numbers; (4) using anchor items and (5) writing different questions based on the same specification table (Pramudita, Rosnawati, & Mam, 2019).

Several factor analyses support the development of quality test and non-test instruments. Two commonly used types of factor analysis are exploratory and confirmatory. A study collected data using an online survey technique. Confirmatory factor analysis was then used to analyze the data. The results of the study show the following six factors that contribute to academic stress among students: academic demands, parent-child relationships, traumatic childhood experiences, peer pressure, financial issues, and self-expectation. It is hoped that further research involving other factors of academic stress among students can be conducted to provide additional information on this topic (Siwi, Anindyarini, & Nahar, 2020).

After analysis, the instrument shows that the construct validity is proven by the standardized loading factor value which is considered significant and the instrument's reliability based on the high construct reliability coefficient, such as 0.890 indicates that the instrument is of good quality (Hudha & Mardapi, 2018). Instrument quality testing can also be done through questionnaire validity analysis using confirmatory factor analysis (CFA) (Imawan, Retnawati, Haryanto, & Ismail, 2024; Ismail, Imawan, & Nadhifah, 2023). The reliability estimation of the questionnaire is done using composite reliability. All instruments must have an appropriate model, construct validity, and reliability that meet academic requirements based on field studies (Gunartha, Sulaiman, Suardiman, & Kartowagiran, 2020). A good instrument is one with strong content validity and construct validity. There is a study in which content validity was achieved using the Lawshe's method and the CVR formula while construct validity was obtained through exploratory factor analysis using the SPSS program and later confirmed with confirmatory factor analysis using the Lisrel program. The construct validity was proven by the significant MSA value and the loading factor value. The fit of the construct model with the data was indicated by a p-value of 0.68 ($\geq$ 0.05), a root mean square error approximation of 0.00 ($\leq$ 0.08), and a goodness of fit index of 0.94 ($>$ 0.90). The reliability of the instrument was 0.774, which means it is reliable (Fardhila & Istiyono, 2019).

One way to select good test items for inclusion in a CAT question bank is through DIF analysis. Item differential function (DIF) describes a situation where test takers with similar abilities but from different demographic groups have different chances of achieving the same result. A DIF study found that through CFA, it can be proven that DIF-free items have better construct validity. The implications of this study are expected to inspire counseling psychologists to be more cautious when using rating scales or instruments (Ismail, Retnawati, Arovah, & Imawan, 2024; Sumin, Sukmawati, & Nurdin, 2022).

Several item analysis methods facilitate polytomous scoring because not all instruments have dichotomous results. A study focused on items with polytomous scoring aimed to investigate the advantages of assessment by utilizing the combination of the MCM/GPCM model compared to the 3PLM/GRM model within a mixed-item format in mathematics tests. The results of the study showed that the combination of the MCM/GPCM model provided more accurate estimations than the 3PLM/GRM model (Abadyo & Bastari, 2015).

One advantage of IRT in assessment analysis is its ability to determine levels or categories achieved by test-takers. This enables a detailed analysis of item characteristics and accurate ability estimates based on an ability scale (Otaya, Kartowagiran, Retnawati, & Mustakim, 2020). A challenge in CAT is the need for a well-constructed item bank as sample sizes can affect analysis quality. The 2PL model can achieve stability in providing item parameter estimates using a minimum of 1000 trial respondents (Ibrahim, Retnawati, Irambona, & Orantes Pérez, 2024). 2PL analysis, namely analysis that produces parameters a (differentiating power) and b (level of difficulty), can provide insight into the academic relevance of the content being tested in the use of CAT (Karimah, Retnawati, Hadiana, Pujiastuti, & Yusron, 2021). The research findings indicate the learning achievement tests for grades VII and VIII have a difficulty level (location) categorized as moderate (0.190 and 0.451), and a discrimination index (slope) in the good category with averages of 0.700 and 0.633 meaning that a good 2PL IRT analysis is required to produce a well-designed test instrument (Nisa & Retnawati, 2018).

Currently, 3PL models (with additional "guess" parameters) and 4PL (with additional "carelessness" parameters) are increasingly being used to obtain test item parameters, although the Rasch, 1PL and 2PL models are still more frequently used (Pardede et al., 2023). This 4PL model is used with consideration of sufficient knowledge, the need for additional information and if the test taker has high ability (Siwi et al., 2020). IRT offers a consistent analysis of test item characteristics because it assesses the inherent properties of the test items without dependence on the test taker. On the other hand, CTT analysis produces test item characteristics that are influenced by certain test takers, so they are not consistent when used on different test takers.

The 3PL analysis has become increasingly popular. For example, a study aimed at analyzing the characteristics of the EPOT instrument, which consists of listening, structure, and reading subtests, and subsequently identifies the quality of each EPOT test item. This test, called the English Proficiency Online Test (EPOT), follows the TOEFL ITP (Institutional Testing Program) framework. The study employed a descriptive quantitative approach by describing the characteristics of EPOT test items in terms of item difficulty index, item discrimination index, test information function and test measurement errors. All subtests fit the 3-PL model based on the analysis results. Most of the EPOT test items showed a good range of difficulty and discrimination indices. The EPOT information function indicates that accurate items are used in the 3-PL model for specific ability ranges (Siwi et al., 2020).

The advantage of applying the 2PL model is that it can classify question items correctly in various difficulty category values, determine the effectiveness of distractors (answer choices) and assess the suitability of the test (Kusumawati & Hadi, 2018). The Rasch and 1PL models are also widely applied in making CAT because they can identify content that is difficult for test takers, are simple to analyze and have been tested for stability (Ulwatunnisa, Retnawati, Muhardis, & Yusron, 2024). The use of IRT can help teachers or lecturers to find out where the test takers' difficulties are so that activities can then be carried out to provide support to improve the test takers' abilities. Besides, IRT can also identify which items are of good and bad quality (Kartianom & Mardapi, 2017). The difficulty level indicates that 100% of the essay questions in the real analysis final exam are categorized as difficult, suggesting that easier questions need to be created in order to achieve a more balanced difficulty level by using the Rasch model for partial credit scoring (Isnani, Utami, Susongko, & Lestiani, 2019). One of the common analyses used to calibrate a CAT question bank is the Rasch model. There are various methods to perform Rasch analysis. A study shows the ease of conducting Rasch analysis using the R program. The research data were analyzed using the Rasch model assisted by the R program. The Rasch model fit the data for 42 items after being calibrated three times. Based on difficulty level, ICC, and item reliability, 28 out of 42 items (66.67%) were considered good (Muchlisin, Mardapi, & Setiawati, 2019).

CAT is also thought to be suitable for large-scale assessments because it uses technology that makes assessment activities easier and faster in contrast to paper and pencil assessments which require more time, effort and costs. The use of information technology in assessment activities can increase the effectiveness and efficiency of the process (Hsu & Wang, 2019). The development of computational theory supports the development of assessment methods to adapt to current needs through algorithms based on Artificial Intelligence (AI) (Liu et al., 2024). The shortcomings of paper- and pencil-based assessments need to be resolved. There are various solutions, one of which is the application of CAT which is relevant to today's skills. There are various skills needed in this modern era full of new problems, one of which is problem-solving skills. Assessment using CAT can also assess problem-solving skills with a high level of accuracy (Istiyono, Dwandaru, & Faizah, 2018).

CAT uses question items analyzed by IRT can carry out a consistent and adaptable assessment process. This adjustment allows the test to meet the needs of various types of test takers' abilities, namely by adjusting the questions based on the responses given, thereby increasing accuracy and fairness in the assessment. CAT makes the assessment process more practical and relevant to the abilities of each individual test taker, especially for students in certain education, who benefit from efficient assessment activities, for example, if CAT focuses on mathematics skills (Ebenbeck & Gebhardt, 2022, 2024). However, using CAT alone is not enough to improve students' mathematical abilities because it needs to be combined with appropriate and interest-based learning activities to truly support students' development. Effective preparatory activities, such as cooperative and interest-based learning models, can strengthen the impact of CAT and ensure that the assessment is meaningful and supports students' educational goals (Huang et al., 2020; Menendez et al., 2013).

Students with special needs can be evaluated for mathematical proficiency using verbal, non-verbal, and mathematical abilities with the CAT exam which has been specially modified in this way (Şenel & Kutlu, 2018). Based on previous research, very few articles discuss the use of CAT to assess the mathematical abilities of students with special needs, and there needs to be a need to research CAT-based test methods to evaluate the mathematical abilities of prospective students precisely. Therefore, the research theme raises the challenges of using CAT to assess the mathematical skills of prospective students (Yuan, Xia, Han, & Hu, 2020). It is necessary to conduct more serious research to photograph the assessment activities, primarily to assess the mathematical abilities of prospective students with special needs, to know the actual situation, and then find the best solution to overcome it because of the importance of this problem. Portraits of the implementation of mathematical assessment activities need to be known as soon as possible so that a solution is found if it is proven that there is a problem. This solution is expected to make educational activities, especially mathematics assessments for students with special needs, more effective and efficient. Appropriate assessment methods/techniques will reduce costs, reduce energy, reduce time use, and improve the quality of education in an educational unit.

Advancements in information technology have made it feasible to conduct assessments with computer assistance, providing numerous benefits over traditional paper-and-pencil methods (Jette et al., 2007). Computer-assisted tests offer several advantages, such as (1) reducing the time required for grading and reporting and (2) eliminating logistical challenges like distributing, storing, and reproducing test materials. Conventional paper-based tests are often subject to issues like test leaks and cheating. Additionally, administering the same test items to participants with varied skill levels can lead to inefficiencies, as a one-size-fits-all approach may not accurately measure abilities and can introduce significant measurement errors (Kristóf & Eszter, 2016). Computer-based assessments now offer a more efficient and effective way of measuring examinee abilities compared to conventional methods (Chalmers, 2016). A concrete example of this can be seen in the implementation of the computer- based national examination which has been taking place in the last few years for elementary school, junior high school and high school students in Indonesia, which shows how technology can be applied in the examination system (Nufus & Wahyuni, 2024). The use of paper-and-pencil tests is decreasing due to the long administration process and delays in providing feedback with increasingly rapid advances in technology (Cook et al., 2008).

CAT is the latest development in computer-based assessment, which offers a more flexible approach and is tailored to the test taker's abilities. It adjusts the questions given based on each individual's ability level, creating a more personal assessment experience rather than a general assessment (Haley et al., 2011). In the CAT system, the types of questions given to examinees depend on their answers to previous questions. A correct answer will lead them to a more difficult question while a wrong answer will lead to an easier question. This approach allows test takers to answer a unique set of questions, according to their ability or skill level, thereby increasing the accuracy and efficiency of test results. For prospective elementary school teachers, CAT is very useful for assessing their mathematical abilities, because (1) it provides high accuracy with fewer questions, (2) guarantees efficient and quality testing, (3) reduces overall exam time, and (4) reduces the possibility of cooperation or cheating between test takers because they will not face the same questions. With its adaptive nature, CAT also provides more direct insight into a participant's abilities, supporting rapid, data-driven decision- making.

The benefits of using computer-based tests, especially with the application of CAT have been widely discussed in the literature. As research results from Stepanek and Martinkova (2020) CAT offers several key advantages, including (1) flexible test management, (2) enhanced test security, (3) increased motivation among test-takers, and (4) the potential to reduce testing time by up to 50% while maintaining reliability (Stepanek & Martinkova, 2020). In addition, CAT can provide the most accurate measurement of examinee abilities. CAT can select questions that suit each participant's ability level, providing relevant questions based on their previous answers with the help of intelligent algorithms (Reckase, Ju, & Kim, 2018).

A researcher from the University of Macedonia in Greece developed a mobile-based CAT system, CAT-MD to assess students' abilities in physics in secondary schools and illustrate the potential of CAT. The system allows students to access the exam through a mobile phone, PDA, or laptop, demonstrating the flexibility of CAT in a variety of technological formats. They used the 1PL model with a question difficulty index ranging from -2 to +2, which shows the CAT's ability to adapt to various student ability levels (Triantafillou, Georgiadou, & Economides, 2008). Another application of CAT was developed at Universiti Sains Malaysia where CAT was used to evaluate science abilities among grade 8 students as part of the International Trend of Mathematics and Science Study (TIMSS). This research emphasizes the efficiency of CAT in achieving a high level of precision with a smaller number of questions which effectively differentiates between students with varying levels of academic ability (Sumin et al., 2022).

Additionally, a researcher developed CAT to assess the mathematics abilities of grade 7 junior high school students using the triangular tree method where the decision to choose questions takes place dynamically during the assessment process. With 193 math questions in its database, this CAT software utilizes the Maximum Likelihood Estimation (MLE) method to accurately measure a student's ability. This adaptive capability ensures that each examinee will be given questions appropriate to their level of ability, resulting in a more precise measurement of skills. These studies on CAT confirm the potential of CAT in a test taker in a more accurate manner, as it can be adapted to his needs and evolve according to his progress efficiently (Winarno, 2012). One method that is often used to estimate capabilities in the CAT algorithm is MLE. However, MLE has a weakness, namely the difficulty in estimating the ability of examinees whose scores do not follow a certain pattern. Therefore, alternative strategies are needed besides MLE to increase the accuracy of question selection and provide a greater variety of questions which in turn can increase test security in CAT (Suhardi, 2020). Various methods are also available for selecting initial questions in CAT and research shows that the low-high method using IRT is in accordance with the principles of adaptive testing and can determine test takers' abilities with only 17 questions (Rukli & Atan, 2024).

