MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2025, 21(1), em2571
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/15897

The impact of computerized adaptive test termination rules on accuracy across different ability estimation methods

Lana Dwahdh [1], Nedal Alshraifin [1*]

[1] Department of Educational Psychology and Counseling, Yarmouk University, Irbid, JORDAN

Received 27 September 2024 ▪ Accepted 14 January 2025

Abstract

The study aimed to compare the termination rules of computerized adaptive testing (CAT), specifically the rule of termination after a fixed number of items versus the rule of termination based on the minimum standard error (SE), using the methods of maximum likelihood estimation (MLE) and maximum a posteriori. The goal was to assess the relative accuracy of each rule to determine which method provides the highest measurement accuracy. In order to address the objectives of the study, the researcher developed a mathematics item bank for the 6th and 7th grades consisting of 275 items. This bank was used to develop 6 achievement tests, three for 6th grade and three for 7th grade, with each test comprising 46 items. In addition, 10 items were common to all the tests and were used as a common core. The tests were conducted with 2612 students of class six and seven. For data and information analysis and processing, BILOG-MG-3.0, SPSS, and Fast Test Web v3.80.26 applications were used. Four different applications were performed on another sample of 403 students in order to evaluate the precision of the CAT procedure through several terminating criteria–a fixed criterion of 25 administered items and a criterion based on a SE of no larger than 0.25. From the results obtained, the accuracy of examinations is independent of the method used in estimating the parameters and that the determination of the fixed period is superior to the determination relating to SE. Moreover, the results revealed that adopting Bayes' theorem and a termination rule determined by standard deviation improves precision of the estimation, though in a case where 25 items are used, MLE is the best.

Keywords: computerized adaptive testing, item bank, item response theory, cat termination rules, ability estimation accuracy

INTRODUCTION

At any level in the education system, assessment and evaluation is critical for the educational program. Without the fundamental skills in measurement tools, a teacher cannot adequately fulfill the role of an evaluator. In this way, the activity of educational evaluation is significant at every point of the education continuum and it is crucial in the process of assuring quality and advancing the learning capabilities of the students. In order to do this, it is necessary to carry out a radical overhaul of the evaluation system to address the problems and dangers of classical skepticism, while maximizing the use of computer technologies to provide trustworthy and ongoing evaluation (Zio, 2018).

Psychological assessment as well as educational measurement has been pervasive with respect to devising constructs that can accurately test individuals and evaluate mental and performance assessment. Their relevance falls in the area to unmask the distinctive and limited skills of a person. In addition, item response theory (IRT) and the production of specialized statistical tools has fostered new developments in computerized adaptive testing (CAT) in comparison to the traditional paper-and-pencil and linear computer tests that ask each candidate the same set of questions irrespective of how good or bad they are performing. In contrast, computerized adaptive tests ensure that the questions asked to relate to the ability of the responder and hence give more precise and reliable predictions. Compared to

✉ lanamohamaddwahdh@gmail.com ✉ nshraifin@yahoo.com **(*Correspondence)**

Contribution to the literature

- This study provides a comparative analysis of fixed-item termination and minimum standard error-based termination in Computerized Adaptive Testing (CAT), offering empirical insights into their impact on measurement accuracy.
- The findings highlight that Bayesian Estimation (MAP) improves precision when paired with a standard error-based termination rule, whereas Maximum Likelihood Estimation (MLE) performs better with fixed-item termination, contributing to the optimization of ability estimation methods.
- By demonstrating that measurement accuracy is more influenced by termination rules than estimation methods, this study enhances the efficiency and reliability of CAT applications in educational assessments.

the paper and pencil test, an adaptive test has shorter time duration and less preparation time required.

Burhanettin and Selahattin (2022) categorize adaptive testing as the process of carrying out assessments whereby test items are constructed about an examinee's objective to answer the items that have been selected from a pool of questions, provided that the examinee has previously answered all the other items. A CAT too. Tests that rely on score computation theory, IRT, enable the comparison among candidates who possess different aptitudes and take the test at different occasions. Luo et al. (2020) state it as a transformation of psychological and educational tests to be used effectively on a computer. The procedure starts with choosing the first item from the bank of items according to certain criteria. When the first item has been given, the candidate's ability is calculated the physics using one of the endorsed methods of scoring which is based on IRT multi models and according to the pre estimated ability of a candidate to a level a new item is picked from item bank.

There are different strategies used in adaptive testing, with the computer-based adaptive testing strategy standing out. This makes use of ability based question-selection algorithms that operate through a computer. A prerequisite for a CAT test is the availability of an item bank that has been constructed and in which the item parameters have been estimated sufficiently closely so that they can be treated as fixed values. These parameters are when the test is conducted on the candidate are used to find out the amount of information contained in each test item and the candidate's ability levels. In the beginning the medium level of the question is administered and thereafter, the computer adapts to the sequential level of the responses from the examinee. The item response is re-estimated after every administration. In the end, the assessment ceases when a stopping rule is reached, for instance, after a specified number of items have been administered or set measurement precision has been achieved (Veldkamp & Verschoor, 2019).

The CAT is implemented through the following steps: All examinees begin with an ability level (θ) of 0, which ensures that the starting point of the adaptive test (CAT) is equal for all conditions, eliminating any chance differences that could influence the study's results. Then, one or several items are selected and presented to the examinee as an initial step. Most studies use the Fisher maximum information criterion for item selection, which increases the efficiency of the CAT by minimizing the standard error of ability (SE [θ]) as quickly as possible (AlAli & Wardat, 2024).

Estimation Process

Items are presented to the examinee and administered repeatedly by the computer. After each answer, the examinee's ability level is estimated. There are several methods for estimating the examinees' ability level such as maximum likelihood estimation (MLE), expected a posteriori (EAP), Bayesian model, MAP, and weighted likelihood estimator. The next item is presented based on the previously estimated ability level. Then the test termination criteria are determined, which indicates the end of the test, followed by the preparation of the final report of the ability level and the report generation options (Jatobá et al., 2020).

There are four criteria for terminating the adaptive test: test length, accuracy, classification, and information. The stopping rule is generally based on whether the attribute level has reached a predetermined precision. Item information is generally used to measure this precision. In the CAT, stopping rules are either fixed or variable length. Using a fixed length stopping rule will end the test once all respondents have received the same number of items, meaning that all subjects have finished the test when the test length has reached the predetermined length. Variable length stopping rules attempt to mitigate measuring subjects with varying degrees of precision by rotating items until a predetermined precision level is met. This means that some examinees will be given fewer items while other examinees will be given more items to achieve similar measurement precision. A common precision criterion used in variable length adaptive tests is the standard error of measurement (SEM), called the SE stopping rule. This stopping rule will continue to administer items until the SE associated with the provisional trait estimate is less than the pre-specified SE, resulting in trait levels estimated with a similar SEM (i.e., precision) for all examinees and thus the test ends (Audrey, 2019).

Stafford et al. (2019) outlined several criteria that can be used to stop administering adaptive tests, including the fixed length criterion, where each examinee is given the same number of test items as determined by the examiner. However, this lacks consistency in the accuracy of ability estimates for examinees, and the accuracy of estimation may vary depending on the examinee's position on the ability continuum. Attempts to improve accuracy using this method require increasing the number of test items, which undermines the purpose of test adaptation. Additionally, this increases the exposure rate of each item in the question bank, which must be considered when adapting tests (Leroux & Dodd, 2014). The variable length criterion depends on reaching a certain level of estimation accuracy based on the examinee's ability, as determined by the SEM. Under this criterion, not all examinees are exposed to the same number of items but rather the number needed to achieve the predetermined accuracy level (Alneyadi et al., 2023).

