Exploring the Effects of Test Length, Number of Alternatives, and Sample Size on the Accuracy of Individual Estimates and Item Parameters within the Framework of Item Response Theory

BY

Dr. Asem Abdelmageed Kamel Ahmed
Assistant Professor, Department of Educational Psychology,
Graduate School of Education, Cairo University
E-mail: Dr.as_hg@cu.edu.eg

Dr. Amr Mohamed Ibrahim Youssef
Assistant Professor, Department of Educational Psychology,
Graduate School of Education, Cairo University
E-mail: Dr_amr@cu.edu.eg

***Study Summary*:**

The present study aimed to investigate the effects of test length, the number of response alternatives, and different sample sizes on the accuracy of individual ability estimates and item parameters based on Item Response Theory (IRT). The researchers generated sample data using the "WinGen" software, considering various sample sizes ranging from 500 to 2000, different test lengths (ranging from 30 to 50 items), and multiple response formats (three-, four-, and five-option items).

The key findings indicated that the accuracy of ability estimation, as well as the accuracy of the difficulty and discrimination parameters, was not affected by test length. However, the guessing parameter was influenced by test length. Additionally, the accuracy of the theta (θ) parameter was higher with four- and five-option items. In contrast, the discrimination, difficulty, and guessing parameters were not affected by the number of response options. The results also revealed that ability estimation was most accurate with a sample size of 500, whereas the accuracy of the discrimination parameter improved with larger sample sizes, particularly with a sample size of 2000.

***Keywords*:** Test length, number of response alternatives, sample sizes, item parameters, individual abilities, Item Response Theory (IRT).

Introduction:

The Classical Test Theory (CTT) has been widely used for a long time in the construction, development, analysis, and interpretation of psychological and educational tests (Adetutu & Lawal, 2022). However, it has faced several criticisms, including major limitations in achieving objectivity, the absence of a fixed measurement unit, and the inability to account for multidimensional constructs. Additionally, CTT relies heavily on the measuring instrument and the sample to which the test is applied, making its estimates dependent on specific conditions(Ali, 2022).

In response to these limitations, researchers have made significant efforts to develop more robust measurement approaches, leading to the emergence of a modern measurement theory known as Item Response Theory (IRT). This theory is also referred to as Latent Trait Theory and is distinguished by the independence of item parameters—difficulty, discrimination, and guessing—from the abilities of the individuals assessed. Additionally, the theta (θ) parameter, which represents individuals' ability, remains independent of the specific items used for(Ali & Istiyono, 2022).

IRT has led to the development of several mathematical models based on specific assumptions to define the relationship between an individual's performance on a test and the underlying abilities that influence that performance. The most common IRT models include:

- The one-parameter model (Rasch Model)
- The two-parameter model (Lord Model)
- The three-parameter model (Birnbaum Model) (Aybek, 2023)

Among these models, the three-parameter model is the most comprehensive, as it describes the item characteristic curve using three parameters: item difficulty, item discrimination, and guessing. This model is particularly advantageous over the other two models because it accounts for the guessing factor, which is likely to occur in multiple-choice tests and may influence the accuracy of the theta (θ) parameter estimation (Apriyani et al., 2023; Wang et al., 2022; Zhang et al., 2022)

A review of previous studies and measurement literature suggests that item parameters and ability estimates may be affected by variations in test length, sample size, and the number of response alternatives.(Aybek, 2023) The availability of computer-based programs has facilitated the calculation of IRT model statistics, including those for the three-parameter model. Therefore, this study aims to examine the impact of test length, the number of response alternatives, and sample size on the accuracy of ability and item parameter estimates using simulated data generated by the "WinGen" software.

Research Problem and Questions:

As the limitations of Classical Test Theory (CTT) continue to grow, the advantages of Item Response Theory (IRT) become increasingly evident. One of its key strengths lies in the independence of item characteristics from the individuals being assessed and the independence of the theta (θ) parameter from the items used. Unlike CTT, which interprets individual scores based on a reference group, IRT evaluates individual performance in relation to the test items themselves. Furthermore, standard error estimation is conducted for each test-taker individually, allowing for meaningful comparisons of individuals' performance even when different measures of the same trait are used (Bjorner et al., 2023; Brucato et al., 2023)

Multiple-choice tests (MCTs) are widely used for assessing ability and achievement, particularly in Egypt's education system. However, these tests face a significant challenge: guessing. When test-takers are unsure of the correct answer, they may resort to guessing, which, if correct, artificially inflates their final score beyond their true ability level. This compromises the accuracy of ability estimation, as it becomes difficult to determine whether a correct response reflects actual knowledge or was simply a result of guessing(Cai et al., 2023).