Another example of CAT development was carried out by a researcher, who developed a CAT-based multiple-choice test which was divided into three batteries and ten subtests. These subtests are organized into a six-level hierarchy designed to measure the abilities of students ranging from seven and a half years old (level A) to nearly sixteen years old (level F). Quantitative batteries cover a variety of topics such as quantitative relationships, number series and creating equations. For example, in the quantitative relationships section, students are asked to determine whether a pair of numbers is greater, equal to or less than another pair. Each level of the test (A to F) is age-adjusted with all six levels combined in one booklet and responses recorded using an optical grade reader sheet for computerized scoring. This system allows assessments to be conducted centrally or decentralized with results typically reported in the form of a standard age score (SAS) similar to intelligence testing, where the mean score is 100 and the standard deviation is 15. Scores are given for each battery as well as for the average. Overall average provides a comprehensive picture of the abilities of each examinee (Marfeo et al., 2019).

The advantages of CAT, such as flexibility, accuracy, and efficiency have driven much research and development in the field of assessment. The use of CAT utilizes technology to make the testing process more effective and efficient. It is important to understand the extent to which mathematics educators in Indonesia, especially mathematics lecturers are prepared and have an understanding in applying CAT to assess the abilities or skills of their students, namely prospective elementary school teachers in Indonesia considering the importance of

CAT in the current digital era. By exploring this perspective, this research focuses on identifying the challenges and opportunities associated with CAT, which in turn can support wider implementation in the world of education.

This research is significant because it provides valuable insights into the challenges of implementing CAT in Indonesia, particularly in the context of mathematics education in higher education. As technology continues to advance rapidly, the potential of CAT to transform traditional assessment practices becomes increasingly apparent. CAT offers a more personalized and efficient approach to testing, adapting to each student's ability level in real time and providing a fairer and more accurate measurement of their knowledge. In Indonesia where the educational landscape is rapidly evolving, the introduction of CAT could significantly enhance the effectiveness and efficiency of assessment processes, reducing biases and improving the overall quality of educational evaluations.

Understanding the challenges in implementing CAT in Indonesia offers significant opportunities to design corrective actions and follow-up steps to overcome the existing barriers. One of the first steps that can be taken is the development of intensive training programs and workshops for lecturers and teachers. Given that many lecturers are unfamiliar with CAT or have limited skills in using the technology, these programs would enhance their knowledge and competencies in designing and implementing CAT effectively. Another crucial follow-up action involves improving technological infrastructure. Since inadequate infrastructure is a major hindrance to the successful implementation of CAT, universities in Indonesia would benefit from investing in the necessary hardware and software that support CAT systems. This would make the adoption of CAT more feasible, ensuring that it can run smoothly and efficiently across campuses.

Moreover, addressing the incompatibility of CAT with essay-based questions is another important aspect. To overcome this challenge, it would be essential to adapt the current assessment models by developing question formats that align better with the CAT system, such as multiple-choice or computer-based questions. This adaptation would ensure that the testing system functions properly while maintaining the integrity of the assessment process. Furthermore, understanding these challenges also provides an opportunity to influence education policies. It is important to encourage stakeholders at both the university and government levels to create policies that support CAT implementation. These could include providing funding for infrastructure upgrades or establishing policies that promote the use of technology in educational assessments.

In addition, there is a clear demand from lecturers for more affordable and accessible CAT systems. Therefore, developing platforms that are cost-effective and easy to access would be a valuable next step. This could involve creating open-source CAT software or collaborating with technology providers to offer more affordable solutions for Indonesian universities. Ultimately, addressing these challenges will significantly improve the quality of educational assessments in Indonesia. By implementing CAT, universities would be able to create fairer, more objective assessments that offer a more accurate measurement of student abilities. This shift toward a more modern, technology-driven assessment system could play a key role in preparing graduates for the demands of a rapidly evolving workforce. Overall, tackling these issues would make education in Indonesia more efficient, effective, and aligned with global advancements in assessment practices. The research questions are as follows: (1) What assessment practices are implemented by mathematics lecturers in Indonesia? (2) What challenges do mathematics lecturers face in implementing CAT in Indonesia? These two questions focus on describing the assessment practices and the barriers encountered in the implementation of CAT.

2. Literature Review

The adoption of computerized adaptive testing (CAT) in educational assessments has garnered considerable attention due to its ability to personalize assessments and enhance their accuracy. CAT is an advanced testing method where the difficulty of the test is adjusted in real-time based on the test-taker's performance. This technology has been shown to improve the precision of student evaluations, especially in subjects such as mathematics, where varying levels of competency are evident. While CAT is increasingly used in many educational contexts worldwide, its implementation in Indonesia, particularly in mathematics assessments, remains a challenge.

Computerized adaptive testing (CAT), which is gaining popularity is built on two main components: computer technology and item response theory (IRT). The core idea of CAT is that an algorithm selects test items that align with the ability level of each individual test-taker (Wainer, 2000). Advances in information technology that are developing at this time make it possible to administer a test using computer assistance (Hambleton & Swaminathan, 2013). Computer-assisted tests have several advantages over paper-and-pencil tests such as (1) reducing the time for test assessment work and writing reports, (2) eliminating logistical work such as distributing, storing and reproducing questions. During the written test, there are often leaks or cheating. In addition, the drawback of conducting conventional tests is that participants with varying abilities are given the same items so that the test device becomes inefficient because the measurement scale is no longer appropriate and allows measurement errors to be quite large (Retnawati, 2016).

Computerization makes individual assessments more efficient and accurate than paper and pencil tests (Boo & Vispoel, 2012). The computer-based national examination (CBNE), which was implemented in schools several years ago, is an example of the use of computers in measuring students' abilities. In the era of advanced and informative technology, it is very feasible to conduct computer-based tests (Ismail, Retnawati, & Imawan, 2022; Khoshsima & Toroujeni, 2017). The impact of technological advances, paper- and pencil-based testing has decreased due to the length of time in administering tests and feedback (Boo & Vispoel, 2012; Imawan, Retnawati, Haryanto, & Ismail, 2025; Ismail, Retnawati, Sugiman, & Imawan, 2025).

The use of computers has occupied a wide scope; recently the adaptive computer-based test media, namely CAT has begun to be widely studied. CAT is an adaptive-based media that provides test items to test takers according to their abilities (Thompson & Weiss, 2011). This allows each test taker to take the test with an unequal number of items due to the different abilities of the participants. The choice of questions is based on the answers of the examinees to the questions that are currently given, if the answers are correct, then questions with a higher level of difficulty will be given but if the answers are incorrect then the next question will have a lower level of difficulty. Items that are either too easy or too difficult for test-takers provide minimal information about their ability so test-takers typically receive questions with a success probability of around 50% (Green, Bock, Humphreys, Linn, & Reckase, 1984).

Meanwhile, CAT-based testing can improve efficiency and accuracy as well as practicality in its implementation. Based on this, CAT can be used as a medium for selecting student abilities because of (1) high accuracy from a smaller number of questions, (2) efficiency and quality of tests, (3) shorter examination times, and (4) cooperation between students. Test takers can be avoided because each test taker does not necessarily get the same questions and test takers can also find out their abilities early.

The advantages of administering tests using computers, especially CAT among others are (1) flexibility of test management, (2) increased test safety, (3) increased motivation, (4) reduce testing time by up to 50% while maintaining the same level of reliability (Georgiadou, Triantafillou, & Economides, 2006). The standard error of measurement (SEM) decreases with each item as more information about the examinee's ability is gathered. This process allows for a reduction in the number of items given without compromising accuracy (Lunz, Bergstrom, & Gershon, 1994; Wainer, 2000). CAT offers numerous advantages over fixed-item tests (FIT). Flens, Smits, Carlier, van Hemert, and de Beurs (2016) found that the number of items in CAT procedures is reduced by 26 to 44 percent compared to FIT while testing efficiency is actually improved (Haley et al., 2011). CAT can produce the most significant information in measuring the ability of test takers. In addition, the CAT model with intelligent algorithms is able to select test items with the appropriate level of difficulty according to the students' answers (Mardapi, Haryanto, & Hadi, 2012).

As an example of the application of CAT, there is a study conducted by Triantafillou et al. (2008) from the University of Macedonia, Greece, namely developing computerized adaptive testing on mobile devices (CAT-MD). CAT application can be accessed using wireless devices such as mobile phones, personal digital assistants (PDAs), or laptops. This CAT-MD is intended for 2nd grade high school students in Greece for physics. The question bank development in this study used the 1 PL model with an item difficulty index between -2 and +2. When conducting a trial for a question bank for CAT, it is usually done in several test packages. In relation to the creation of test packages, a study aimed to prove that parallel tests can be constructed by randomizing the question numbers and the order of answer alternatives. The results of the study show that the item analysis using both classical test theory and item response theory approaches revealed no significant difference in the difficulty index between package 1 and package 5. However, according to classical test theory, there was a shift in the difficulty index category for packages 2 to 5 when compared to package 1, the original package which was not a good package because it contained questions that were too easy (Kartowagiran et al., 2019).

In addition, there is also research on other CAT, namely the evaluation of the ability of 8th grade science students in the International Trend of Mathematics and Science Study (TIMSS) using computerized adaptive testing conducted by Samsudin, Chut, and Ismail (2019) from University Sains Malaysia. CAT is used as an alternative paper and pencil test instrument to determine whether the application of CAT can produce a high level of precision with fewer items given and distinguish different academic levels among groups of students, where CAT is configured in Concerto.

There is also a study by Winarno (2012) who developed a CAT using the decision triangle tree method in the procedure for selecting test items during the implementation of the test. The questions developed are for Junior High School/Islamic Junior High School, grade VII mathematics which are entered in the CAT software database, consisting of 193 questions. The estimation of the test taker's ability is done by using the maximum likelihood (MLE) method. The developed CAT is able to provide adaptive questions based on the responses of test takers and can measure the ability of test takers accurately and precisely.

Some of the advantages of carrying out tests using the CAT have made this application a lot of research and development. Utilization of technology for test purposes is aimed at the effectiveness and efficiency of the implementation and administration of tests. Because of the importance of CAT in the digital era as it is today, it is necessary to examine how the implementation of tests by mathematics lecturers in Indonesia has been so far, and it is also necessary to examine the understanding of mathematics lecturers in Indonesia regarding CAT. Numerous studies have been conducted on the psychometric and technical aspects of CAT (for a review, see (van der Linden & Pashley, 2000)). Topics include the development of the item pool (Imawan et al., 2025; Ismail, Retnawati, Sugiman, & Imawan, 2024; Lee & Dodd, 2012; Wise & Kingsbury, 2000), comparisons of item selection methods (Finkelman, Kim, Weissman, & Cook, 2014; Van Der Linden, 2005) and stopping rules (Choi, Grady, & Dodd, 2011).

In Indonesia, mathematics assessment in higher education traditionally relies on conventional methods such as paper-and-pencil tests. Computer-based testing (CBT) has seen limited implementation, though it became more common during the COVID-19 pandemic as institutions shifted to remote learning. The transition to more adaptive systems like CAT remains slow despite the growing prevalence of CBT. Traditional assessments often fail to address the diverse learning levels of students and do not fully leverage the advantages of technology-driven assessment tools.

Online testing platforms do not yet offer the adaptability that CAT provides, which is critical for accurately gauging the mathematical abilities of students while taking a step toward digitalization. CAT, by dynamically adjusting question difficulty based on individual student performance, promises a more efficient and personalized assessment method. However, the implementation of such systems is constrained by both technical and pedagogical challenges in Indonesia.

The implementation of CAT in Indonesia is hindered by several key challenges. Researchers have emphasized the need for collaborative efforts between educational institutions, technology developers, and policymakers to create solutions that address both the technical and pedagogical needs of the Indonesian education system to overcome these barriers. Such collaboration could help bridge the gap between the potential of CAT and its practical application in the Indonesian context. CAT also comes with some disadvantages, such as higher development costs, the necessity for regular maintenance of the item bank, and complex technical requirements (Tan, Cai, Li, Zhang, & Tu, 2018).

One important issue that is often overlooked is the psychological impact of CAT on test-takers. It has been argued that because the items in CAT are tailored to the test-taker's ability, it can be more motivating and cause less anxiety compared to traditional fixed-item tests (Linacre, 2000; Mead & Drasgow, 1993; Wainer, 2000; Weiss, 1982). While the accuracy and efficiency of CAT compared to FIT are highly significant from the perspective of test developers, these advantages are not always recognized by test-takers (Kimura, 2017). Meta-analytic results

indicated that there was no overall effect of test type on anxiety and motivation when comparing CAT to FIT. However, specific modifications in CAT administration can have positive psychological effects on test-takers (Akhtar, Silfiasari, Vekety, & Kovacs, 2023).