The most common criterion is the variable length of the test and is based either on the minimum SE rule, where this method depends on designing the test so that it ends when it reaches a certain SE level or an equivalent level of accuracy, for example the test ends when the SE becomes 0.25 or less. This criterion is characterized by providing accurate equivalent scores for all students, assuming that the question bank is constructed correctly. The second rule is the minimum information (MI) rule, where the test stops when there are no available items capable of providing a pre-determined minimum of information for the tester's ability. This rule is characterized by the fact that the tester is not exposed to unnecessary items that achieve high measurement accuracy, while the disadvantage of this method is that it provides less measurement accuracy than the SE rule because it does not provide additional items when the available information is insufficient (Weiss & Şahin, 2024).

Ayanwale and Ndlovu (2024) noted that CAT can reduce the number of items administered by 50% compared to linear tests, while Ebenbeck et al. (2024) explained that the length of the test can be shortened by 50-90%. Therefore, it increases students' motivation, contributes to reducing their test anxiety, lack of concentration, and boredom (Ayanwale & Ndlovu, 2024; Ebenbeck et al., 2024).

A test cannot be terminated if too few items have been administered to the examinee, as they may not feel their ability has been sufficiently assessed unless they answer between 10-20 items. Additionally, the test cannot end unless it is confirmed that all topics measured by the test have been covered. To accurately assess an individual's ability, all topics the test aims to measure must be presented to them (Tian et al., 2020).

A key component of CAT is the stop rule, which determines when the vertebrae can no longer be administered. CAT can be either a fixed-length test or a variable-length test. Fixed-length tests end once a predetermined number of passages have been administered, while this may be considered a very simple way to turn off adaptive testing, it does not measure the latent traits of examinees with the same accuracy. In contrast, variable length stopping rules attempt to provide equal measurement accuracy across attribute levels. By administering items until a predetermined level of measurement accuracy is achieved, computerized adaptive tests provide more reliable estimates of ability using fewer items compared to traditional tests (Bao et al., 2021). CAT is one of the greatest reflections of developments in ICT in education and contributes to more efficient and effective assessments. Unlike traditional paper-and-pencil tests, adaptive testing uses different test formats simultaneously based on the performance of individuals with different levels of ability (Bao et al., 2021).

The IRT is a theoretical framework for psychological and educational measurement, used for item analysis and building adaptive tests. One of the reasons for the emergence of this theory is the weaknesses found in classical test theory. To address these weaknesses, psychometric research has focused on developing a theory that can achieve the following: item statistics that do not depend on the group of examinees, ability scores that are independent of test difficulty, and a foundation for linking test items to the ability levels of individuals (Stoeckel et al., 2021).

The IRT differs from the classical theory in the assumptions on which it is based, in order to reach reliable results, the most prominent of which are, as follows:

1. The assumption that an individual possesses one or more traits underlying his or her responses to test items, which are used to explain these responses. These traits cannot be directly observed, so they are referred to as latent traits.
2. Another assumption is statistical independence, which means that the responses of two examinees to test items are statistically independent at a given ability level. In other words, examinees' response to one item should not be as positive or negative as their response to another item (Bichi & Talib, 2018).

The literature highlights several advantages of computerized adaptive tests over traditional tests, the most important of which are Weiss and Şahin (2024), Wulandari et al. (2020), and Mujtaba and Mahapatra (2020).

Immediate scoring and storage of results, increased measurement accuracy, and allowing the individual to respond at their own pace according to their ability. The

individual's response to one item determines the difficulty of the next item, so the individual immediately knows whether their response was correct or incorrect. The use of computers in adaptive testing also offers various opportunities to diversify item formats, such as graphical representations and animations. Computers can be used to increase the accuracy of test scores through CAT, where the individual's ability estimate is updated after each new response, and the next item is selected to optimize test characteristics.

To develop a computerized adaptive test, five components or steps need to be carried out. The first component is an item pool, which is designed according to the subject of the test (for instance, math ID's for a math test). The other four components, used in the construction of a CAT system, are non-content based. These include permanent item pools, the first use of measurement, a humorous use of measurement devices, and the census as such (Wang et al., 2020).

To be able to carry out computerized adaptive tests, use of an item bank is essential, It consists of an organized and classified database of test items in the same manner as books in a library. This arrangement is based upon the nature of the test item as well as its psychometric characteristics such as difficulty, reliability, validity etc. To achieve this, when proposing to set up a teacher evaluation bank, a specification matrix and evaluation map must be created. The table of test specifications indicates the power of a test and the appropriate question format to be employed (Weiss & Şahin, 2024). Here, how the previous item answer relates to the subsequent one is essential.

1. Maximum information technique and Bayesian technique are by far the most popular and widely used techniques in item selection (Ayanwale & Ndlovu, 2024).
2. The first and foremost step of a computerized adaptive test lies in estimating the tester's ability. The estimate of the ability is made on the basis of the response of the examinee to the items which have parameters that are known according to the IRT model that has been used to calibrate the data. There are two methods for estimating ability (Hambleton et al., 1991).
3. MLE: An examiners' ability is predicted by this technique through the likelihood estimation function. This is the method which is widely used for estimating ability. The method or technique uses the examinee response to the items and predicts the level of ability which maximizes the likelihood function of the examinee's response pattern. The equation below shows the likelihood function for the response pattern. L = One of the major disadvantages of this method is the inability to estimate the ability in severe boundaries such as where the respondent has addressed all the items either correctly or incorrectly. In these instances the ability estimate can be 4 or -4 (Hori et al., 2021).

$$L(UI\theta)\prod_{t=1}^{n} P_i(\theta_s^{ui} Q_i \theta^{1-ui}), \quad (1)$$

where u_i is 1 for a correct response and it is 0 for an incorrect response.

4. Bayesian methods: These methods are typically used when an examinee answers all items correctly or incorrectly, making the MLE method unsuitable for extreme ability levels. In this approach, if prior information about the distribution of abilities is available, the ability estimate becomes more meaningful. Bayesian estimation methods include the EAP and the MAP methods. This approach is known for its accuracy in estimating ability and can estimate ability for all examinees, even at extreme ability levels (Rios, 2022).

In adaptive testing, a key step is determining the test termination rule, which typically ends either after a specified number of items or when a predefined minimal SE is reached. The termination rule is checked after each ability estimate to ensure whether the test should end or if a new item should be presented (Wainer, 2000).

1. Fixed length testing: In this scenario, each examinee receives the same number of items that match their ability. One of the advantages of fixed-length testing is that it simplifies the decision of whether to present a new item to the examinee by counting the items already administered. In contrast, in a variable-length test, examinees who receive fewer items may feel that they were not adequately assessed. However, a disadvantage of fixed-length testing is that it does not provide the same level of measurement accuracy at all points along the ability distribution. The ability estimates for examinees at the extremes of the distribution are less accurate than for those in the middle of the distribution (Stemler & Naples, 2021). The fixed-length criterion sets a maximum number of items that must be administered, and the CAT stops once this limit is reached. While longer tests generally increase the accuracy of ability estimates, shorter tests may be considered to address certain issues early in the CAT process, such as the impact of early errors on item selection (Magis & Raiche, 2012).
2. Variable length testing: The test ends when a certain level of measurement accuracy for ability is achieved, typically based on the SEM for the estimated ability. After each item is presented to the examinee, the SE is calculated, and items continue to be administered until a predetermined SE level is reached. An advantage of this criterion is that the ability of all examinees is estimated

with the same level of accuracy (Ebenbeck & Gebhardt, 2022).