To address this issue, measurement researchers have refined IRT models as a means of achieving more accurate ability estimates, even in the presence of guessing. IRT conceptualizes an individual's response to a test item as a function of both the person's latent ability and the characteristics of the test item itself. This allows for predicting an individual's performance on a given item based on their estimated ability level (Cotter et al., 2023; Gikaro et al., 2024; Huang, 2023).

However, the practical application of IRT requires that its assumptions be met, such as One-dimensionality and the item characteristic curve (ICC) structure. Data may not fit the one-parameter (Rasch) model if the discrimination or guessing parameters vary across items. Similarly, the two-parameter model may be inadequate if the guessing parameter is not properly accounted for (Garcia et al., 2023; Hanzlová & Lynn, 2023; Huang et al., 2023)Notably, low-ability test-takers often resort to guessing in multiple-choice tests, leading to an increase in the guessing parameter beyond zero. This violates the assumptions of one- and two-parameter models, making the three-parameter model a more suitable choice for handling guessing effects and improving measurement accuracy.

Despite extensive research on IRT applications, there remains a need for more empirical data to assess how various testing conditions influence IRT-based estimates. Based on the theoretical considerations discussed, this study aims to investigate the impact of test length, number of response

alternatives,(Huang, 2023) and sample size on item and ability parameter estimates using the three-parameter model.

Specifically, the study seeks to answer the following main research question: "**How do variations in test length, number of response alternatives, and sample size affect the accuracy of the theta (θ) parameter and item parameter estimates under the three-parameter model?**"

Sub-Questions of the Study:

1. What is the effect of test length on the ability parameter of individuals according to the three-parameter model?
2. What is the effect of test length on item parameters according to the three-parameter model?
3. What is the effect of the number of response options on the ability parameter of individuals according to the three-parameter model?
4. What is the effect of the number of response options on item parameters according to the three-parameter model?
5. What is the effect of sample size on the ability parameter of individuals according to the three-parameter model?
6. What is the effect of sample size on item parameters according to the three-parameter model?

Study Objectives:

The present study aims to achieve the following objectives:

1. Verify the assumptions of Item Response Theory (IRT) based on the responses of the study sample to different test models.
2. Examine the estimates of individuals' abilities in multiple-choice tests used in the study within the framework of the three-parameter model.
3. Assess the accuracy of the three-parameter model in estimating the Theta (θ) parameter of individuals and item parameters under specific hypothetical conditions, including varying test lengths and different sample sizes.
4. Investigate the impact of test length, sample size, and the number of response options on the accuracy of the three-parameter model in estimating the Theta (θ) parameter of individuals and item parameters.

Significance of the Study:

- The significance of this study stems from its focus on multiple-choice questions, a format that has become increasingly prevalent in student assessment in recent years.

- It contributes to improving our understanding of the factors influencing the accuracy of individual score estimation based on Item Response Theory (IRT).
- It enhances our comprehension of the factors affecting item parameters according to Item Response Theory (IRT).
- Its findings contribute to the development of test models with high psychometric properties using multiple-choice questions.

Theoretical Framework and Related Studies:

An **achievement test** is defined as:

"A set of questions or stimuli representing a specific trait or ability, formulated in a structured manner to determine the student's level of skills and knowledge acquired in a particular subject through their responses." (Reise & Moore, 2023)

An **objective test** is one type of achievement test, which is relatively modern compared to essay-based tests. It is termed "objective" due to its accuracy, reliability, and resistance to examiner bias (Silveira et al., 2023)Among the most widely used types of objective test questions in education is the **multiple-choice question (MCQ)**. A well-constructed MCQ can measure both simple and complex learning objectives.

Advantages of Multiple-Choice Tests:

- High efficiency, especially when test items are well-structured.
- Versatile applications in research and different educational stages.
- Allows broader coverage of the behavioral domain being assessed(Siraji et al., 2023; Tang et al., 2023).