3. Method

3.1. Research Design

This research is a type of qualitative research with a phenomenological approach. This research aims to explore information on two main aspects, namely: (1) mathematics assessment practices currently being implemented at universities in Indonesia and (2) the challenges faced by mathematics lecturers in implementing CAT to assess the mathematics abilities of prospective elementary school teachers. Through a phenomenological approach, this research seeks to understand the experiences and perceptions of mathematics lecturers regarding the application of CAT in the assessment process.

This research is a qualitative study that adopts a phenomenological approach to explore two key aspects in depth. The first aspect focuses on the current mathematics assessment practices being implemented at universities in Indonesia. Specifically, it aims to examine the methods and tools used by mathematics lecturers to evaluate the mathematical abilities of prospective elementary school teachers. This includes understanding the types of assessments employed, such as traditional paper-and-pencil tests, computer-based assessments, or alternative forms of evaluation, and how these practices align with the educational goals and standards in Indonesian higher education. The second aspect of the study investigates the challenges faced by mathematics lecturers in implementing computerized adaptive testing (CAT) as a tool to assess the mathematics abilities of prospective elementary school teachers. This includes identifying the technical, pedagogical, and infrastructural barriers that hinder the adoption of CAT. The research will explore lecturers' experiences with the technology, their level of familiarity with CAT, and the perceived difficulties in adapting this advanced testing method to the context of Indonesian universities.

This research aims to capture the lived experiences and perceptions of mathematics lecturers, focusing on how they perceive the potential benefits and limitations of CAT in the assessment process by utilizing a phenomenological approach. The goal is to gain a deeper understanding of the factors that influence the acceptance and implementation of CAT as well as the underlying reasons for its challenges. Ultimately, this research seeks to contribute to the development of more effective assessment strategies and offer insights that could guide future efforts to integrate CAT into the higher education system in Indonesia.

3.2. Data Collection Techniques

Data for this research was collected through a two-stage approach to ensure a comprehensive understanding of the mathematics assessment practices and the challenges in implementing CAT in Indonesia. The first stage involved the distribution of questionnaires to 17 mathematics lecturers representing diverse regions, including Western, Central, and Eastern Indonesia. These lecturers were selected to provide a broad perspective on the current mathematics assessment practices employed across the country. The questionnaire aimed to gather initial information on the methods, frequency, and tools used in mathematics assessment, as well as their familiarity with and use of CAT in their teaching. The responses helped establish a foundational understanding of the state of mathematics assessments and served as a starting point for further investigation.

In the second stage, four lecturers were selected for in-depth interviews based on their responses indicating a limited understanding or experience with CAT. These lecturers were chosen to gain a more nuanced perspective on the specific challenges they face when attempting to implement CAT in their assessment practices. The interviews were designed to probe deeper into their individual experiences, exploring not only the difficulties they faced but also their perceptions of the potential benefits and drawbacks of CAT in the context of mathematics education.

To guide these interviews, a structured interview guide was developed, consisting of open-ended questions designed to elicit detailed responses. The interview guide focused on two key areas: the lecturers' current assessment practices and their views on the feasibility and challenges of integrating CAT into these practices. The aim was to uncover insights regarding the technological, pedagogical, and logistical barriers they face, as well as any support or training they may have received to help them adapt to CAT. This structured approach allowed for the collection of in-depth, qualitative data that provides a richer understanding of the lecturers' experiences, perceptions and the challenges in applying CAT for mathematics assessment.

3.3. Data Analysis Techniques

The data analysis process in this study follows the framework outlined by Creswell (2014) which consists of four distinct stages: (1) data reduction, (2) data presentation, (3) data verification, and (4) drawing conclusions. This systematic approach ensures that the analysis is rigorous and thorough, facilitating the identification of significant themes and patterns that are relevant to the research objectives.

In the "data reduction" stage, the first step involves organizing and simplifying the vast amount of data collected from both the questionnaires and the in-depth interviews. This process includes reviewing the raw data, transcribing interviews, and categorizing responses to identify key areas of focus. During this stage, irrelevant or extraneous information is filtered out, and the data is condensed into manageable segments. This allows for a clearer focus on the core issues related to the research questions, particularly the assessment practices and challenges in implementing CAT. Coding techniques may be applied to the responses, grouping similar ideas and concepts together to create a structured framework for further analysis.

Next, in the "data presentation" stage, the reduced data is organized into a more coherent and accessible format. The goal is to present the data in a way that highlights key findings and makes the information easier to interpret. This may involve the use of tables, charts, or narrative summaries to visually represent patterns and trends that emerge from the data. In this stage, the researcher synthesizes the information, ensuring that the various aspects of the data are presented in a clear and organized manner. This helps to facilitate a deeper

understanding of the respondents' experiences and perceptions regarding mathematics assessment practices and the implementation of CAT.

The "data verification" stage involves checking the reliability and validity of the findings. This step ensures that the interpretations of the data are accurate and trustworthy. Techniques such as member checking, where interview participants review the findings for accuracy, or triangulation where multiple data sources are cross-checked, may be used to confirm the consistency of the results. The researcher also revisits the raw data and compares it with the presented themes and patterns to ensure that the analysis aligns with the original data. This verification process helps to minimize biases and ensures that the conclusions drawn are grounded in the data itself. A quality instrument has undergone testing processes, including Aiken's content validity analysis, concurrent validity, and Cronbach's alpha reliability (Wardani, Prihatni, Negeri, & Jl Jogja-Solo Km, 2018).

Finally, in the "concluding" stage, the researcher synthesizes all the analyzed data to address the research questions. This stage involves interpreting the data to identify overarching themes and insights, conclusion about the current state of mathematics assessment practices in Indonesia, and the specific challenges faced by lecturers in implementing CAT. The researcher reflects on how the findings relate to existing literature and theories, drawing connections between the data and the broader educational context. This stage leads to the formulation of recommendations for improving mathematics assessment practices and the integration of CAT into higher education in Indonesia.

This analysis ensures a rigorous, systematic approach to understanding the complex data, providing rich insights into the experiences of mathematics lecturers and the challenges they face with CAT implementation by following Creswell's framework.

3.4. Participants

The research participants consisted of 17 mathematics lecturers who had filled out questionnaires. From this group, four lecturers were selected for interviews based on their responses regarding their understanding of CAT. All respondents interviewed met the criteria of having at least five years of mathematics teaching experience to provide in-depth information. The distribution of participants based on their understanding of CAT can be seen in Table 1.

Table 1. The distribution of participants.

Category	**Understanding of CAT**			**Little understanding of CAT**			**Don't understand CAT.**			**Total participants**
Number and code name of interviewed	-			1(A)		1(B)	2(C & D)			4
University and region of interviewed.	-			Yogyakarta State University, W		19 November University, Kolaka C	Cenderawasih University, E			
Number of questionnaires	6	1	-	3	2	1	-	-	4	17
Region of questionnaires	W	C	E	W	C	E	W	C	E	

Note: W: Western Indonesia region.
C: Central Indonesia region.
E: Eastern Indonesia region.

3.5. Ethical Clearance

The study received ethical approval from the Institutional Review Board of Yogyakarta State University, Indonesia (approval number: T/60.1/UN34.9/PT.01.04/2024). Throughout the research process, ethical standards were strictly adhered to, including obtaining informed consent from all participants ensuring data confidentiality, and maintaining voluntary participation.

4. Result

4.1. Description of Questionnaire Results

Although the questionnaire was not the primary data collection technique in this study, it serves as an important complementary tool that provides valuable insights and supports the qualitative findings obtained through in-depth interviews. The following sections describe the results of the questionnaire completed by 17 mathematics lecturers.

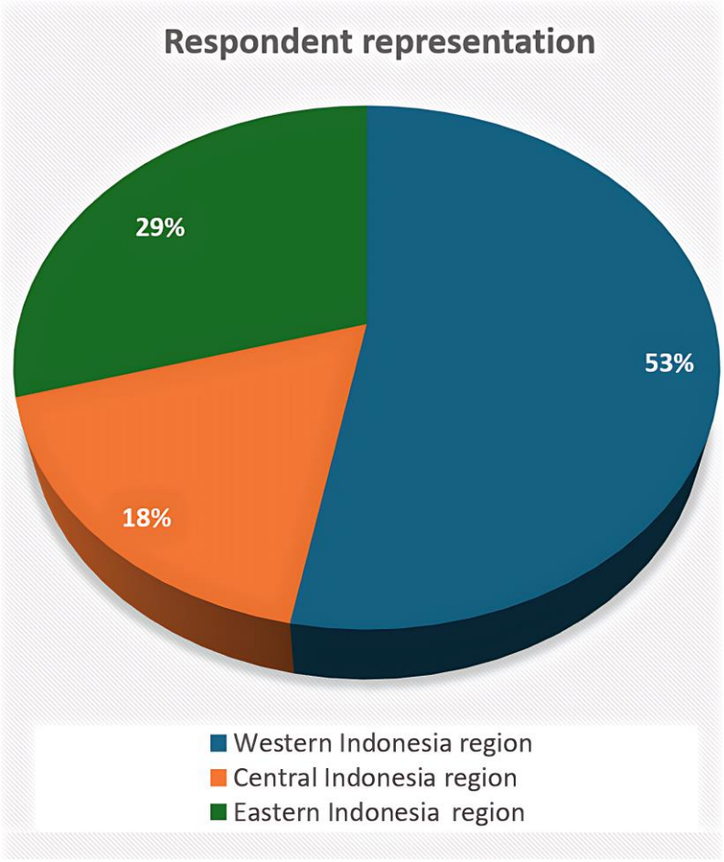

Figure 1. Percentage of respondent representation.

Figure 1 illustrates the percentage of respondent representation across different regions in Indonesia. The demographic analysis shows a diverse distribution with 53% of respondents being mathematics lecturers from Western Indonesia, 18% from Central Indonesia, and 29% from Eastern Indonesia. This distribution underscores the broad geographical reach of the study ensuring that perspectives from different educational contexts within Indonesia are represented. This diversity is critical to understanding regional differences in the application of mathematics assessment techniques and the challenges that arise as a result of these differences.

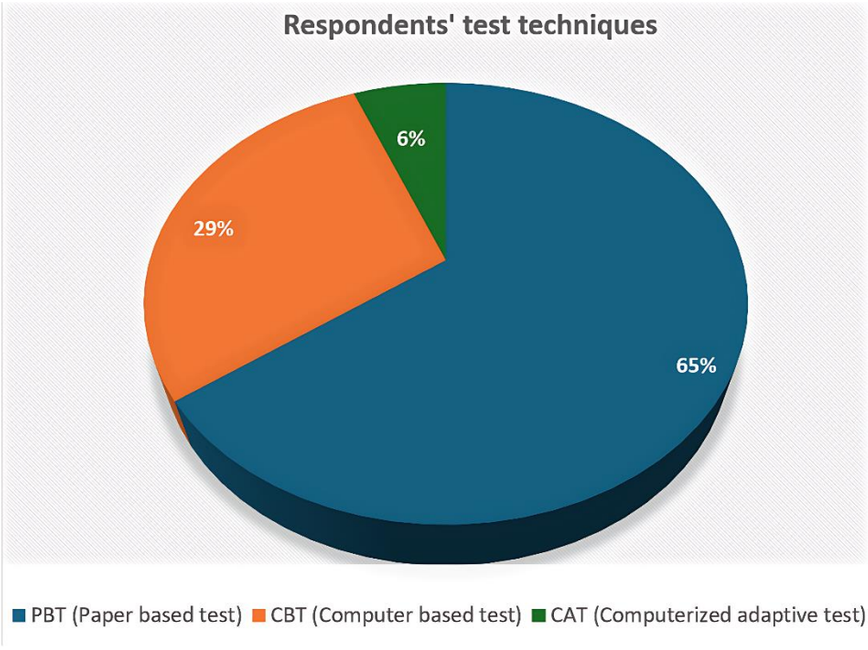

Figure 2. Percentage of respondents' test techniques.

Figure 2 illustrates the percentage of respondents' test techniques. The questionnaire results reveal that 65% of mathematics lecturers use paper-based tests, 29% employ computer-based testing (CBT) such as Google Forms, and only 6% implement computer adaptive testing (CAT). This distribution highlights the dominance of traditional assessment methods. The low adoption of CAT suggests significant potential for development, indicating a gap in the use of modern testing technologies that could enhance assessment processes and improve educational outcomes.