Lee and Kim (2020) pointed out that variable-length CATs, which stop when a certain SE is reached, reduce the number of items administered by approximately 91-93%, while fixed-length tests that stop after administering a set number of items reduce the number by around 89-91%.

Estimation-based CATs require a predefined stopping rule to determine when the test should end. Among the various stopping rules available, the precision criterion is frequently used (Magis et al., 2017, p. 47-48). This criterion terminates the test once the ability estimate reaches a predetermined level of accuracy, defined by an SE that must be equal to or less than the target SE. The precision criterion, for example, with a SE between 0.3 and 0.5, can achieve a good balance between test accuracy and test length in a CAT, as can be seen in the studies by Hol et al. (2008), Stafford et al. (2019) and Ebenbeck et al. (2024).

Research Problem

The concern that warrants an exploration into the success of adaptive test termination methods stems from the recommendations in multiple research that encourage the broader application of adaptive testing within evaluation and measurement systems. Moreover, there is also a question towards retaining the validity of measurement when applying either fixed or variable termination rules. As in the case of Choi et al. (2011), the recommendation saw, "… a modification with the aim of presenting a new termination rule is implemented during a computerized adaptive test and hence minimizes the expected SE of the posterior predictive variance…" This appendix took the position that future studies on shorter CAT applications should center on termination rules for the control of item presentation during ability estimation. In addition, Yildiz et al. (2024) proposed using the "minimum SE" termination rule and changing its requirements assessment procedures to report limits for judging the CAT effectiveness. He also suggests using CAT in the "fixed number of items" mode with varying item counts that are dependent on the method of ability assessment.

In the previous literature regarding CAT, there is a discrepancy regarding the appropriate ability estimation method and stopping rule. It follows from the studies conducted by Leroux and Dodd (2014), Babcok and Weiss (2012), and Stafford et al. (2019) that relying on the SEM criteria, a variable length rule gives a more accurate picture. The trimmed rule, on the other hand, does measure the latent traits of the examinees, but not with the same level of accuracy, introducing a variation in the estimation of abilities across the range of capabilities. Such a problem exists for high stake practical examinations like certification or licensing examinations where the test ends up being less accurate, an over dependence on CAT leads to a larger item exposure and a larger burden on the examinees.

Choi et al. (2011) pointed out that although the fixed-length rule is simple, examinees are measured with varying degrees of accuracy, resulting in larger measurement errors at extreme ability levels. Moreover, the fixed-length rule limits the efficiency of CAT when unnecessary items, which provide little information about the examinee's ability level, are administered. Kalender and Berberoglu (2017) found no significant differences between the fixed-length and SE termination rules in CAT and suggested its use for university admissions in Turkey.

Although the idea of CAT is theoretically simple, planning and constructing it are complex processes. Essential components must be considered, including the item bank, item selection method, ability estimation method, and test termination rule. Any flaw in one component can affect the others, resulting in unsatisfactory outcomes.

A mathematics item bank was developed and validated for item selection based on the three-parameter logistic model. Constructing an item bank involves compiling items, testing them, calculating their parameters, and storing them in specialized software that provides items with specific characteristics suited to the test's purpose and the group of examinees.

In light of the above, there are multiple methods for estimating the examinees' abilities in CAT and various ways to terminate the computerized adaptive test. This study contributes real data from commonly used tests that are relatively free from cultural bias. A mathematics ability estimation test was selected, and the data extracted from it were used in the first phase, then developed into a computerized adaptive test in the second phase by converting the traditional (paper-and-pencil) test into a CAT. The study examined the effectiveness of CAT using the minimum SE termination rule with two ability estimation methods: the MLE and the MAP. It also investigated the effectiveness of CAT using the fixed-length termination rule with both MLE and MAP methods.

This CAT study fills in the gaps existing in the current research by providing ability estimation and test termination approach insights and is on par with the latest during this timeframe. Unlike many prior works which use simulated data, this research incorporates data from culturally fair mathematics ability assessments which adds to its realism. The work is also stimulating as it changes a conventional math paper test into a CAT test format, which is pioneer in the field of math education. MLE and MAP methods are widely used for ability estimation and this study compares the efficacy of these two methods with respect to two termination rules: fixed length and SE minimum. This

dual-layered technique ensures that any comparison which is proposed is assessed to the maximum possible extent. In this way, the research makes contribution to the theory life while looking for ways to improve offline CAT design and implementation within various educational practices.

Given this framework, the problem can be summarized in answering the following research questions regarding the effectiveness of CAT across different conditions and experimental designs, varying in termination rule and ability estimation method:

1. Does the accuracy of CAT in mathematics for students differ based on the ability estimation method used?
2. Does the accuracy of CAT in mathematics for students differ based on the test termination rule?
3. Does the accuracy of CAT in mathematics for students differ based on the interaction between the ability estimation method and the test termination rule?

Research Objectives

This study aims to compare the termination rules of CAT–the fixed number of items rule and the minimum SE rule–using two methods of ability estimation: MLE and MAP. It seeks to evaluate the relative accuracy (effectiveness) of CAT termination rules in measuring ability to determine which methods provide the most accurate measurement under a specific item pool.

Research Importance

The importance of this study lies in two aspects:

1. Theoretical importance: Given the widespread use of CATs to measure attitudes and abilities, numerous studies have examined the accuracy of CAT through comparisons of different item selection methods. Therefore, it is crucial to study the measurement accuracy of CAT using different termination rules (fixed number of items and minimum SE) with the three-parameter logistic model under IRT, employing the MAP and MLE methods, as in this study. The significance also stems from comparing different estimation methods using CAT termination rules, which has not been extensively addressed in Arab studies. This makes the study valuable in helping researchers build adaptive computerized tests under various research conditions for practical applications, which is an emerging area in CAT.
2. Practical importance: Most studies on item banks and adaptive computerized tests are scarce. This study helps guide researchers in building CATs and assists in their practical application. It can also provide insights to researchers focusing on constructing and developing educational and psychological tests and measures, offering guidance on the best method for test termination when estimating ability in CAT. Moreover, the study's results can serve educational test administration, particularly in the Ministry of Education and educational measurement and evaluation centers within government and private institutions, which analyze educational and psychological test data.

Conceptual and Operational Definitions

1. CAT: A precise test that maximizes time and effort by presenting items tailored to the examinee's ability level. Operationally, in this study, it refers to a mathematics achievement test prepared by the researcher for sixth and seventh-grade students.
2. Effectiveness (relative accuracy): The ability to achieve accurate measurements, as demonstrated by several statistical indicators: the average number of items administered, the test's information function, the SE of ability estimation, the SEM, the root mean square error, and bias in estimated ability. Operationally, it is defined in this study by assessing confidence through the correlation between the estimated and true ability of the examinee, determining the most efficient method, i.e., the one that uses the fewest items with the least SE in estimating ability.
3. MLE: A method that estimates ability based on the examinee's response pattern (1 or 0 for each item). It is one of the ability estimation methods in IRT and derives parameter estimates through likelihood maximization. Operationally, it refers to the estimation of ability values for each examinee using MLE based on responses to mathematics test items, which follow the three-parameter logistic model.
4. MAP estimation: A method that relies on prior information about the distribution of ability and assumes the distribution follows a known form, typically the standard normal distribution. Operationally, it refers to the estimation of ability values using MAP for mathematics test items.
5. Minimum SE: The method where ability estimation continues until the SE reaches a pre-specified value, at which point the test is terminated. It is a rule for CAT termination. Operationally, in this study, the minimum SE is set to 0.25.

Research Limitations

This study was limited to:

1. Sixth and seventh-grade students in public and private schools under the directorates of

education in Irbid Governorate for the academic year 2023/2024.