Multiple-Choice Tests:

A **multiple-choice question** consists of a **stem** followed by a set of possible answers, from which the examinee must select the correct one. These response options are called **alternatives**, with only one correct answer, while the remaining options function as **distractors**. Distractors serve to mislead test-takers who lack sufficient knowledge, making them appear plausible to low-ability individuals but not to those with high ability(Ayanwale et al., 2024).

The primary purpose of **distractors** is to **challenge examinees who do not know the correct answer, assess their need for additional knowledge, and identify their weaknesses when selecting incorrect alternatives**. Therefore, distractors should be **closely related to the correct answer** to ensure their effectiveness (Ayasse & Coon, 2024; Gilbert et al., 2025).

(Hu & Valdivia, 2024; Jones et al., 2024; Young et al., 2025)emphasize the necessity of ensuring homogeneity among response alternatives to prevent examinees from eliminating certain options as a strategy to deduce the correct answer. Consequently, the length of the correct answer or key should be equal to the length of the distractors, as variations in their length may provide examinees with unintended cues. Additionally,

all alternatives should align with the level of information presented in the item stem. Furthermore, it is crucial to consider the examinee's age and educational background, as what is appropriate for one age group may not be suitable for another.

The **psychometric properties of the test and item parameters** are influenced by several factors, including:

1. **Item Wording:** The item stem must be precisely formulated to ensure clarity, specificity, and uniform interpretation across all examinees. Ambiguous or vague stems may lead to varied understandings of the required response, increasing the likelihood of incorrect answers (Ma et al., 2024).
2. **Selection of Response Alternatives:** When designing distractors, their homogeneity should be carefully considered. A set of homogeneous alternatives enhances examinees' engagement and increases item difficulty. In contrast, heterogeneous alternatives lower item difficulty, thereby reducing its discriminative power. Notably, items that are either too difficult or too easy tend to exhibit weak discrimination(Mangold, 2024; Uto, Tomikawa, et al., 2023; Wang et al., 2023).
3. **Structure of Alternatives:** Response alternatives vary across different tests. According to(Harrison et al., 2023; Shibata & Uto, 2022; Wortham et al., 2023), some multiple-choice items require selecting a single correct response from the given options, which is the simplest and most commonly used format, particularly in achievement tests. Other items may require examinees to choose the **best possible answer**, allowing for multiple plausible solutions, necessitating the selection of the most appropriate one. Additionally, some response formats involve **compound alternatives**, where a single response option combines two choices(Silvia, 2022; Siraji & Haque, 2022). Some items also require selecting multiple correct answers; however, this format is generally discouraged in achievement tests.
4. **Number of Response Alternatives:** The number of alternatives in multiple-choice items varies across tests. However, it should be sufficient to minimize guessing while maintaining accuracy. Although the number of response options should generally not exceed five, they must be carefully selected to ensure relevance and precision. The emphasis should be on the **quality** of alternatives rather than their quantity. Furthermore, for younger students, reducing the number of alternatives is recommended to enhance comprehension and decision-making(Soland, 2022; Stavropoulos et al., 2022; Toraman et al., 2022).

Item Response Theory (IRT):

Item Response Theory (IRT) emerged as an extension of Classical Test Theory (CTT), complementing its principles and addressing many measurement challenges that CTT could not fully overcome(Vaganian et al., 2022; Wang et al., 2022).

IRT and its associated mathematical models aim to estimate both item parameters and individual abilities. The more closely the dataset aligns with the chosen model, the more precise these estimates become. Items and individuals are positioned on an ability scale through estimation processes, provided that a plausible relationship exists between the expected probabilities of individuals' responses and their actual performance at each ability level(Kawakubo et al., 2024; Raykov, 2023; Wolcott et al., 2022).

IRT shares commonalities with CTT, such as the presence of a **latent trait continuum**. The likelihood of an individual correctly responding to an item can be predicted based on their position on this continuum. This probability is a monotonically increasing function of the individual's ability level, meaning that higher ability levels are associated with a greater probability of answering correctly(Charamba et al., 2023; Cotter et al., 2023; Zhang et al., 2022).