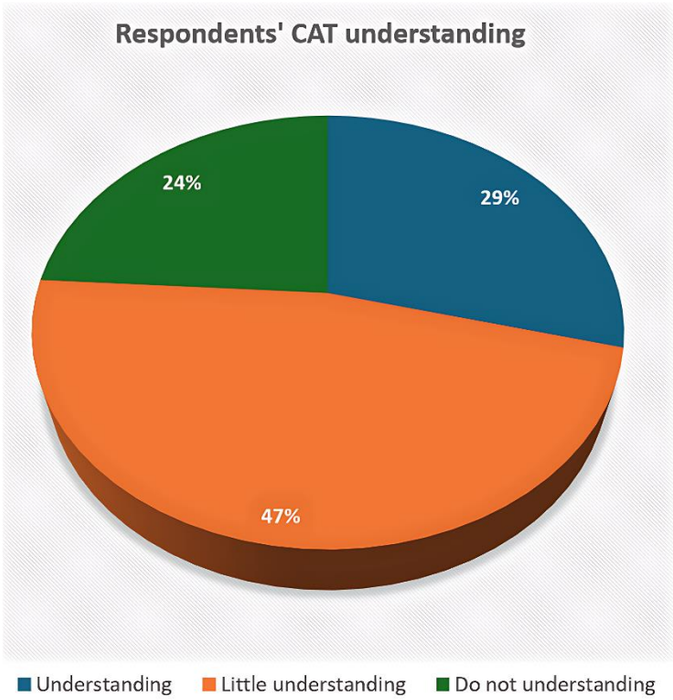

Figure 3. Percentage of respondents' CAT understanding.

Figure 3 illustrates the percentage of respondents' understanding of CAT. The questionnaire results indicate that 29% of lecturers reported understanding CAT, while 47% had limited understanding, and 24% did not understand it at all. This indicates that most lecturers have only limited knowledge of CAT, which could be a barrier to its effective application in mathematics assessment. To gain a deeper understanding of these views and identify specific barriers to understanding and implementing CAT, in-depth interviews were conducted with representatives of groups who had limited understanding and those who did not understand CAT. This qualitative approach aims to explore challenges and knowledge gaps about CAT as well as provide more complete insight into the potential role of CAT in mathematics assessment for prospective elementary school teachers. Overall, the questionnaire results provide important context for the qualitative findings of this study, highlighting current assessment practices and demonstrating the need for further education and training regarding innovative assessment methods such as CAT.

4.2. Challenges of Using CAT- Based on Questionnaire Results

The questionnaire in this study also produced findings in the form of respondents' answers which could be grouped into several themes in addition to producing a description of the characteristics inherent in the questionnaire respondents described in the previous section. The answers of respondents who are grouped are answers related to the challenges of mathematics lecturers in understanding, using, or making CAT applications. The following is the grouping of themes based on the results of the questionnaire.

4.2.1. Theme 1: Lack of Understanding of Mathematics Lecturers Regarding CAT

The responses summarized in Table 2 indicate a significant lack of understanding among mathematics lecturers regarding CAT. Many respondents expressed uncertainty about the procedures and practical applications of CAT, preferring to rely on conventional testing methods. This lack of familiarity presents a barrier to the effective implementation of CAT in their assessment practices. Several respondents noted that the perceived complexity of CAT contributed to their reluctance to engage with this testing method. As a result, there is a critical need for targeted training and educational resources to enhance lecturers' understanding of CAT.

Table 2. Lack of knowledge of mathematics lecturers regarding CAT.

Respondent's answer	Theme
1. The procedure or sequence is often forgotten.	Lack of understanding of mathematics lecturers regarding CAT.
2. I need help understanding how to use CAT. It sounds difficult to make/ use it.	
3. Because I don't understand, I usually prefer to use conventional techniques, or during a pandemic, I use Google forms (CBT).	
4. I've never heard of it or used it, so I don't know its difficulties.	
5. I don't need to learn how to use CAT.	
6. If the device is ready, it will be easy to use. However, I need to understand how to complete a CAT.	
7. I don't know what application to input the questions because I have never used or created this CAT-based question.	
8. Someone who uses this application in the assessment must really understand how to use it because otherwise, it will be confusing.	

4.2.2. Theme 2: CAT Requires Adequate Hardware and Software

Table 3 illustrates that respondents highlighted the necessity for adequate hardware and software to facilitate CAT. Several lecturers noted that the existing computer facilities were insufficient which poses a significant challenge to implementing CAT effectively. The lack of access to appropriate technology can hinder both the

administration of CAT and the overall testing experience for students. Moreover, the lengthy preparation process for CAT questions and the need for extensive question banks were also emphasized. Addressing these hardware and software needs is essential for the successful deployment of CAT in educational settings.

Table 3. CAT requires sufficient hardware and software.

Respondent's answer	Theme
1. Students need a laptop or computer for the CAT exam, and the computer lab on campus needs to be improved. 2. Hardware and software are needed to run CAT, and a question bank is available. 3. The main difficulty is that the number of computers still needs to be improved. Moreover, there is no computer lab in the study program or the department. 4. School regulations students are not allowed to bring cellphones and have computer limitations. 5. The process of becoming a ready-to-use question takes a long time. 6. We should also have many question banks or a collection of questions to make it easier. 7. Typing mathematical equations/ characters is challenging if you access them using a smartphone.	CAT requires adequate software and hardware.

4.2.3. Theme 3: CAT Requires an Adequate Internet Network

According to Table 4, respondents expressed concerns about the reliance on a stable internet connection for CAT administration. Interruptions in the network can severely impact students' ability to complete assessments leading to inconsistencies in test results and increased anxiety for both students and lecturers. This highlights the importance of ensuring reliable internet access in educational institutions to facilitate the effective use of CAT and to support a smooth testing experience.

Table 4. CAT requires an adequate internet network.

Respondent's answer	Theme
1. There is no difficulty, which becomes a problem if the network is disrupted. 2. Internet network problems in working on issues.	CAT requires an adequate internet network.

4.2.4. Theme 4: CAT Cannot Be Used for Essay Tests

The responses in Table 5 reveal a consensus among respondents that CAT is not well-suited for evaluating essay or descriptive questions. This limitation stems from the inherent nature of CAT, which is primarily designed for objective testing formats. As a result, mathematics lecturers expressed concerns about the feasibility of using CAT for comprehensive assessments that require in-depth written responses. This challenge points to the need for developing alternative assessment methods that can effectively evaluate students' abilities in solving complex mathematical problems.

Table 5. CAT cannot be used for essay tests.

Respondent's answer	Theme
1. It is not suitable to be applied to the description problem. 2. Difficulty in entering questions in image form. 3. Measuring students' ability in solving description questions.	CAT cannot be used for essay tests.

4.3. Challenges of Using CAT- Based on Interview Results

The following are the results of in-depth interviews with 4 mathematics lecturers who are indicated to be in the slightly understand and don't understand CAT categories based on the results of filling out the questionnaire. The answers of the four mathematics lecturers were then summed up to facilitate understanding. Furthermore, each interviewer's questions and answers will be presented to each participant.

The following section summarizes the results of in-depth interviews with four mathematics lecturers who had varying levels of understanding of the CAT based on their answers to the questionnaire. The insights they shared provide a more complete picture of the challenges faced in implementing CAT in educational settings.

4.3.1. Overview of Participant Responses

All four participants (P1, P2, P3, and P4) provided valuable feedback regarding their experiences and perceptions of CAT. Their responses revealed a general lack of understanding of CAT reflecting significant gaps in understanding its basic principles and application in education.

4.3.2. Lack of Understanding of CAT

Interview question 1 is symbolized by Q1, namely, "is it true that you have not heard or known or understand about computerized adaptive testing?" Suppose the four participants are symbolized by P1, P2, P3, and P4. Each participant's answer is P1: "Yes, I don't understand about computerized adaptive testing, which I know is just CBT, computerized- based test." P2: "I've heard the name CAT, but don't know how to make and use it." P3: "I have ever heard of CAT, only understand a little. It is usually used in teacher professional education program (TPEP), if I'm not mistaken." and P4: "Have heard a little." From the answers of the four participants, it can be concluded that the lecturer in mathematics has not understood CAT. These responses indicate a clear need for more comprehensive training on CAT among mathematics lecturers, highlighting the necessity for educational institutions to provide resources and workshops to familiarize staff with this testing method.

4.3.3. Experience with Conventional and Computer-Based Testing

Q2, namely, "have you ever used CBT, meaning to give tests to students using CBT or have you conducted or carried out tests with CBT yourself?" The answer of each participant is P1: "I once gave it to students at that time

I used Google Forms, Kahoot, and Quizizz." P2: "Once, when the COVID-19 pandemic started, I used the Google Form to test mathematics for students" P3: "Once, I used the CBT Google Form" and P4: "No, I'm already a senior, so it's difficult to operate a computer and the like". From the answers the four participants can be concluded that mathematics lecturers usually prefer CBT or conventional.

These insights reveal that while some lecturers have experience with CBT, there is a notable reluctance among others, particularly those less familiar with technology.

4.3.4. Preference for Conventional Testing

Q3 is "in your opinion, which is better to use CBT or conventionally, like we used to use pencil and paper?" Each participant's answer is P1: "In my opinion, all of them have their own perspectives, if the advantages of using CBT itself make it easier for me as a lecturer to carry out the assessment process and also save time, especially during the pandemic period where face-to-face cannot be done on campus so the exam uses CBT. But another obstacle when using CBT is that there are still some students who have problems how to operate computers and their density and also if they don't have a data package." P2: "CBT requires a data package, whereas conventional requires paper and a printer/photocopy, so each one is used according to existing conditions." P3: "If CBT is more practical, fast and efficient" and P4: "I prefer conventional because it is easier and more familiar to me, CBT is not yet common". From the answers of the four participants, it can be concluded that the lecturer in mathematics believes that both CBT and conventional each have advantages and disadvantages.

Overall, lecturers expressed a mixed preference for testing methods, balancing the benefits of technology against their comfort with traditional assessments.

4.3.5. Identifying Weaknesses in Conventional Testing

Q4 is, "What are the weaknesses of the conventional method using pen and paper, sir/madam?" Each participant's answer is P1: "Yes, if the conventional-based test also has advantages and disadvantages, I feel that the assessment process is not as efficient as using CBT, then face-to-face is needed and also more wasteful of paper", P2: "Waste of paper and costs a printer/photocopy of test questions, besides correcting student answers it is also manual, the process usually takes a long time to correct, especially the description test." P3: "It takes more time for respondents to fill out, costs are more expensive, besides distributing it and taking it is more difficult." and P4: "If the multiple-choice test is easier to correct, but if the old description test corrects it". From the answers of the four participants, it can be concluded that the lecturer in mathematics believes that the weakness of the conventional method is that it wastes time and money.

From these answers, it is evident that lecturers recognize the limitations of conventional testing methods, including time inefficiencies and resource wastage.

4.3.6. Assessment and Grading Practices

Q5, namely, "how to determine student scores after students take the CBT test?" Each participant's answer is P1: "If you use Kahoot, it's already been determined during the manufacturing process which answer is the right answer, so when I download the results from the application, the score is already listed as well as Quizizz. But for the Google Form, I first downloaded the Excel file and then I entered it into the Excel formula to form or generate scores for each student but this only applies to multiple-choice questions that I can use, if the description questions I still ask students to write on the paper are then photographed and uploaded to another application, such as WhatsApp." P2: "Student scores are the same as the scores obtained by multiplying the total score multiplied by 100." P3: "Student scores are the same as the scores obtained by the total score multiplied by 100 " , and P4: "Never, maybe it will be printed and then I will manually correct it and determine the value". From the answers of the four participants, it can be concluded that the lecturer in mathematics determines the student's score as usual, final score = (score obtained/total score) x 100.

These responses highlight a reliance on traditional grading methods suggesting a need for training in more modern assessment techniques.

4.3.7. Approach to Difficulty Levels in Questions

Q6 is "when making the question package, did you have a difficulty level for each question?" The answers of each participant are P1: "Yes sir", P2: "Yes, I made some easy, medium, and difficult", P3: "Yes, I made all the level of difficulty equal, I prefer to make description questions that are all at the same level the difficulty and the score is the same for each question", and P4: "Yes, I will adjust it according to the ability of my students". From the answers of the four participants, it can be concluded that the lecturer in the field of mathematics makes a package of questions containing questions of different levels of difficulty based on the lecturer's own perspective, without CTT or IRT item analysis.

All participants affirmed that they do consider difficulty levels but with varying degrees of structure and rigor. For example, P1 and P2 create questions of varying difficulties while P4 adjusts questions based on students' abilities.

4.3.8. The Questions Based

Q7 is " what are the questions that you make based on what you are, if not based on the level of difficulty, based on the test grid based on indicators like that?" Each participant's answer is P1: "Yes, that's right, so I will first describe the indicators then I will develop the items that will be raised. Also, see that I didn't analyze the results of last year's exam because every year I change the questions and I only see the level of difficulty from Bloom's Taxonomy", P2: "If I have more time then I make a grid, but if I'm in a hurry then I just make questions according to the material the students are learning." P3: "Sometimes I make a grid first, and sometimes I just go straight to make questions according to the material studied by students, because they are used to it and understand because they have taught for many years." and P4: "Yes, I have taught for years, I am used to making test questions, or you

can also use test questions from the previous year. the previous year". From the answers of the four participants, it can be concluded that the lecturer in mathematics makes test questions based on the grid if they are not busy.

Participants emphasized the importance of test grids and indicators, with some acknowledging the influence of previous years' exams on their question design.