2. The use of the study tool, which consisted of achievement tests in mathematics for sixth and seventh graders in Jordan.
3. The use of two methods for terminating the CAT: the first involves administering a predefined number of items (47) to all examinees, and the second continues administering items until the examinee's ability estimation reaches the minimum SE, which has been reported by some researchers to be 0.25.

PREVIOUS STUDIES

This section reviews previous studies related to the effectiveness of adaptive testing and CAT termination methods. It includes both international and Arab studies that directly or indirectly address the topic, presented chronologically from oldest to most recent.

van der Linden and Glas (2010) sought to improve CAT by using item cloning techniques. They proposed using a multilevel IRT model to increase the number of items and reduce the cost of item writing. They employed marginal likelihood and Bayesian methods for parameter estimation and developed an item selection method that involved choosing an optimal set of cloned items, then selecting an item randomly from the set. Simulations with law school admission test data demonstrated the accuracy of this method in item calibration and the effectiveness of adaptive testing.

Žitný (2011) conducted a study to examine the accuracy, validity, and efficiency of CAT by reviewing the results of 15 research studies in the areas of ability testing, clinical psychology, personality testing, and healthcare. The findings showed that CAT is effective in efficiently providing the required information, reducing both time and the number of necessary items. The study also provided evidence for the reliability and validity of adaptive tools, based on simulation studies. It recommended further direct research to confirm CAT results with actual examinees.

Saleh et al. (2023) conducted a study that tested the effectiveness of CAT in terms of measurement accuracy and its psychometric properties. The researcher used 48 multiple-choice questions and administered the test to students in two forms: a paper-based test and a computerized adaptive test. The results showed that CAT was more effective than the linear test, requiring fewer items (20 questions) to achieve higher accuracy. The study also found that the MAP estimation method was more accurate than the MLE method. CAT provided 20% higher measurement accuracy than the linear test and reduced the number of items by over 50%.

Stafford et al. (2019) aimed to compare CAT termination rules using the generalized partial credit model (GPCM). One key consideration for any CAT program is the criterion used to stop item administration (termination rule), ensuring that all examinees are evaluated under the same standard. This study compared the performance of three variable-length termination rules: SE, MI, and change in theta (CT) (used either separately or in combination with minimum and maximum item count requirements), along with the fixed-length termination rule. The results indicated that the MI criterion produced biased estimates and showed considerable variability in measurement quality across the data distribution. The CT rule performed strongly when paired with a lower bound and minimum test length, while the SE rule consistently provided the best balance between measurement accuracy and operational efficiency. It required the least number of items to obtain accurate theta (θ) estimates, especially when paired with the maximum item count termination rule.

Ebenbeck and Gebhardt (2022) aimed to develop a computerized adaptive test for students with special needs, with the aim of reducing test time and improving its accuracy. The researchers used question banks for mathematics and reading comprehension, The math question bank consists of 80 items created for students ages 8 to 12 who have basic arithmetic skills, while the reading comprehension question bank consists of 219 items and relied on the Rasch model to achieve a balance between test length and accuracy. The results showed that using question banks and the Rasch model can help reduce test time and improve test accuracy.

Janpla and Piriyasurawong (2023) aimed to build an item bank of multiple-choice questions in geography for use in constructing two tests: one computerized linear test and one computerized adaptive test. The researcher developed 375 multiple-choice items and distributed them across two models (A and B), with 54 items in each model, sharing 10 common items. The models were applied to two different samples of examinees, and the items were stored in a database using Microsoft Access 2000. The item bank was then used to generate both a computerized linear and adaptive test for eighth-grade geography.

Alkan and Deniz (2023) aimed to develop a computerized adaptive test model for an occupational interest inventory that was initially created in paper-and-pencil format. The paper-and-pencil version of the occupational field interest inventory (OFII) was administered to 1,425 high school students, and subsequent simulations were conducted using the collected data. According to the simulation results, the most optimal criteria for CAT implementation were the GPCM under IRT, a SE value of 0.40 as the test termination rule, and maximum Fisher information (MFI) as the item selection method. The OFII concluded with an average of 59 items, and the correlations between the paper-and-pencil scores and the estimated θ from the simulation ranged from 0.91 to 0.97.

Following the simulation results, the CAT was administered to 150 students, and the correlations between the students' online test scores and their estimated θ levels from the CAT ranged from 0.73 to 0.91.

Yildiz et al. (2024) conducted a study to examine the effectiveness of CAT in estimating cognitive ability using Raven's matrices tests. The study divided the test items into two sets, each containing 70 items, and applied them to 2,695 students. The results showed that the fixed number of items termination rule provided more accurate estimates and a higher information function compared to the minimum SE rule. Additionally, the adaptive test terminating at the minimum SE offered more accurate estimates and required 70% fewer items than the linear test, while also providing a higher information function.

Alkan and Deniz (2023) aimed to develop a computerized adaptive test model for an occupational interest inventory that was initially created in paper-and-pencil format. The paper-and-pencil version of the OFII was administered to 1,425 high school students, and subsequent simulations were conducted using the collected data. According to the simulation results, the most optimal criteria for CAT implementation were the GPCM under IRT, a SE value of 0.40 as the test termination rule, and MFI as the item selection method. The OFII concluded with an average of 59 items, and the correlations between the paper-and-pencil scores and the estimated θ from the simulation ranged from 0.91 to 0.97. Following the simulation results, the CAT was administered to 150 students, and the correlations between the students' online test scores and their estimated θ levels from the CAT ranged from 0.73 to 0.91.

In reviewing previous studies, it was observed that their objectives varied. Some aimed to investigate the effectiveness of adaptive tests, while others examined CAT effectiveness by proposing new methods for item selection, such as van der Linden and Glas (2010, Žitný (2011), Ebenbeck and Gebhardt (2022). Several studies focused on comparing computerized adaptive tests with linear tests,such as those by Janpla and Piriyasurawong (2023), Amin (2018), Alkan and Deniz (2023). Some studies aimed to compare CAT termination methods, while others focused on the impact of fixed and variable-length CAT designs on reducing test anxiety. Additionally, studies compared variable termination rules, such as those by Yildiz et al. (2024), Alkan and Deniz (2023), Stafford et al. (2019).

From the review of previous studies, it is evident–within the knowledge of the researchers–that no study has investigated the effectiveness of CAT termination methods according to the ability estimation method. Furthermore, none of the studies have compared ability estimation methods with CAT termination rules. This study, therefore, seeks to examine the effectiveness of CAT termination methods based on the ability estimation method by developing an adaptive test according to the estimation method and termination rule. It also aims to provide a theoretical framework for adaptive testing based on modern theory and to compare ability estimation methods according to CAT termination rules, as well as compare CAT termination rules based on ability estimation methods.

METHODOLOGY

This is a psychometric study that employs the descriptive survey method to collect data. Achievement tests in mathematics for sixth and seventh-grade students were designed based on IRT. The tests were analyzed in terms of item difficulty, discrimination, and guessing parameters, as well as the ability parameters of the individuals.

Population and Sample

The study population consisted of all sixth and seventh-grade students in public and private schools under the directorates of education in Irbid Governorate for the academic year 2023/2024, totaling 27,275 students, according to the 2023 statistics from the Ministry of Education.

To answer the research questions, two samples were selected:

1. The first sample included 2,612 sixth and seventh-grade students, used to build an item bank, calibrate the items, estimate their parameters, and match the model. This sample was chosen using a simple random sampling method, ensuring its distribution across variables like grade (sixth and seventh), gender (male and female), and school type (public and private).
2. The second sample consisted of 403 seventh and eighth-grade students, also selected through simple random sampling, distributed across the variables of grade (seventh and eighth), gender (male and female), and school type (public and private). This sample was used to apply the CATs.