One fundamental distinction between IRT and CTT lies in the differentiation between an individual's **true ability** and the **estimated ability score**. Unlike CTT, where ability estimates are sample-dependent, IRT assumes that an individual's ability remains constant regardless of the sample's characteristics (Miller et al., 2022; Moreta-Herrera et al., 2024). However, IRT requires complex and extensive mathematical computations, making it impractical before the advent of computer-based statistical programs(Apriyani et al., 2023; Uto, Tomikawa, et al., 2023).

A key advantage of **modern measurement theories (IRT) over classical approaches (CTT)** is that IRT focuses on individual response patterns to test items, whereas CTT relies on total raw scores. Additionally, IRT allows for the estimation of measurement error at **any ability level**, unlike CTT, which assumes a uniform error distribution across all test scores(Charamba et al., 2023; Donaldson et al., 2023).

Assumptions of Item Response Theory (IRT):

IRT is built upon strong assumptions that are not always easily met in real-world data. For IRT models to yield reliable results, the dataset must adequately satisfy these assumptions before evaluating model fit(Effatpanah & Baghaei, 2023; Fernandes et al., 2023; Kawakubo et al., 2024; Veldkamp et al., 2025). These assumptions include:

1. **One-dimensionality:**

IRT assumes that a single latent trait accounts for individuals' responses to test items. The test developer presumes that test-takers' performance on the items is explained by **one underlying ability**. However, achieving perfect One-dimensionality is challenging due to potential influences from **personal or cognitive factors**, such as motivation, test anxiety, and other external variables. To satisfy this assumption, a **dominant factor** must primarily influence test performance—representing the ability being measured. Ensuring that the test assesses only **one trait** enhances the accuracy and validity of its interpretations (Huang et al., 2023; Jiang et al., 2023).

When test items are homogeneous and measure the same trait, answering any given item requires **similar cognitive and behavioral processes**. Factor analysis is often employed to identify the **primary factor** influencing performance. If multiple traits are detected, items can be grouped into homogeneous clusters through factor analysis, after which **IRT models** can be applied separately to each homogeneous set (Gilbert et al., 2024; Kiliç et al., 2023).

2. **Local Independence:**

   Local independence assumes that an individual's response to a particular test item **should not be influenced** by their responses to other items in the same test. Instead, the individual's **latent ability** should be the **only factor** determining their responses. This assumption is closely linked to **One-dimensionality**—in fact, local independence is considered an equivalent assumption. However, they are not identical concepts.

   A test may be **multidimensional** if two or more latent traits influence item responses, provided that the items remain **independent** within groups of individuals with similar trait levels. The number of dimensions in a test corresponds to the number of **latent traits required** to achieve local independence(Gewily et al., 2024; Santos et al., 2023).

3. **Item Characteristic Curve (ICC):**

   The **ICC** mathematically describes the relationship between an individual's probability of answering an item correctly and their **latent ability level**. This relationship follows a **nonlinear regression function**. Given individuals' observed scores at different ability levels, it is possible to plot the ICC, which represents the regression curve passing through the **conditional distribution means** at each ability level(Cook & Wind, 2024; Uto, Aomi, et al., 2023).

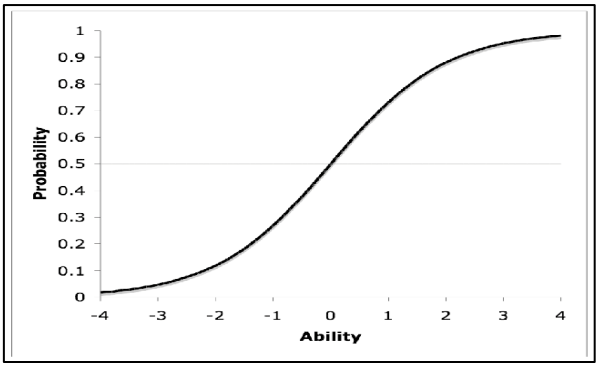


Figure 1: Item Characteristic Curve (ICC)

In the figure above, the **horizontal axis** represents the **ability continuum (θ)** measured by the item, while the **vertical axis** represents the **probability of answering the item correctly (P(θ))**. As an individual's ability increases, so does the probability of correctly responding to the item.