4.3.9. Personal Bias in Assessing Question Difficulty

Q8 is "the level of difficulty is based on your own perspective or subjectivity, do you feel which questions are difficult and which ones are easy, based on personal opinion?" The answers of each participant were P1: "Yes, sir, based on my understanding", P2: "Yes, that's right, I was wondering whether the questions I made were easy, medium, or difficult for my students, because I feel that I understand the abilities of my students", P3: "Yes, that's right, I guess, because I have taught for many years, I feel that I understand the abilities of my students.", and P4: "Yes". From the answers of the four participants, it can be concluded that the lecturer in mathematics estimates the level of difficulty of the questions made based on a personal perspective.

All participants indicated that their assessments of difficulty are indeed subjective based on their understanding of student capabilities.

4.3.10. Disconnect Between Perceived and Actual Student Ability

Q9, which is "based on your experience, have you ever felt that you made an easy question but it turns out that the students couldn't answer it?" The answers of each participant were P1: "Yes, I once found that easy questions could not be answered by students correctly because the diversity of my students' abilities was quite striking in the class, there were those with high, medium and low abilities", P2: "Once a few times, I didn't expect it turns out that there are my students who don't understand the basic concepts, so they can't answer even simple questions, maybe I should pay more attention and help those students who are still lacking in the future." P3: "Once a few times, I didn't think that there were students who were not able to answer easy questions, maybe students are lazy to read or study or their students' abilities are indeed lacking". P4: "Yes, but I have predicted it, because there are indeed some students in the class who are very weak in their abilities." From the answers of the four participants, it can be concluded that the lecturer in the field of mathematics once made a question that had an easy level of difficulty but turned out to be a difficult level when it was done by students.

Participants expressed frustration at this disconnect, indicating that despite their intentions, students often struggled with what were deemed easy questions.

Q10, which is "means that you mean that students need to be given questions that match their abilities, can't they give questions carelessly?" Each participant's answer was P1: "For students who have high abilities, maybe for them it is an easy question, but for my students who have low abilities, they are easy questions that could be considered as difficult questions so that during the evaluation there are still some that are not answered. Answer the easy question with the wrong answer. If for multiple- choice questions, yes, that's the assessment if you answer nine, it means you get 90 if the maximum score is 100. For the questions in the description, I give weights based on the level of difficulty of each and there is an assessment rubric for each concept requested from the question", P2: "Yes, that's true but in fact I don't think the problem is wrong but it's my fault that I haven't been able to teach the basic concepts properly, so there are my students who haven't been able to answer the questions that I think are correct. The package of questions that I made is for standard 1 class, I feel that whoever my student is, at least he should be able to work on the questions that I make, because usually the questions I make are not difficult, only measure the minimum abilities of students." P3: "Yes, that's right", and P4: "Yes". From the answers of the four participants, it can be concluded that the lecturer in mathematics is of the opinion that students should be given items according to their abilities.

Responses confirmed a shared belief in the necessity of aligning question difficulty with student capabilities to ensure fair assessments.

4.3.11. Fairness in Grading Practices

Q11, which is "back to the process of giving student grades, sir / madam, it means that all students, for example, answered nine out of 10 numbers, and the score is 90. In your opinion, is that fair because each student can answer a question with a different level of difficulty?" Each participant's answer is P1: "Yes, it could be unfair because I also divide the multiple-choice questions for medium and difficult questions but the other important points remain the same if yes, it could be unfair because for multiple-choice questions, I also give medium-young questions and it's difficult but other important points remain the same if it's correct eight means the value is 80 for a maximum score of 100", P2: "I think it's fair, answering 9 numbers out of 10 numbers means it's fair to get a score of 90, if it's multiple-choice yes because every correct answer is given score of 1, the story is different from the description test which has an assessment rubric.", P3: "I think it's fair", and P4: "Yes". From the answers of the four participants, it can be concluded that the lecturer in mathematics is of the opinion that the process of giving final grades to students that has been used so far has not been fair because it does not pay attention to the level of difficulty of the test questions that can be done by students.

Participants recognized the potential for unfairness in their current grading practices, especially concerning multiple-choice assessments.

Q12, which is "meaning for the case of description or essay questions, there is no problem, sir/madam, because there is a fair assessment rubric, but for multiple choice tests, it means that there is a possibility of unfairness, sir/madam, student scores?" Each participant's answer was P1: "Yes, how else can it be, that's why in addition to questions and doubles, I also gave questions to distinguish the level of student ability." P2: "Yes, because between one question and another, the difficulty level is actually different." P3: "That's true too", and P4: "Yes, I rarely use multiple choice tests". From the answers of the four participants, it can be concluded that the lecturer in mathematics is of the opinion that the scoring of students on the description test is not a problem/it is fair, but for the multiple-choice test it is still not fair.

The consensus indicated a belief that descriptive tests could be assessed more equitably, while multiple-choice assessments posed challenges due to varied difficulty levels.

4.3.12. Need for Improved Assessment Tools

Q13 is "it means that CBT uses Google Forms if it is in the form of a multiple-choice test, it still has shortcomings, sir / madam in determining the final score of students because it is not based on the level of difficulty of the questions". The answers of each participant were P1: "If there really is an application like that, it would be something extraordinary in my opinion if he could give a fair assessment according to the ability of the student." P2: "Yes, that's true, but is there a new way that is better?", P3: "Yes, that's right", and P4: "Yes, like that". From the answers of the four participants, it can be concluded that mathematics lecturers do not know that there is a CAT that is able to give a final score that is fairer than CBT and conventional. Participants acknowledged the limitations of CBT in fairly assessing student abilities based on question difficulty.

Q14, namely "if there is an application that is more thorough than CBT which is as easy as CBT to use but in the process the student will get questions according to his abilities and he will get a score according to the level of difficulty of the questions he is working on, what do you think, if any? Such an application?" Each participant's answer was P1: "I'm interested but is it as easy as making CBT? If I can CBT", P2: "Good, the student test scores can be fairer." P3: "I want to use it if it is easy to use" and P4: "Yes, that's good but not all students here have internet packages and cell phone, plus the study programs or faculties do not yet have adequate computer labs." From the answers of the four participants, it can be concluded that lecturers in mathematics are interested in using CAT.

There was enthusiasm among participants for a more sophisticated assessment tool, provided it was user-friendly and accessible.

4.3.13. Willingness to Implement CAT

Q15, namely, "whether the application seems to be able to be realized with a computerized adaptive test or CAT, do you have the desire to make CAT?" Each participant's answers were P1: "If it's easy to use maybe I'll try." P2: "Want but busy is there anything I can use right away?", P3: "I want to but I'm not used to it, for my age this is a difficult job" and P4: "It seems difficult, I just want to use it if there is one, or ask a young lecturer for help to enter my test questions into the application". From the answers of the four participants, it can be concluded that the lecturer in mathematics wants to make a CAT application if the process is easy. Responses indicated a willingness to explore CAT, provided the process was straightforward and manageable.

Interview Question 16 (Q16): "What challenges do you anticipate in using online tests like CAT?"

Q16, namely, "basically, CAT is similar to CBT in its implementation, which requires an internet network computer device, and the most important thing is that the application is available according to you. What are the difficulties in using an online test like this earlier?" Each participant's answer was P1: "Based on my experience giving a test using CBT, the student was not 100% able to take it because at the time of the pandemic, there were one or two students who did not have an internet connection at home so they could not take CBT if indeed CAT was similar to CBT in terms of implementation, it means that the shortcomings will probably be more or less the same as CBT", P2: "It's difficult if we as lecturers are busy, we want a fast test application, and it seems that CAT is still rarely able to make it, yes, it might be difficult to make it. Then we in Papua, not all students have cell phones and internet packages that are adequate for online tests, not to mention if you look at the existence of a computer lab on campus that has not been able to accommodate the needs of all students to take online tests." P3: "I personally haven't used to making such applications, so I can imagine the difficulty, and for applications related to technology, I usually ask my younger colleagues for help to make it faster and more efficient. In addition, our study programs, departments, or faculties do not yet provide adequate computers". P4: "It's difficult here because of the limited number of students and the infrastructure, not all students can be active in online lectures. Maybe it can be done specifically for students who have cell phones and internet packages." From the answers of the four participants, it can be concluded that the lecturer in mathematics believes that there are several weaknesses in the use of online tests such as CBT or CAT. Weaknesses include network availability, laptops/computers/handphones, question banks and CAT applications. Participants highlighted infrastructural challenges, including internet access and technology availability, particularly in Papua.

From the insights gathered from the four participants, it is clear that mathematics lecturers identified several key areas for improvement in the use of online assessments, such as computer-based testing (CBT) and computerized adaptive testing (CAT). They highlighted network availability as a significant barrier, noting that unreliable internet access often prevents students from participating in assessments. Additionally, the accessibility of devices such as laptops, computers, and smartphones were a concern, as not all students have the necessary equipment to engage in online testing, especially in economically disadvantaged areas. The lecturers also emphasized the need for a robust question bank tailored to the CBT and CAT formats, as the lack of a well-organized repository can lead to inconsistencies in assessments. Finally, challenges related to the usability of CAT applications were mentioned with participants stressing the importance of ease of use for both lecturers and students; complex applications can hinder engagement. These insights can be organized into themes that align with the questionnaire results, reflecting a cohesive understanding of the challenges in adopting online testing methodologies and underscoring the necessity for targeted improvements to enhance the effectiveness and fairness of mathematics assessments.

In Table 6, the results of the interview conclusions were organized into sub-themes that correspond with the themes derived from the questionnaire results. The challenges faced by mathematics lecturers in utilizing computerized adaptive testing (CAT) can be categorized into four primary areas: (1) a lack of understanding regarding CAT which can hinder its effective implementation; (2) the necessity for adequate software and hardware to support CAT systems; (3) the requirement for a stable internet network, which is crucial for the functionality of online testing; and (4) the limitation that CAT cannot be utilized for essay tests.

Table 6. Grouping of themes based on interview conclusions.

Interview conclusion	Subtheme	Theme
1. Lecturers in mathematics need help understanding CAT. 2. Lecturers in mathematics make question packages containing questions of different difficulty levels based on the lecturer's perspective, without CTT or IRT item analysis. 3. Mathematics lecturers estimate the difficulty level of the questions based on personal perspectives. 4. A mathematics lecturer once created a question that was manageable in difficulty but difficult for students to answer. 5. Mathematics lecturers determine student grades as usual, final grade = (Score obtained/ Total score) x 100	Lack of understanding of mathematics lecturers regarding CAT.	Challenges: He uses CAT in mathematics tests to assess the performance of students he has taught and to measure the mathematical abilities of students he will teach in the future.
1. Lecturers in mathematics usually prefer to use conventional or CBT. 2. Mathematics lecturers believe that CBT and conventional have advantages and disadvantages. 3. Mathematics lecturers believe that online tests such as CBT or CAT have several areas for improvement. These weaknesses include network availability, laptops/computers/hand phones, question banks, and CAT applications.	1. CAT requires adequate software and hardware. 2. CAT requires an adequate internet network. 3. CAT cannot be used for essay tests.	
1. Lecturers in mathematics argue that the conventional method's weakness is that it wastes time and money. 2. Lecturers in mathematics make test questions based on the grid if they are not busy. 3. Lecturers in mathematics believe that students should be given test questions according to their abilities. 4. Lecturers in the field of mathematics believe that the process of giving final grades to students that have been used so far has not been fair because it does not consider the level of difficulty of the test questions that students can do. 5. Lecturers in mathematics are of the opinion that students' scoring on the essay test is not a problem and is fair, but it is still not fair for the multiple-choice test. 6. Some mathematics lecturers do not know that the CAT can give a fairer final score than CBT and conventional. 7. Mathematics lecturers are interested in using CAT, especially if it is easy and cheap. A mathematics lecturer wants to make a CAT application if the process is easy.	Mathematics lecturers recognize the advantages of CAT and are interested in using it if it is easy and cheap.	He needs to develop an interest in using CAT to carry out assessments that provide fairer and more accurate scores in measuring the mathematical abilities of the students he has taught and the prospective students he will teach.

Besides the challenges in using CAT, the results of the interviews also produced other sub-themes which showed that mathematics lecturers were actually very interested in the advantages of CAT and needed a CAT application that was easy to use and inexpensive. Thus, the answers to in-depth interviews in this study resulted in 2 major themes, namely (1) the challenges of using CAT in mathematics tests and (2) the need and interest of mathematics lecturers to use CAT.

5. Discussion

As explained in the previous section, the results of the study are in 2 major themes, namely (1) the challenges of using CAT in mathematics tests and (2) the need and interest of mathematics lecturers to use CAT. Further discussion regarding the research results as follows:

5.1. Mathematics Lecturer Challenge to use CAT in Mathematics Tests

This theme was obtained from the results of filling out a questionnaire and the results of in-depth interviews. The results of the two data collection techniques yielded the following conclusions.