Instrument

To achieve the study's objectives, the researchers developed a study tool comprising achievement tests in mathematics for sixth and seventh grades. A total of 275 items were prepared, distributed across six achievement tests–three for sixth grade and three for seventh grade–each containing 46 items.

Statistical Analysis

The following steps were taken to verify the uni-dimensionality assumption and select the appropriate logistic model for the data:

1. Verification of uni-dimensionality for achievement tests: SPSS was used to confirm the

Table 1. Number of primary factors in achievement tests and the ratio of the eigenvalue of the first factor to the eigenvalue of the second factor

Achievement tests	Number of factors	Eigenvalue first factor	Eigenvalue second factor	Ratio first to second factor
Sixth-1	8	13.613	2.533	5.374
Sixth-2	9	15.570	3.098	5.025
Sixth-3	5	20.976	1.978	10.604
Seventh-1	7	16.145	2.330	6.929
Seventh-2	6	15.186	2.463	6.165
Seventh-3	4	22.073	2.017	10.943

Table 2. Items excluded based on Chi-square criterion and biserial correlation

Criterion	Number of items not fitting the three-parameter model
Chi-square (χ^2)	25
Biserial correlation	0
Total	25

Table 3. Number of individuals whose data did not fit the three-parameter logistic model for each achievement test

Subtest (test versions)	Number of individuals not fitting the model
Sixth-1	30
Sixth-2	48
Sixth-3	18
Seventh-1	34
Seventh-2	27
Seventh-3	62

uni-dimensionality of the tests and to extract the factor structure using principal components analysis with orthogonal rotation (Varimax) on the data collected from the study sample through the administration of the achievement tests. High values for the sample adequacy index indicated that the data were suitable for analysis and for selecting the appropriate logistic model.

The principal components and their numbers for each test, as well as the ratio of the first factor's eigenvalue to the second factor's eigenvalue, were obtained, as shown in **Table 1**.

Table 1 shows that the ratio of the eigenvalue of the first factor to the second factor was greater than 2, indicating uni-dimensionality and the presence of a dominant factor. According to Hambleton and Swaminathan (1985), this dominant factor could be the mathematical ability factor.

Fitting the Three-Parameter Logistic Model to the Data

The researchers fitted the three-parameter logistic model to the items of the achievement tests. The data obtained from the application of the achievement tests were analyzed using the BILOG-MG-3.0 software, and the following criteria were applied:

1. Chi-square statistics: Used to determine the fit of the item to the model.
2. Item-total correlation (biserial correlation): To assess the correlation between the item and the total score.

This process ensures the appropriateness of the model for the data, evaluating how well each item fits the theoretical expectations of the three-parameter logistic model.

Table 2 shows that 25 items did not fit the three-parameter logistic model for the achievement tests based on the Chi-square criterion, while no items were excluded based on the biserial correlation criterion.

Table 3 shows that the number of individuals whose data did not fit the three-parameter model was **219**, indicating that they did not contribute to the information function. Since the data fit the three-parameter logistic model, the averages of the item parameters (difficulty, discrimination, and guessing) were calculated for the items of each test according to the three-parameter logistic model.

Table 4 shows the averages of item parameters (difficulty, discrimination, and guessing) for each test according to the three-parameter logistic model.

Scaling of Items Fitting the Model

1. Achievement tests were scaled on a common metric using the BILOG-MG-3.0 software.
2. Error in scaling was found to be low, with a value of 0.303.
3. Reliability of ability estimation was high, at 0.925, indicating a strong level of precision in ability estimation.

Table 4. Averages of item parameters (difficulty, discrimination, and guessing) for each test according to the three-parameter logistic model

Achievement tests	Difficulty parameter	Discrimination parameter	Guessing parameter
Sixth-1	-0.542	2.416	0.216
Sixth-2	-0.517	2.873	0.209
Sixth-3	-0.898	3.159	0.203
Seventh-1	0.145	3.653	0.224
Seventh-2	0.061	3.362	0.214
Seventh-3	-0.240	4.163	0.203

Table 5. Minimum, maximum, mean, and standard deviation of item parameters (difficulty, discrimination, and guessing) according to IRT

Parameter	Maximum value	Minimum Value	Mean	Standard Deviation
Difficulty	0.851	-1.589	-0.248	0.465
Discrimination	4.765	1.351	3.377	1.186
Guessing	0.459	0.105	0.212	0.050

Extracting Characteristics of the Achievement Test Items

1. The item parameters (difficulty, discrimination, and guessing) were estimated according to IRT after scaling the subtests.

Table 5 shows the minimum, maximum, mean, and standard deviation of item parameters (difficulty, discrimination, and guessing) according to IRT.

Item Storage and Preparation

1. Final test bank: The total number of items included in the mathematics test across its six versions was 250 items.
2. The items that fit the model were stored using the Fast Test Web v3.80.26 software, with items categorized according to different dimensions.
3. Once the test bank was finalized, the researchers proceeded with official procedures to apply for the CAT using computer labs in various schools. The items and their parameters were entered into the Fast Test Web v3.80.26 software, and test sessions were prepared accordingly.

Four Different Applications of the Adaptive Tests

1. First application:
   a. Each participant took an adaptive computerized test.
   b. The test ended after 25 items were administered.
   c. The test started at different ability levels for each examinee.
   d. Ability was estimated using MLE.
2. Second application:
   a. Each participant took an adaptive computerized test.
   b. The test ended when an SE of 0.25 or less was achieved.
   c. The test started at different ability levels for each examinee.
   d. Ability was estimated using MLE.
3. Third application:
   a. Each participant took an adaptive computerized test.
   b. The test ended after 25 items were administered.
   c. The test started at different ability levels for each examinee.
   d. Ability was estimated using MAP.
4. Fourth application:
   a. Each participant took an adaptive computerized test.
   b. The test ended when a SE of 0.25 or less was achieved.
   c. The test started at different ability levels for each examinee.
   d. Ability was estimated using MAP.,

These applications tested different stopping criteria and estimation methods, ensuring flexibility in the assessment process. Let me know if you'd like further details or clarification!

RESULTS

Results Related to the First Question

"Does the accuracy of the computerized adaptive mathematics test for students differ based on the ability estimation method?"

To answer this question, data from two applications of the test were used on a sample of 402 students:

1. First application: Ability was estimated using the MLE method with 196 students.
2. Second application: Ability was estimated using the MAP method with 206 students.

Table 6. MANOVA analysis to determine the effect of estimation method on test accuracy

Independent variable	Dependent variable	Sum of squares	Degrees of freedom	Mean squares	F-value	Significance level
Estimation method	Score	20.042	1	20.042	0.468	0.494
Standard error	0.151	1	0.151	3.751	0.053	
Error	Score	17,129.988	400	42.825		
Standard error	16.101	400	0.040			
Corrected total	Score	17,150.030	401			
Standard error	16.252	401				

Table 7. Results of MANOVA to determine the effect of termination method on test accuracy

Independent variable	Dependent variable	Sum of squares	Degrees of freedom	Mean squares	F-value	Significance level
Termination method	Score	13,074.903	1	13,074.903	1,283.386	0.000
Standard error	0.057	1	0.057	1.401	0.237	
Error	Score	4,075.127	400	10.188		
Standard error	16.195	400	0.040			
Corrected total	Score	17,150.030	401			
Standard error	16.252	401				

The accuracy of the test was evaluated using multivariate analysis of variance (MANOVA) to compare the accuracy between the two methods based on student scores and SE. **Table 6** presents the results of this analysis.