5.1.1. Lack of Understanding of Mathematics Lecturers Regarding CAT

One of the predominant challenges identified was the lack of understanding among mathematics lecturers regarding CAT. This difficulty often stems from the complexity of CAT systems, which can lead to confusion and a lack of clarity about their operational processes. Furthermore, the limited availability of accessible information and resources on CAT exacerbates this issue. Many mathematics lecturers expressed that they found it challenging to understand how to implement CAT effectively, leading to a preference for more familiar assessment methods, such as conventional techniques (paper and pencil) or, during the COVID-19 pandemic, computer-based testing (CBT) using platforms like Google Forms.

Although CBT offers certain psychometric advantages as noted by Triantafillou et al. (2008) which include enhanced standardization, increased test security, improved display capabilities, reduced measurement errors, and faster scoring and interpretation it shares fundamental similarities with traditional paper-based testing (PBT). Both CBT and PBT present the same test items to all test-takers. In CBT, the estimation of a test-taker's ability is performed using a classical approach, where the number of correctly answered, questions is compared to the total number of questions completed. This lack of adaptability in traditional testing formats contrasts sharply with the tailored approach of CAT which dynamically adjusts to the test-taker's ability level, thereby offering a more personalized assessment experience.

Furthermore, regarding CBT, psychometrically, some of the advantages of CBT as stated by Yulianto (2008) are (1) increasing standardization, (2) increasing test safety, (3) increasing test display ability, (4) reducing error of measurement, and (5) speeding up scoring and interpretation. However, in principle, CBT is the same as PPT. The test takers receive the same set of test items. The working principle of CBT only transfers the PPT paradigm to

the computer screen. Estimation of the ability of test takers in CBT is done classically, namely by counting the number of questions that can be answered correctly compared to all the questions done by the test takers. Meanwhile, in contrast to CBT, according to Hadi and Romadhon (2013) adaptive testing is a test held for test participants with the questions/items determined based on the participants' initial answers/responses. This is what distinguishes CAT from PPT and CBT. According to Santoso (2010) CAT is not just moving items into the computer, but the computer is set to select, present items according to the estimated level of ability of test takers, and calculate test taker's answer scores. Another mathematics lecturer said that he had never heard of and used CAT, so he did not know the difficulties in using it. This shows that information about CAT has not been extended to all regions of Indonesia. In addition, it is estimated that because of the heavy workload and busyness of the mathematics lecturers, they do not have more time to add knowledge about the latest information in the realm of assessment.

In addition, there was a mathematics lecturer who stated that the difficulty of making CAT caused a reluctance to study it and hoped that there would be a ready to-use CAT application, making it easier to use. Meanwhile, another mathematics lecturer is of the opinion that someone who uses the CAT application in an assessment must really understand the basics of CAT first, otherwise, it will confuse them and will not be able to account for it.

5.1.2. CAT Requires Adequate Software and Hardware

The challenges of using CAT are reviewed in terms of facilities and infrastructure. According to the mathematics lecturer, for exams using CAT, students need a laptop or computer while on campus the computer laboratory is inadequate. CAT as an information system is a structured system that relies on computer hardware and software technology to collect, process, store and generate information. The development of CAT follows the criteria for developing an information system by adopting the waterfall model which is also known as the linear sequential model (Pressman, 2012).

It's not enough to stop there; other mathematics lecturers are of the opinion that in using CAT, apart from requiring hardware, it also requires software to run CAT. In fact, it must also be supported by the availability of a question bank. Not everyone is able to develop CAT software because it requires special skills. The difficulty of making CAT applications is due to the nature of artificial intelligence (AI), so that CAT applications have human-like intelligence. The use of CAT in every test has the aim of utilizing the invariant property of IRT in creating an algorithm, namely, each test taker will receive test items that have been adjusted to the individual abilities of the test takers so that the questions given are not questions that are too difficult or too easy for individuals test takers (Embretson & Reise, 2013).

The same thing is also stated by Özyurt, Özyurt, Baki, and Güven (2012) that adaptive assessment provides an opportunity to find out students according to their ability level. An expert system is one of the applications of artificial intelligence. The basic principle of intelligent systems is to adopt human knowledge to computers, so that computers have the ability to complete work or solve certain problems as done by an expert (Turban, Aronson, & Liang, 2005).

In addition, the creation of a question bank also takes a lot of time because the test items that enter the question bank must first be tested for quality using IRT analysis. The importance of this question bank is in accordance with what is stated by Masters (1999) which states that in an adaptive test, test items are selected from the question bank based on the rules for selecting test items that have been previously determined. Computers with programmable capabilities have high ease of use in the adaptive testing process.

In line with the previous opinion regarding the question bank, other mathematics lecturers argue that the process of developing an instrument in the form of test questions that are ready to be used takes a long time and requires a large number of trial samples. In addition, there are mathematics lecturers who argue that if there is no laptop or computer, both lecturers and students have difficulty typing equations in online tests if they only use Android phones. The development of adaptive tests needs to pay attention to and evaluate six components, namely: 1) item response model, 2) question bank, 3) initial item selection, 4) method of estimating the level of ability, 5) procedure for selecting items, 6) rules for stopping the test (Wise & Kingsbury, 2000). When developing the CAT, the criteria for the six components must be considered, because each has different criteria and options.

In contrast to computerised-based testing (CBT), adaptive testing is distinguished by its ability to adjust questions based on the participant's initial responses (Hadi & Romadhon, 2013). This feature sets computerized-based testing (CAT) apart from both traditional paper- pencil testing (PPT) and CBT. While CBT mainly involves transferring test items to a digital format, CAT uses algorithms to select and present questions that are tailored to the estimated skill level of each individual test taker. This adaptive approach not only improves the testing experience but also allows for a more precise assessment of a student's abilities, as the system dynamically adjusts according to the participant's performance throughout the test (Tang et al., 2021).

5.1.3. CAT Requires Adequate Software and Hardware

A key factor in the successful implementation of CAT is ensuring the availability of sufficient software and hardware. As noted by one mathematics lecturer, students must have access to laptops or computers for CAT-based exams. However, many campus computer labs require substantial upgrades to meet this need effectively. As an information system, CAT operates as a structured framework that depends on computer hardware and software technologies to gather, process, store, and produce relevant data. The development of CAT systems follows established guidelines for information system development, frequently using the waterfall model or linear sequential model as described by Quinlan (2008).

In addition to the requirement for suitable hardware, other mathematics lecturers emphasized the importance of advanced software to support the functioning of CAT. Furthermore, a comprehensive question bank is crucial to ensure the system can provide a diverse and challenging assessment experience. Developing effective CAT software is a complex undertaking that demands specialized skills and expertise which are not always readily accessible.

The complexity of designing CAT applications is further intensified by the inherent nature of artificial intelligence (AI). CAT systems utilize AI to simulate human-like intelligence, allowing them to adaptively evaluate

test-takers. The main goal of incorporating CAT into assessments is to leverage the consistent characteristics of Item Response Theory (IRT) to create an algorithm that customizes the testing experience. This approach ensures that each participant receives questions suited to their individual ability levels, thereby preventing situations where they are presented with questions that are either too challenging or too simple (Yim et al., 2024). Furthermore, research from Şenel and Kutlu (2018) affirm that adaptive assessment provides a valuable opportunity to accurately assess students' ability levels (Şenel & Kutlu, 2018).

An expert system, a form of artificial intelligence is based on the principle of incorporating human knowledge into computer systems, enabling computers to perform tasks or solve specific problems with expertise comparable to that of a human expert. While the challenges of implementing CAT are substantial, the potential benefits it offers, particularly in terms of personalized assessments and a more accurate evaluation of student abilities, are undeniable. Overcoming the current barriers to understanding and accessibility will be essential for fostering wider adoption of CAT in mathematics education.

Creating a question bank is a substantial task that demands considerable time and effort. Before test items can be added to the bank, they must undergo rigorous quality testing through item response theory (IRT) analysis. This process ensures the items are both reliable and valid, which is essential for adaptive testing. The selection of test items in an adaptive test follows predefined rules, with the items drawn from the established question bank. The use of computers with programmable capabilities significantly streamlines the adaptive testing process, enabling the smooth application of these item selection rules. Additionally, to fully leverage the potential of CAT, the question bank must include items with high reliability estimates, as emphasized by researchers such as Ebenbeck and Gebhardt (2024) and Mujtaba and Mahapatra (2021). In addition to validity, reliability, and item characteristics, differential item functioning (DIF) is also a crucial consideration for CAT. DIF analysis can identify items that may favor specific groups, such as those based on gender. For instance, a study revealed 36 items affected by DIF in Indonesia's national exams, indicating a bias towards certain groups (Setiawan, Kassymova, Mbazumutima, & Agustyani, 2024). This reliability is essential to ensure that the adaptive assessment accurately reflects a test taker's ability. Technology in education is crucial, particularly in assessment activities. Research indicates that the interaction between the ease of access to digital devices in schools and the completeness of learning facilities positively impacts students' academic achievements, especially in mathematical literacy (Liestari & Muhardis, 2021).

In agreement with the importance of a robust question bank, several mathematics lecturers have expressed that developing effective test items is a time-intensive process that necessitates the use of numerous trial samples. Additionally, they highlighted that access to technology poses a challenge. If either lecturers or students lack a laptop or computer, they face difficulties in inputting equations for online tests, particularly when only using Android phones. Triantafillou et al. (2008) outline the essential components to consider in the development of adaptive tests, which include: (1) item response model, (2) question bank, (3) initial item selection, (4) method of estimating ability levels, (5) procedure for item selection, and (6) rules for terminating the test. It is vital to address each of these components with distinct criteria and options when developing CAT, as their individual characteristics significantly impact the overall effectiveness of the assessment.

5.1.4. CAT Requires Adequate Internet Network

Another critical challenge associated with using CAT for testing is its reliance on an online platform, necessitating a stable internet connection. While some mathematics lecturers believe that CAT is relatively straightforward to implement, they underscore the potential complications that arise from internet disruptions. In various regions of Indonesia, particularly in remote areas, reliable internet access can be problematic. For instance, in Papua Province, connectivity issues may persist for weeks due to damage to underwater optical cables, significantly hindering the ability to conduct online assessments.

5.1.5. CAT Cannot Accommodate Essay Tests

A further limitation of CAT is its current inability to facilitate descriptive tests effectively. Several mathematics lecturers prefer using descriptive assessments over multiple-choice formats, believing that such tests provide a more accurate measurement of students' mathematical abilities by reducing the influence of guessing. This reliance on descriptive tests emphasizes the need for CAT systems to evolve and integrate formats that allow for comprehensive evaluation of student understanding, which currently remains a challenge.

5.2. Needs and Interests of Mathematics Lecturers to Use CAT in Mathematics Tests

The next central theme that emerged from the research results though not initially articulated in the study's objectives centres on the expressed needs and interests of mathematics lecturers in employing CAT for mathematics assessments. The goal is to achieve fairer and more accurate evaluations of students' mathematical capabilities, whether for those they have taught or for prospective students they may teach in the future. This revelation highlights one of the strengths of qualitative phenomenological research which allows for an in-depth exploration of participants' experiences, feelings, and expectations many of which might not have been anticipated by the researcher prior to conducting the study (Şenel & Kutlu, 2018). Participants shared their thoughts candidly, resulting in intriguing findings that contribute significantly to the understanding of the potential for CAT in mathematics education. The genuine interest in using CAT reflects a desire among lecturers to enhance their assessment strategies and improve student outcomes, thus showcasing a readiness to embrace innovative approaches to teaching and evaluation in the mathematics domain.

Mathematics lecturers involved in this study face significant challenges in understanding, utilizing, and developing CAT applications. However, despite these obstacles, they recognize the benefits that CAT offers, particularly its ability to assess test takers more fairly compared to traditional paper- based testing (PBT) or computer- based testing (CBT) methods. The interest in CAT among mathematics lecturers is driven by its ability to customize test questions based on the abilities of the test takers, resulting in more accurate and skill-based final scores. One key advantage of CAT is its capacity to reduce boredom for high-ability test takers by preventing them

from encountering overly simple questions, while ensuring that low-ability test takers are not overwhelmed by excessively difficult ones. This adaptive feature is consistent with findings from Walker, Böhnke, Cerny, and Strasser (2010) which highlight the importance of matching test difficulty with the individual capabilities of each student.

Moreover, CAT can be tailored through various adjustments to assess a wide range of mathematical competencies, including verbal, non-verbal, and numeracy skills, offering a comprehensive evaluation of student abilities (Ebenbeck & Gebhardt, 2024). Mathematics lecturers express a strong desire for a future CAT application that is both user-friendly and cost-effective. Such an application could greatly improve the quality of measurement, assessment, and evaluation processes in education overall. This viewpoint underscores the potential of CAT-based assessments to enhance efficiency and accuracy while ensuring practicality in their implementation.

Given its advantages, CAT stands out as a powerful tool for evaluating student capabilities. It offers high accuracy with fewer questions, ensuring that both efficiency and test quality are maintained. Additionally, CAT shortens examination times and reduces the potential for collusion among students, as each participant receives a unique set of questions tailored to their individual abilities. Furthermore, research by Istiyono et al. (2018) indicates that CAT can effectively manage (a) question banks, (b) automatically organize test items based on students' abilities, (c) randomly generate answer choices, and (d) keep records of test results, both individually and collectively.

The benefits of administering tests through computers, particularly CAT are also emphasized by Stepanek and Martinkova (2020) who highlight several advantages, including (1) flexible test management, (2) enhanced test security, (3) increased student motivation, and (4) a reduction in testing time by up to 50%, all while maintaining a consistent level of reliability. Triantafillou et al. (2008) further assert that CAT generates the most informative data in assessing test takers' abilities. The intelligent algorithms embedded in CAT systems enable the selection of test items that are aligned with students' proficiency levels. Study in Africa provides empirical evidence for selecting ability estimation methods for CAT as part of efforts to design accurate testing programs for use in higher education (Oladele, Ndlovu, & Spangenberg, 2022).

The findings from this study provide valuable insights into the challenges and opportunities of using CAT in mathematics education. The interest expressed by mathematics lecturers in adopting CAT reflects a readiness to embrace new, innovative assessment methods that have the potential to enhance educational outcomes. By tackling the challenges identified, creating more user-friendly applications, and leveraging the unique advantages of CAT, mathematics education can shift towards a more flexible and personalized assessment system that better captures students' true mathematical abilities. As the field of education continues to evolve, it's important for all stakeholders to stay committed to exploring and implementing adaptive assessment approaches that not only support educators' needs but also foster meaningful learning experiences for students. In addition to being supported by CAT-based assessments, learners should also be provided with learning media that can enhance their abilities or skills so that the results can be maximized. There is research that develops a problem-based local history module as a learning media to improve students' critical thinking skills. The problems presented in the module are contemporary and related to the students' environment. The process of connecting new knowledge with existing knowledge allows students to develop their own understanding and improve their critical thinking skills in problem-solving (Mujiyati, Warto, & Sutimin, 2019).

6. Conclusion

The results of this study indicate that the current landscape of mathematics assessment at universities in Indonesia predominantly relies on traditional (paper and pencil) methods and computer- based testing (CBT). Online mathematics assessments have gained increasing prominence, especially since the onset of the COVID-19 pandemic in early 2020 as educational institutions adapted to remote learning. Despite the shift towards digital assessments, the integration of more advanced assessment technologies, such as computerized adaptive testing (CAT) remains limited.

The study identifies several challenges faced by mathematics lecturers in implementing CAT, which can be grouped into four primary themes. First, there is a significant lack of understanding among lecturers regarding CAT, its purpose, and its potential benefits. Second, the effective use of CAT requires adequate software and hardware, which are not always available or accessible in many institutions. Third, a reliable and stable internet network is essential for the successful implementation of CAT, but many universities in Indonesia face connectivity issues. Finally, CAT has limitations in accommodating descriptive or open-ended questions, which are integral to assessing higher-order thinking skills in mathematics.

Despite these challenges, the findings highlight a strong demand for user-friendly and cost-effective CAT applications that can deliver fairer, more personalized, and accurate assessments of students' mathematical abilities. Lecturers and educational stakeholders recognize the potential of CAT to improve the quality of assessment by adapting to individual student abilities, offering a more tailored and efficient evaluation process. The demand for such tools underscores the need for further innovation in the development of accessible CAT systems that can meet the unique needs of the Indonesian educational context.

Based on the findings, this study recommends the urgent development of an accessible and affordable CAT application specifically designed for use in Indonesian universities. Given that CAT applications remain relatively scarce in the country and awareness of their advantages is still limited, it is crucial to disseminate information regarding the benefits of CAT in enhancing educational assessments. Raising awareness about these advantages, coupled with fostering innovation among the younger generation of developers will be vital in creating CAT systems that address the challenges identified in this study.

The adoption of CAT in Indonesian higher education has the potential to significantly improve the quality and fairness of mathematics assessments. CAT could play a pivotal role in enhancing the overall quality of mathematics education in Indonesia by ensuring that assessments are more aligned with individual students' learning needs. It is essential to support the development of affordable, accessible, and effective CAT solutions that not only address the current challenges but also foster a more equitable and efficient educational environment for future generations.

7. Recommendations

Based on the findings of this study, several recommendations can be made to improve the implementation of computerized adaptive testing (CAT) in mathematics education in Indonesia. First, it is crucial to invest in comprehensive training programs for mathematics lecturers to enhance their understanding and proficiency in using CAT. These programs should focus on both the technical aspects of CAT systems and their pedagogical applications, ensuring that lecturers are equipped with the necessary skills to effectively integrate CAT into their assessment practices.

Second, efforts should be made to improve the technological infrastructure at Indonesian universities. This includes providing adequate hardware and software that support CAT as well as ensuring stable and reliable internet connectivity. Public and private stakeholders, including government institutions and technology providers, should collaborate to address these infrastructure challenges, making CAT more accessible to a wider range of universities.

Third, the development of user-friendly and cost-effective CAT applications should be prioritized. These applications must be tailored to the specific needs of the Indonesian educational context, ensuring that they are both affordable and easy to use for lecturers and students alike. By creating applications that are more adaptable to different teaching and learning environments, the adoption of CAT can be facilitated more effectively.

8. Implications

The findings and recommendations of this study have several important implications for the future of mathematics education and assessment in Indonesia. First, the integration of CAT into the assessment process has the potential to significantly enhance the fairness and accuracy of mathematics evaluations, offering a more personalized approach that adjusts to the abilities of individual students. This could lead to more reliable and equitable assessments, ensuring that students are evaluated based on their true abilities rather than being subject to biases inherent in traditional testing methods.

Moreover, the adoption of CAT could foster a shift towards more technology-driven and innovative assessment practices in Indonesian universities. This would not only improve the quality of mathematics education but also better prepare students for the demands of a technology-based workforce. As such, the successful implementation of CAT could serve as a model for other disciplines, encouraging the adoption of advanced assessment tools across various fields of study.

Finally, addressing the challenges identified in this study and implementing the recommended solutions would have a broader impact on the overall quality of higher education in Indonesia. By embracing modern assessment technologies, universities can provide a more effective and efficient learning environment that meets the diverse needs of students and contributes to the long-term development of the nation's educational system.

References

Abadyo, A., & Bastari, B. (2015). Estimation of ability and item parameters in mathematics testing by using the combination of 3PLM/GRM and MCM/GPCM scoring model. *REiD (Research and Evaluation in Education), 1*(1), 55-72. https://doi.org/10.21831/reid.v1i1.4898

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The effect of computerized adaptive testing on motivation and anxiety: A systematic review and meta-analysis. *Assessment, 30*(5), 1379-1390. https://doi.org/10.1177/10731911221100995

Boo, J., & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports: Mental & Physical Health, 1*(4), 443–460. https://doi.org/10.2466/01.08.21.PR0.110.4.443-460

Castillo-Paredes, A. (2021). The training of the physical education teacher in special educational needs, towards reflection and action, from international agreements and laws in Chile. *Retos, 44*, 709–715. https://doi.org/10.47197/RETOS.V44I0.91283

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1-38. https://doi.org/10.18637/jss.v071.i05

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*(1), 37-53. https://doi.org/10.1177/0013164410387338

Cook, K. F., Choi, S. W., Crane, P. K., Deyo, R. A., Johnson, K. L., & Amtmann, D. (2008). Letting the CAT out of the bag: Comparing computer adaptive tests and an 11-item short form of the roland-morris disability questionnaire. *Spine, 33*(12), 1378-1383. https://doi.org/10.1097/BRS.0b013e3181732acb

Creswell, J. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks: SAGE Publications.

Dasopang, M. D. (2017). Study and learning. *Fitrah: Jurnal Kajian Ilmu-ilmu Keislaman, 3*(2), 333-352. https://doi.org/10.24952/fitrah.v3i2.945

Ebenbeck, N., & Gebhardt, M. (2022). Simulating computerized adaptive testing in special education based on inclusive progress monitoring data. *In Frontiers in Education Frontiers Media SA, 7*, 945733. https://doi.org/10.3389/feduc.2022.945733

Ebenbeck, N., & Gebhardt, M. (2024). Differential performance of computerized adaptive testing in students with and without disabilities–a simulation study. *Journal of Special Education Technology*, 1-10. https://doi.org/10.1177/01626434241232117

Embretson, S. E., & Reise, S. P. (2013). Item response theory. In: Psychology Press. https://doi.org/10.4324/9781410605269.

Fardhila, R. R., & Istiyono, E. (2019). An assessment instrument of mind map product to assess students' creative thinking skill. *Research and Evaluation in Education, 5*(1), 41-53. https://doi.org/10.21831/reid.v5i1.22525

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*(1-4), 59-76. https://doi.org/10.7333/1412-0204059

Fleming, D., Wilson, M., & Ahlgrim-Delzell, L. (2018). Using item response theory to describe the nonverbal literacy assessment (NVLA). *Psychology in the Schools, 55*(4), 341-349. https://doi.org/10.1002/pits.22110

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychological Assessment, 28*(8), 953. https://doi.org/10.1037/pas0000240

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology, 37*(2), 261-278. https://doi.org/10.1111/j.1467-8535.2005.00525.x

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Gunartha, I. W., Sulaiman, T., Suardiman, S. P., & Kartowagiran, B. (2020). Developing instruments for measuring the level of early childhood development. *Research and Evaluation in Education, 6*(1), 1-9. https://doi.org/10.21831/reid.v6i1.21996

Hadi, S., & Romadhon, T. (2013). *Development of web-based computerized adaptive test*: Aswaja Pressindo.

Haley, S. M., Coster, W. J., Dumas, H. M., FRAGALA-PINKHAM, M. A., Kramer, J., Ni, P., . . . Ludlow, L. H. (2011). Accuracy and precision of the pediatric evaluation of disability inventory computer-adaptive tests (PEDI-CAT). *Developmental Medicine & Child Neurology, 53*(12), 1100-1106. https://doi.org/10.1111/j.1469-8749.2011.04107.x

Hambleton, R., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. New York: Springer Science & Business Media.

Hamdi, S., Suganda, I. A., & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using lombok local cultures as contexts for junior secondary school mathematics. *Research and Evaluation in Education, 4*(2), 126-135. https://doi.org/10.21831/reid.v4i2.22089

Hsu, C.-L., & Wang, W.-C. (2019). Multidimensional computerized adaptive testing using non-compensatory item response theory models. *Applied Psychological Measurement, 43*(6), 464-480. https://doi.org/10.1177/0146621618800280

Huang, C.-Y., Chen, S.-S., Chen, C.-T., Lee, P.-S., Yu, T.-Y., & Chen, K.-L. (2020). Psychometric properties and efficiency of the Computerized Adaptive Testing System for measuring Self-Care Performance in Taiwanese children with developmental disabilities. *Archives of Physical Medicine and Rehabilitation, 101*(8), 1332-1337. https://doi.org/10.1016/j.apmr.2020.01.014

Hudha, S. A., & Mardapi, D. (2018). Developing an instrument for measuring the spiritual attitude of high school students. *Research and Evaluation in Education, 4*(1), 4. https://doi.org/10.21831/reid.v4i1.20304

Ibrahim, Z. S., Retnawati, H., Irambona, A., & Orantes Pérez, B. E. (2024). Stability of estimation item parameter in IRT dichotomy considering the number of participants. *Research and Evaluation in Education, 10*(1), 114-127. https://doi.org/10.21831/reid.v10i1.73055

Imawan, O. R., Retnawati, H., Haryanto, & Ismail, R. (2024). *Confirmatory factor analysis and differential item functioning analysis on mathematical literacy instruments for prospective Indonesian elementary school teachers*. Paper presented at the AIP ConferenceProceeding, 080009. https://doi.org/10.1063/5.0228174.

Imawan, O. R., Retnawati, H., Haryanto, H., & Ismail, R. (2025). Innovations in assessment methods: Computerized adaptive testing (CAT) for sustainable energy efficiency. Dalam Lecture Notes in Civil Engineering. In (2nd ed., Vol. 557, pp. 161–168): Springer Nature Link. https://doi.org/10.1007/978-981-97-8348-9_20.

Ismail, R., Imawan, O. R., & Nadhifah, A. (2023). *The influence of interest in learning mathematics on mathematics learning achievement of junior high school students in Papua*. Paper presented at the AIP Conference Proceedings, 2811(1). https://doi.org/10.1063/5.0142257.

Ismail, R., Retnawati, H., Arovah, N. I., & Imawan, O. R. (2024). Contexts proposed by teachers in Papua for developing mathematics HOTS assessment instruments: A phenomenological study. *Journal of Education and e-Learning Research, 11*(3), 548-556.

Ismail, R., Retnawati, H., & Imawan, O. R. (2022). Cluster analysis of the national examination: School grouping to maintain the sustainability of high school quality. *Research and Evaluation in Education, 8*(1), 13-23. https://doi.org/10.21831/REID.V8I1.45872

Ismail, R., Retnawati, H., Sugiman, & Imawan, O. R. (2024). *Construct validity of mathematics high order thinking skills instrument with cultural context: Confirmatory factor analysis*. Paper presented at the AIP Conference Proceeding, 080008. https://doi.org/10.1063/5.0228143.