Table 6 indicate that the accuracy of the computerized adaptive test in mathematics is not significantly affected by the estimation method used. The ability estimates and SE values show similar results using both the MLE and MAP methods. This suggests that the computerized adaptive test is reliable in providing accurate ability estimates regardless of the estimation method.

The small difference in significance (with an F value of 3.751 and p = 0.053 for SE) indicates that while there is a minor variation, it is not statistically significant enough to favor one method over the other in terms of accuracy.

In conclusion, the results demonstrate the importance of the computerized adaptive test in producing precise estimates, regardless of whether MLE or MAP is used.

Results Related to the Second Question

"Does the accuracy of the computerized adaptive mathematics test for students differ based on the test termination rule?"

The study sample included 402 students from the seventh and eighth grades, divided into two groups:

1. First group: Used the fixed termination rule after 25 items (199 students).
2. Second group: Used the termination rule when the SE reached 0.25 or less (203 students).

A MANOVA was used to compare the accuracy and evaluation of the test between the two termination methods. The results are shown in **Table 7**.

Table 7 shows that the accuracy of the computerized adaptive test in mathematics differs based on the test termination method used, particularly with respect to the examinee's score. This is evident from the F value of 1283.386, which is statistically significant (p = 0.000). This means that test accuracy significantly varies depending on whether the fixed 25-item rule or the SE rule (0.25 or less) is applied.

On the other hand, the SE did not show a significant difference between the two termination methods (F = 1.401, p = 0.237), suggesting that the level of precision in terms of SE remains consistent regardless of the termination rule.

To further investigate the significance of these differences and to determine which termination method is more accurate, a t-test for two independent samples was conducted, as shown in **Table 8**.

Table 8 highlight significant differences between the two termination methods, with the fixed number of items method showing a statistically significant advantage (p = 0.00) in terms of examinee scores. The t value of 35.824 further emphasizes the superiority of the fixed termination method over the SE-based termination.

The researchers concluded that choosing an appropriate termination method is crucial for improving the accuracy of the test. Notably, the fixed-item termination method outperformed the SE method regarding test accuracy, which suggests that using a fixed number of items might provide more reliable results for adaptive computerized tests.

This indicates that, when designing adaptive tests, the termination rule plays a significant role in achieving precise and consistent estimates of ability.

Table 8. Results of the t-test for independent samples to identify differences between termination methods and their effect on examinee scores as an indicator of test accuracy

Dependent variable	Independent variable (termination method)	N	Mean	Standard deviation	T value	Degrees of freedom	Significance level	Favored method
Score	Fixed	199	16.387	3.318	35.824	400	0.00	Fixed
Standard error		203	4.980	3.064				

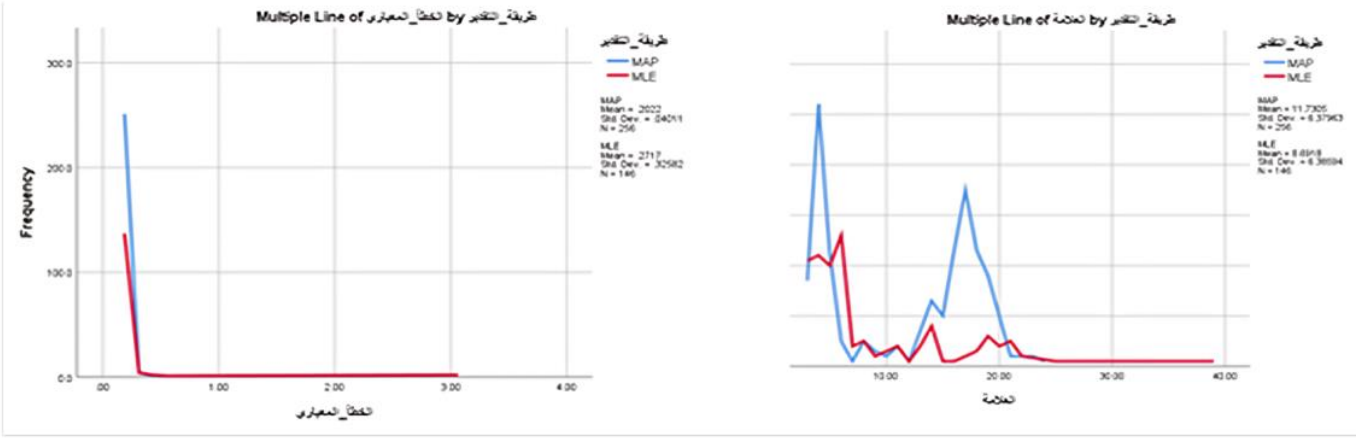

Figure 1. The effect of termination methods on test accuracy based on the score indicator and the standard error indicator (Source: Authors' own elaboration)

Table 9. Results of MANOVA to determine the effect of the interaction between estimation methods and termination rules on test accuracy

Independent variable	Dependent variable	Sum of squares	Degrees of freedom	Mean squares	F-value	Significance level
Estimation methods * termination rules	Score	13,285.944	3	4,428.648	456.150	0.000
Standard error	0.392	3	0.131	3.276	0.021	
Error	Score	3,864.086	398	9.709		
Standard error	15.860	398	0.040			
Corrected total	Score	17,150.030	401			
Standard error	16.252	401				

Figure 1 shows the effect of termination methods on test accuracy based on the score indicator and the SE indicator.

Results Related to the Third Question

"Does the accuracy of the computerized adaptive mathematics test for students differ based on the interaction between the ability estimation method and the test termination rule?"

The test was administered four times to a sample of 402 seventh- and eighth-grade students:

1. First application: Ability estimation using MLE with a fixed termination rule of 25 items, applied to 98 students.
2. Second application: Ability estimation using MLE with a termination rule at a SE level of ≤ 0.25, applied to 98 students.
3. Third application: Ability estimation using MAP with a fixed termination rule of 25 items, applied to 101 students.
4. Fourth application: Ability estimation using MAP with a termination rule at a SE level of ≤ 0.25, applied to 105 students.

The MANOVA was used to compare test accuracy based on the estimation methods and termination rules, as shown in **Table 9**.

Table 9 indicate that the interaction between ability estimation methods and test termination rules significantly affects the accuracy of the adaptive test through two indicators:

1. Examinee's score: The F value is 456.15, which is statistically significant with a p-value of 0.00.
2. SE: The F value is 3.276, also statistically significant with a p-value of 0.021.

These findings suggest that there are meaningful differences that warrant further investigation through post-hoc comparisons using the least significant difference method, as shown in **Table 10**.

Table 10. Post-hoc comparisons

Dependent variable	Comparison	Mean difference	SE	Sig.	Favored method
Score	Score	Fixed * MAP	-1.9084*	0.44181	0.000
	SE * MAP	10.8088*	0.43765	0.00000	Fixed * MLE
	SE * MLE vs. Fixed * MAP	-11.9492*	0.44181	0.00000	Fixed * MAP
	SE * MAP	0.7680	0.43765	0.08000	-
	Fixed * MAP vs. SE * MAP	12.7172*	0.43427	0.00000	Fixed * MAP
Standard error	SE	Fixed * MLE vs. SE * MLE	0.01950	0.02852	0.495
	Fixed * MLE vs. Fixed * MAP	0.0820*	0.02831	0.00400	Fixed * MLE
	Fixed * MAP vs. SE * MAP	0.0163	0.02804	0.56100	-
	SE * MLE vs. Fixed * MAP	0.0625*	0.02831	0.02800	SE * MLE
	SE * MAP	-0.0032	0.02804	0.91000	-
	Fixed * MAP vs. SE * MAP	-0.0657*	0.02782	0.01900	SE * MAP

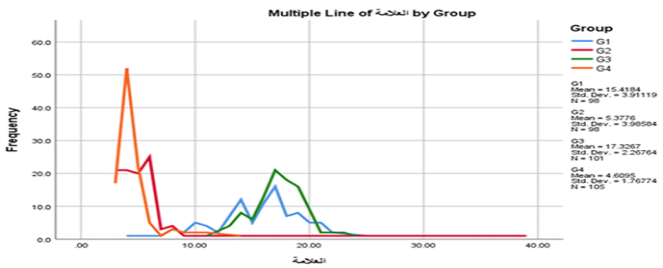

Figure 2. The effect of interaction between estimation methods and termination rules on test accuracy based on the score indicator (Source: Authors' own elaboration)

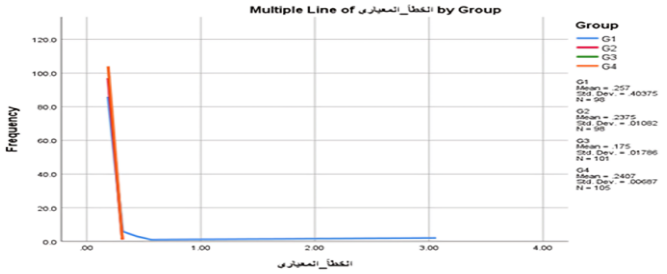

Figure 3. The effect of interaction between estimation methods and termination rules on test accuracy based on the standard error indicator (Source: Authors' own elaboration)

The researchers' findings regarding the accuracy of the test based on the examinee's score indicator highlight the following points:

1. Significant differences between (fixed * MLE) and (SE * MLE), favoring the fixed * MLE method. This indicates that when using MLE, the fixed termination method is better than the SE-based termination method.
2. Significant differences between (fixed * MLE) and (fixed * MAP), favoring the fixed * MAP method. This suggests that when using the fixed termination method, the MAP estimation method performs better than the MLE method.
3. Significant differences between (fixed * MLE) and (SE * MAP), favoring the fixed * MLE method. This indicates that test accuracy is better when using the combination of MLE and a fixed termination method compared to the MAP method with a SE-based termination.
4. Significant differences between (SE * MLE) and (fixed * MAP), favoring the fixed * MAP method. This suggests that test accuracy is better when using fixed * MAP compared to SE * MLE.
5. Significant differences between (fixed * MAP) and (SE * MAP), favoring the fixed * MAP method. This indicates that test accuracy is better when using fixed * MAP compared to SE * MAP.
6. No significant differences between (SE * MLE) and (SE * MAP), indicating that test accuracy is similar when using SE * MLE compared to SE * MAP.

Figure 2 shows the effect of interaction between estimation methods and termination rules on test accuracy based on the score indicator.

The researchers also noted the following results regarding test accuracy based on the SE indicator:

1. Significant differences between (fixed * MLE) and (fixed * MAP), favoring the fixed * MLE method. This indicates that when using the fixed termination method, MLE is more accurate than MAP.
2. Significant differences between (SE * MLE) and (fixed * MAP), favoring the SE * MLE method. This suggests that test accuracy is better with SE * MLE compared to fixed * MAP.
3. Significant differences between (fixed * MAP) and (SE * MAP), favoring the SE * MAP method. This indicates that test accuracy is better with SE * MAP compared to Fixed * MAP.
4. No significant differences were found for the remaining comparisons.

Figure 3 shows the effect of interaction between estimation methods and termination rules on test accuracy based on the SE indicator.

DISCUSSION

Commentary of the Findings Responsive to the First Research Problem

For this question, the results indicated that the accuracy of CAT math was not dependent on the method of ability estimation applied. This can be explained by the fact that this research used two indicators namely the examinee's mark of the SE. The F-test value was 0.468 which as indicated by the p-value of 0.494,was not significant. This means that regardless of the method used in estimating the accurateness of ability, MLE or MAP, the result remains the same in both conditions. This confirms what was reported by Chen et al. (1998) where it was established that the average number of items administered using minimum SE rule across the two ability estimation methods estimation methods, MLE and MAP, was almost the same. The findings also showed that the mean information functions of the MLE and MAP models of the tests using the minimum SE termination rule were roughly the same.

Interpretation of the Findings Associated With the Second Research Question

The accuracy of CAT in mathematics was dependent on the termination method used, as indicated by the results to this question. This can be so due to the fact that the study used the candidate's score as a parameter in the different samples of the study. The calculated value of the F-test was 1,283.386 and was found to be significant at $p = 0.00$. This shows that the accuracy of the test varies depending on the termination method. In order to find out which termination method was superior, a t-test of independent samples was carried out, the results of which showed that the termination method based on a fixed number of items provided higher accuracy of ability estimation than the termination method with a SEM.

Findings, herein discussed, corroborate the outcomes of Yildiz et al. (2024) which sought to test the hypothesis that CAT can satisfactorily estimate a candidate's level of cognitive ability using Raven's matrices tests. The research concludes that for all methods retained, the termination rule which was a fixed number of items yielded better estimates of ability and had higher information function than the minimum SE rule.

Discussion of the Results Related to the Third Research Question

The results for this question were attributed to the interaction between the two indicators, the ability estimation methods and test termination methods. They were able to demonstrate statistically significant differences in the accuracy of CAT. Statistical significance was evident with F-test: $F(4, 20) = 456.15$.

The results also showed sufficient evidence in support of the claim that CAT accuracy when MAP method is used for estimating abilities and when termination rule of SE is minimized to 0.25 is high. This indicates that MAP combined with SE makes the performance of CAT more accurate than Fixed Items combined with MAP. Stafford et al. (2019), Dodd (1989), and Babcock and Weiss (2012) also supported this result, which stated that the rule of variable length, based on the SEM criterion, has a higher accuracy.

Also, the outcomes indicated that the MLE method of ability estimation coupled with a point estimate of 25 items yielded better CAT accuracy. This result is in line with Özyurt and Özyurt (2015) findings who mentioned that CATs created for purposes of probability unit in the mathematics for the 11th-grade students proved to be reliable when MLE was used to assist in the estimation of moderation and a fixed termination rule ranging between 15-20 units was employed.

CONCLUSIONS

The study reached several key conclusions, including the followings:

1. The accuracy of the CAT in mathematics does not vary depending on the ability estimation method used. Whether the ability is estimated using the MLE method or the MAP method, both provide a comparable level of accuracy based on the indicators of the score and the SE.
2. The accuracy of the CAT in mathematics does vary depending on the termination method used, with the fixed number of items termination method being more accurate in estimating abilities compared to the SE-based termination method.
3. The accuracy of CAT estimation, based on the indicators of the examinee's score and the SE, is influenced by the interaction between the ability estimation method and the test termination method. Specifically, CAT accuracy is higher when using MAP with SE compared to MAP with Fixed Items. Additionally, CAT accuracy is higher when using the MLE method for ability estimation with a fixed termination rule of 25 items.

Recommendations

In light of the study results, the researchers recommend the following:

1. Use the fixed number of items termination method in CAT, as it provides better accuracy in adaptive testing.
2. Use the MLE method for ability estimation with a fixed termination rule of 25 items, thereby increasing the accuracy of CAT.

3. Use the MAP method with the SE rule to enhance the accuracy of CAT.
4. Educational institutions, including schools and universities, should adopt computerized adaptive tests.
5. Design high-quality item banks for different subjects.
6. Conduct further studies on the review process for student responses in CAT.