Ismail, R., Retnawati, H., Sugiman, & Imawan, O. R. (2025). A sustainable development approach to math higher-order thinking skills: Culture and green technology. Dalam Lecture Notes in Civil Engineering. In (2nd ed., Vol. 557, pp. 137–143): Springer Nature. https://doi.org/10.1007/978-981-97-8348-9_17.

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *Research and Evaluation in Education, 5*(2), 95-102. https://doi.org/10.21831/reid.v5i2.20924

Istiyono, E., Dwandaru, W. S. B., & Faizah, R. (2018). Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT. *Research and Evaluation in Education, 4*(2), 144-154. https://doi.org/10.21831/reid.v4i2.22218

Jette, A. M., Haley, S. M., Tao, W., Ni, P., Moed, R., Meyers, D., & Zurek, M. (2007). Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. *Physical therapy, 87*(4), 385-398. https://doi.org/10.2522/ptj.20060121

Karimah, U., Retnawati, H., Hadiana, D., Pujiastuti, P., & Yusron, E. (2021). The characteristics of chemistry test items on nationally-standardized school examination in Yogyakarta city. *Research and Evaluation in Education, 7*(1), 1-12. https://doi.org/10.21831/reid.v7i1.31297

Kartianom, K., & Mardapi, D. (2017). The utilization of junior high school mathematics national examination data: A conceptual error diagnosis. *Research and Evaluation in Education, 3*(2), 163-173. https://doi.org/10.21831/reid.v3i2.18120

Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. *Research and Evaluation in Education, 5*(2), 169-182. https://doi.org/10.21831/reid.v5i2.23721

Khoshsima, H., & Toroujeni, S. M. H. (2017). Computer-based testing: Score equivalence and testing administration mode preference in a comparative evaluation study. *International Journal of Emerging Technologies in Learning, 12*(10), 35–55. https://doi.org/10.3991/ijet.v12i10.7159

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions, 14*, 12. https://doi.org/10.3352/jeehp.2017.14.12

Kristóf, K., & Eszter, T. (2016). Computerized adaptive IQ-testing. *Magyar Pszichologiai Szemle, 71*(1), 143–163. https://doi.org/10.1556/0016.2016.71.1.8

Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *Research and Evaluation in Education, 4*(1), 70-78. https://doi.org/10.21831/reid.v4i1.20202

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement, 72*(1), 159-175. https://doi.org/10.1177/0013164411411296

Liestari, S. P., & Muhardis, M. (2021). Hierarchical linear modeling for determining the effect of ict literacy on mathematics achievement. *Research and Evaluation in Education, 7*(1), 78-87. https://doi.org/10.21831/reid.v7i1.39181

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come (No. 69, p. 58). MESA memorandum*. Retrieved from https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf

Liu, Z., Wu, J., Yao, F., Liu, Q., Chen, E., Sha, J., . . . Su, Y. (2024). Computerized adaptive testing method based on reinforcement learning for series diagnosis. *Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence, 37*(1), 13–26. https://doi.org/10.16451/j.cnki.issn1003-6059.202401002

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21*(6), 623-634. https://doi.org/10.1016/0883-0355(94)90015-9

Mardapi, D., Haryanto, H., & Hadi, S. (2012). Pengujian hasil belajar dan penilaian pendidikan berbantuan komputer. *Jurnal Kependidikan Penelitian Inovasi Pembelajaran, 42*(2), 130–143. https://doi.org/10.21831/jpai.v10i1.923

Marfeo, E. E., McDonough, C., Ni, P., Peterik, K., Porcino, J., Meterko, M., . . . Chan, L. (2019). Measuring work related physical and mental health function: Updating the work disability functional assessment battery (WD-FAB) using item response theory. *Journal of Occupational and Environmental Medicine, 61*(3), 219-224. https://doi.org/10.1097/JOM.0000000000001521

Masters, G. G. K., J. P. (1999). *Advances in measurement in educational research and assessment*. New York: Pergamon Press.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological bulletin, 114*(3), 449. https://doi.org/10.1037/0033-2909.114.3.449

Menendez, M. E., Bot, A. G., Hageman, M. G., Neuhaus, V., Mudgal, C. S., & Ring, D. (2013). Computerized adaptive testing of psychological factors: Relation to upper-extremity disability. *JBJS, 95*(20), e149. https://doi.org/10.2106/JBJS.L.01614

Muchlisin, M., Mardapi, D., & Setiawati, F. A. (2019). An analysis of Javanese language test characteristic using the Rasch model in R program. *Research and Evaluation in Education, 5*(1), 61-74. https://doi.org/10.21831/reid.v5i1.23773

Mujiyati, N., Warto, W., & Sutimin, L. A. (2019). Developing a problem-based local history module to improve the critical thinking ability of senior high school students. *Research and Evaluation in Education, 5*(1), 30-40. https://doi.org/10.21831/reid.v5i1.13334

Mujtaba, D. F., & Mahapatra, N. R. (2021). *Multi-objective optimization of item selection in computerized adaptive testing*. Paper presented at the GECCO 2021 - Proceedings of the 2021 Genetic and Evolutionary Computation Conference, 1018–1026. https://doi.org/10.1145/3449639.3459334.

Nisa, C., & Retnawati, H. (2018). Comparing the methods of vertical equating for the math learning achievement tests for junior high school students. *Research and Evaluation in Education, 4*(2), 164-174. https://doi.org/10.21831/reid.v4i2.19291

Nufus, A. H., & Wahyuni, L. (2024). Correlation of language skills on UTBK subtest with students' productivity skills. *World Journal of Advanced Research and Reviews, 21*(3), 2331-2342. https://doi.org/10.30574/wjarr.2024.21.3.0974

Oladele, J. I., Ndlovu, M., & Spangenberg, E. D. (2022). Simulated computer adaptive testing method choices for ability estimation with empirical evidence. *International Journal of Evaluation and Research in Education, 11*(3), 1392-1399. https://doi.org/10.11591/ijere.v11i3.21986

Otaya, L. G., Kartowagiran, B., Retnawati, H., & Mustakim, S. S. (2020). Estimating the ability of pre-service and in-service Teacher Profession Education (TPE) participants using Item Response Theory. *Research and Evaluation in Education, 6*(2), 160-173. https://doi.org/10.21831/reid.v6i2.36043

Özyurt, H., Özyurt, Ö., Baki, A., & Güven, B. (2012). Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. *Expert Systems with Applications, 39*(10), 9837-9847. https://doi.org/10.1016/j.eswa.2012.02.077

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *Research and Evaluation in Education, 9*(1), 86-117. https://doi.org/10.21831/reid.v9i1.63230

Pramudita, K., Rosnawati, R., & Mam, S. (2019). Methods used by mathematics teachers in developing parallel multiple-choice test items in school. *Research and Evaluation in Education, 5*(1), 10-20. https://doi.org/10.21831/reid.v5i1.22219

Pressman, R. S. (2012). *Software engineering a practitioner's approach* (7th ed.). Yogyakarta: Andi.

Quinlan, T. H. (2008). College of education college of education. *Educational Psychology, 5650*(813), 1–12.

Reckase, M. D., Ju, U., & Kim, S. (2018). *Some measures of the amount of adaptation for computerized adaptive tests. Dalam G. J., J. R., W. M., M. D., & C. S. (Ed.)*. Paper presented at the Springer Proceedings in Mathematics and Statistics (Vol. 233, hlm. 25–40). Springer New York LLC. https://doi.org/10.1007/978-3-319-77249-3_3.

Retnawati, H. (2016). *Quantitative analysis of research instruments*. Yogyakarta: Parama Publishing.

Rukli, R., & Atan, N. A. (2024). Simulation of low-high method in adaptive testing. *Research and Evaluation in Education, 10*(1), 35–49. https://doi.org/10.21831/reid.v10i1.66922

Samsudin, M., Chut, T. S., & Ismail, M. (2019). Evaluating computerized adaptive testing efficiency in measuring studentsâ€™ performance in science TIMSS. *Jurnal pendidikan ipa indonesia, 8*(4), 547-560.

Santoso, A. (2010). Development of computerized adaptive testing to measure learning outcomes of Open University students. *Jurnal Penelitian dan Evaluasi Pendidikan, 14*(1), 62–83. https://doi.org/10.21831/pep.v14i1.1976

Şenel, S., & Kutlu, Ö. (2018). Computerized adaptive testing design for students with visual impairment. *Egitim ve Bilim, 43*(194), 261–284. https://doi.org/10.15390/EB.2018.7515

Setiawan, A., Kassymova, G. K., Mbazumutima, V., & Agustyani, A. R. D. (2024). Differential Item Functioning of the region-based national examinationequipment. *Research and Evaluation in Education, 10*(1), 113. https://doi.org/10.21831/reid.v10i1.73270

Siwi, E. J., Anindyarini, R., & Nahar, S. (2020). Item parameters of Yureka education center (YEC) English proficiency online test (EPOT) instrument. *Research and Evaluation in Education, 6*(1), 51-65. https://doi.org/10.21831/reid.v6i1.31013

Stepanek, L., & Martinkova, P. (2020). *Feasibility of computerized adaptive testing evaluated by Monte-Carlo and post-hoc simulations*. Paper presented at the Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, 21, 359–367. https://doi.org/10.15439/2020F197.

Suhardi, I. (2020). Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm. *Research and Evaluation in Education, 6*(1), 32-40. https://doi.org/10.21831/reid.v6i1.30508

Sumin, S., Sukmawati, F., & Nurdin, N. (2022). Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression. *Research and Evaluation in Education, 8*(1), 55-66. https://doi.org/10.21831/reid.v8i1.50809

Tan, Q., Cai, Y., Li, Q., Zhang, Y., & Tu, D. (2018). Development and validation of an item bank for depression screening in the Chinese population using computer adaptive testing: A simulation study. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01225

Tang, X., Schalet, B. D., Hung, M., Brodke, D. S., Saltzman, C. L., & Cella, D. (2021). Linking oswestry disability index to the PROMIS pain interference CAT with equipercentile methods. *Spine Journal, 21*(7), 1185–1192. https://doi.org/10.1016/j.spinee.2021.02.012

Thompson, N., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1. https://doi.org/10.7275/wqzt-9427

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). CAT-MD: Computerized adaptive testing on mobile devices. *International Journal of Web-Based Learning and Teaching Technologies, 3*(1), 13–20. https://doi.org/10.4018/jwltt.2008010102

Turban, E., Aronson, J. E., & Liang, T.-P. (2005). *Decision support systems and intelligent systems* (7th ed.). Noida: Pearson/Prentice Hall.

Ulwatunnisa, M., Retnawati, H., Muhardis, M., & Yusron, E. (2024). Revealing the characteristics of Indonesian language test used in the national-standardized school examinations. *REiD (Research and Evaluation in Education), 9*(2), 210–222. https://doi.org/10.21831/reid.v9i2.31999

Van Der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*(3), 283–302. https://doi.org/10.1111/j.1745-3984.2005.00015.x

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. Dalam Computerized Adaptive Testing: Theory and Practice. In (pp. 1–25): Springer Netherlands. https://doi.org/10.1007/0-306-47531-6_1.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). New York: Routledge.

Walker, J., Böhnke, J. R., Cerny, T., & Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing-A practical method based on a systematic review. *Critical Reviews in Oncology/Hematology, 73*(1), 47–67. https://doi.org/10.1016/j.critrevonc.2009.03.007

Wardani, R. E. A., Prihatni, Y., Negeri, S., & Jl Jogja-Solo Km, K. (2018). Developing assessment model for bandel attitudes based on the teachings of Ki Hadjar Dewantara. *REiD (Research and Evaluation in Education), 4*(2), 117–125.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492.

Wilson, D. M., & Narasuman, S. (2020). Investigating teachers' implementation and strategies on higher order thinking skills in school based assessment instruments. *Asian Journal of University Education, 16*(1), 70–84. https://doi.org/10.24191/ajue.v16i1.8991

Winarno. (2012). The results of software testing for computerized adaptive testing (CAT) in Islamic religious education (PAI) for creating learning medical records (RMP) to diagnose students' abilities in schools. *Jurnal Penelitian Sosial Keagamaan, 6*(2), 183–204.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*(1), 135-155. https://doi.org/10.1111/emip.12094

Yim, L. W. K., Lye, C. Y., & Koh, P. W. (2024). A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis. *REiD (Research and Evaluation in Education), 10*(1), 18–34. https://doi.org/10.21831/reid.v10i1.65284

Yuan, Y., Xia, H., Han, Y., & Hu, M. (2020). *Advances in computerized adaptive testing*. Paper presented at the Proceedings - 2020 International Conference on Intelligent Computing and Human-Computer Interaction, ICHCI 2020, 202–205.

Yulianto, A. (2008). *Computerized adaptive testing: The psychological utilization of computers for psychological testing*. Paper presented at the Proceedings of the National Information Systems Conference, 7–12.