Author contributions: LD: analysis, data collection, statistical analysis, writing – original draft; NA: writing – review & editing, analysis, proofreading. Both authors approved the final manuscript. Both authors agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study adhered to the highest ethical practices. In relation to the conduct of the research, the authors have obtained permission from Yarmouk University (IRB\ 2024\642). The authors further stated that the privacy and anonymity of the research participants were strictly observed. Written informed consents were obtained from the participants.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

AlAli, R., & Wardat, Y. (2024). Enhancing student motivation and achievement in science classrooms through STEM education. *STEM Education, 4*(3), 183-198. https://doi.org/10.3934/steme.2024012

Al-Barbari, R. S. I. (2020). The design patterns of fixed and variable length adaptive electronic tests and their impact on reducing test anxiety and developing attitudes towards electronic tests among students of the faculty of education. *Educational Technology, 30*(1), 23-87. https://doi.org/10.21608/tesr.2020.91492

Alkan, V., & Deniz, K. (2023). Developing a computerized adaptive test form of the occupational field interest inventory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 47-61. https://doi.org/10.21031/epod.1153713

Alneyadi, S., Abulibdeh, E., & Wardat, Y. (2023). The impact of digital environment vs. traditional method on literacy skills; reading and writing of Emirati fourth graders. *Sustainability, 15*(4), Article 3418. https://doi.org/10.3390/su15043418

Audrey, J. (2019). Study of exposure control methods using a variable-length computerized reducer using the partial credit model. *Applied Psychological Measurement, 43*(8), 624-638. https://doi.org/10.1177/0146621618824856

Ayanwale, M. A., & Ndlovu, M. (2024). The feasibility of computerized adaptive testing of the national benchmark test: A simulation study. *Journal of Pedagogical Research, 8*(2), 95-112. https://doi.org/10.33902/JPR.202425210

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1-18. https://doi.org/10.7333/1212-0101001

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. https://doi.org/10.1177/0146621620965730

Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education, 7*(2), 142-151. https://doi.org/10.11591/ijere.v7i2.12900

Burhanettin, O. & Selahattin, G. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies, 27*, Article 62736294. https://doi.org/10.1007/s10639-021-10853-0

Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58(4)*, 569-585. https://doi.org/10.1177/0013164498058004004

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

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*(2), 129-143. https://doi.org/10.1177/014662168901300202

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

Ebenbeck, N., Bastian, M., Mühling, A., & Gebhardt, M. (2024). Duration versus accuracy–What matters for computerised adaptive testing in schools? *Journal of Computer Assisted Learning, 40*(6), 3443-3453. https://doi.org/10.1111/jcal.13074

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Nijhoff Publishing. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R., Swamination, H., & Rogers, H. (1991). *Fundamentals of item response theory*. SAGE.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Zeitschrift Für Psychologie/Journal of Psychology, 216*(1), 12-21. https://doi.org/10.1027/0044-3409.216.1.12

Hori, K., Fukuhara, H., & Yamada, T. (2021). Item response theory and its applications in educational measurement part II: Theory and practices of test equating. *Wiley Interdisciplinary Reviews: Computational Statistics, 14*(7), e1543. https://doi.org/10.1002/wics.1543

Janpla, S., & Piriyasurawong, P. (2023). The architecture of an intelligent multilevel item bank system for higher education graduate standardized testing. *Journal for ReAttach Therapy and Developmental Diversities, 6*(9s[2]), 1393-1405.

Jatobá, V. M. G., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society, 26*, Article 4. https://doi.org/10.1186/s13173-020-00098-z

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students admission to higher education programs in Turkey? *Educational Sciences: Theory & Practice, 17*(2), 573-596.

Lee, Y., & Kim, J. (2020). Efficiency of computerized adaptive testing in reducing item exposure. *Psychological Methods, 25*(3), 537-553.

Leroux, A. J., & Dodd, B. G. (2014). A comparison of stopping rules for computerized adaptive screening measures using the rating scale model. *Journal of Applied Measurement, 15*(3), 213-226.

Luo, H., Cai, Y., & Tu, D. (2020). Procedures to develop a computerized adaptive testing to advance the measurement of narcissistic personality. *Frontiers in Psychology, 11*. https://doi.org/10. 3389/fpsyg.2020.01437

Liu, Q., Zhuang, Y., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y., Pardos, Z. A., Ma, H., Zhu, M., Wang, S., & Chen, E. (2024). Survey of computerized adaptive testing: A machine learning perspective (arXiv:2404.00712). arXiv. https://doi.org/10.48550/arXiv.2404.00712

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package cat R. *Journal of Statistical Software, 48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Magis, D., Yan, D., & Von Davier, A. (2017). *Computer adaptive testing and multistage testing using R: Using packages cat R and mst R*. Springer. https://doi.org/10.1007/978-3-319-69218-0

Mujtaba, D. F., & Mahapatra, N. R. (2020). Artificial intelligence in computerized adaptive testing. In *Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence* (pp. 649-654). https://doi.org/10.1109/CSCI51800.2020.00116

Nour ElDin, A. (2019). The effectiveness of computerized adaptive measurement in assessing university students' achievement. *Saudi Journal of Educational and Psychological Sciences, King Saud University, 64*(5), 29-47.

Özyurt, H., & Özyurt, Ö. (2015). Ability level estimation of students on probability unit via computerized adaptive testing. *Eurasian Journal of Educational Research,* (58), 27-44.

Rios, J. (2022). An examination of individual ability estimation and classification accuracy under rapid guessing misidentifications. *Applied Measurement in Education, 35*(4), 300-312. https://doi.org/10.1080/08957347.2022.2155653

Saleh, S., AlAli, R., Wardat, Y., Al-Qahtani, M., Soliman, Y., & Helali, M. (2023). Structural relationships between learning emotion and knowledge organization and management processes in distance learning environments: An applied study. *European Journal of Investigation in Health, Psychology and Education, 13*(9), 1569-1589. https://doi.org/10.3390/ejihpe13090114

Stafford, R., Runyon, C., Casabianca, J., & Dodd, B. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior Research Methods, 51*(3), 1305-1320. https://doi.org/10.3758/s13428-018-1068-x

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation, 26*, Article 11.

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(1), 181-203. https://doi.org/10.1017/S027226312000025X

Tian, X., & Dai, B. (2020). Developing a computerized adaptive test to assess stress in Chinese college students. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00007

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In B. P. Veldkamp, & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 291-305). Springer. https://doi.org/10.1007/978-3-030-18480-3_15

Wainer, H. (2000). C*omputerized adaptive testing: A primer*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410605931

Wang, N., Wang, D., & Zhang, Y. (2020). Design of an adaptive examination system based on artificial intelligence recognition model. *Mechanical Systems and Signal Processing, 142*, Article 106656. https://doi.org/10.1016/j.ymssp.2020.106656

Weiss, D. J., & Şahin, A. (2024). *Computerized adaptive testing: From concept to implementation*. The Guilford Press.

Wulandari, F., Hadi, S., & Haryanto, H. (2020). Computer-based adaptive test development using fuzzy item response theory to estimate student ability. *Computer Science and Information Technology, 8*(3), 66-73. https://doi.org/10.13189/csit.2020.080302

Yildiz, H., Tunaboylu Demir, C., Ulku, S., Giray, G., & Kelecioğlu, H. (2024). Investigation of measurement precision and test length in computerized adaptive tests under different conditions. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 5-17. https://doi.org/10.21031/epod.1068572

Zio, E. (2018). The future of risk assessment. *Reliability Engineering & System Safety, 177*, 176-190. https://doi.org/10.1016/j.ress.2018.04.020

Žitný, P., (2011). Presnosť, validita a efektívnosť počítačového adaptívneho testovania [Accuracy, validity, and effectiveness of computer-based adaptive testing]. *Československá Psychologie, 55*(2), 167-179.