The probability of an individual answering an item correctly depends solely on the **shape of the ICC**, rather than the **number of individuals** at the same ability level. This property, known as **invariance of item characteristic curves**, ensures that item parameters remain stable across different examinee populations for whom the items have been calibrated. This is a fundamental feature of **IRT models(Gao et al., 2024; Gewily et al., 2024)**.

4. **Independence from Speededness**

This assumption is implicitly related to **unidimensionality**, as IRT models assume that tests are **not administered under strict time constraints**. In other words, individuals who fail to answer items correctly do so due to **limited ability**, rather than insufficient time to complete the test. If speed influences performance, then two factors—**processing speed and the measured trait**—affect responses, violating the assumption of unidimensionality(Gikaro et al., 2024; Gilbert et al., 2024).

Item Response Theory (IRT) Models

IRT models aim to establish the relationship between an **individual's performance** on test items (which is directly observable) and their **latent ability** (which explains this performance). Since these models are **probabilistic**, they rely on **probability theory** to define response patterns) Guo et al., 2024(.

The choice of an appropriate **IRT model** depends on the **nature of test items, the number of items, and the sample size**. However, the most critical factor is the **type**

of data, which can be either **binary (0,1)** or **polytomous (more than two response options)**. Among the most widely used binary **IRT models** are:

1. The One-Parameter Logistic Model (1PLM) (Rasch Model)) Howe et al., 2024(

This is the simplest IRT model and the most commonly applied. It is also known as the **Rasch Model**, named after the Danish mathematician **Georg Rasch**.

- The **1PLM** assumes that only **item difficulty** distinguishes individuals' performance.(Gilbert et al., 2024)
- All items are assumed to have **equal discrimination** power.
- The lower asymptote of the **ICC** is set to **zero**, meaning that individuals with **low ability have no chance of guessing the correct answer**—thus eliminating the effect of guessing.

The **Rasch Model** is particularly useful when dealing with **small sample sizes**, as models with fewer parameters require **less precise data** for parameter estimation.(Guo et al., 2024)

2. The Two-Parameter Logistic Model (2PLM)

In this model, developed by Lord, the item characteristic curve takes the form of a **logistic distribution with two parameters**: **discrimination** and **difficulty**. This is an extension of the **Rasch Model** by adding the **discrimination parameter**, as it is difficult to find multiple items that distinguish consistently between ability levels measured by the test. At the same time, there is no room for guessing, as individuals' responses on test items are not influenced by guessing (Frick et al., 2024; Gao et al., 2024; Howe et al., 2024)

3. The Three-Parameter Logistic Model (3PLM)

This model extends the **two-parameter model** by adding a third parameter, **guessing**, to the logistic distribution of item characteristics, which includes the **difficulty**, **discrimination**, and **guessing** parameters.(Gibbons et al., 2024)

Test designers must determine the model they will use in advance to verify the **data's fit** to the chosen model. This is done by examining the item characteristic curves, which help ensure that the assumptions of the chosen model are met(Gilbert et al., 2024).

In the study conducted by(Young et al., 2025), the goal was to examine the impact of sample size and test length on the reliability of the test and test calibration using the **partial estimation model**. Three sample sizes (200, 500, 1000) and five test lengths (2, 4, 8, 12, 20 items) were used, with data generated via simulation. The **Root Mean**

Square Error of Differences (RMSD) was used to evaluate the accuracy of parameter estimates. The study's results indicated that parameter accuracy increases with the number of items in the test (test length).

In (Yiğiter & Boduroğlu, 2024)study, the goal was to examine the effect of sample size and number of test items on the **theta coefficient** for individuals and item parameters according to the **one-parameter model**. The sample sizes were (50, 100, 500) individuals, with tests of different lengths (25, 50, 300 items). The results showed significant differences in the interaction of sample size and number of items in the accuracy of the individuals' **theta coefficients**, attributed to the interaction between sample size and number of items.

(Tomikawa et al., 2024) focused, in part, on examining the effect of sample size on the accuracy of **item parameter estimates**. The **three-parameter model** was used to calibrate 360 binary items, using the simulation method. It was assumed that the first 120 items formed the common part of the test, while the remaining items formed four subtests, each containing 60 items. The sample sizes ranged from 250 to 1500 individuals, assuming that individual estimates were normally distributed. A total of 14 conditions were formed based on sample size and ability distribution. For each condition, the test with 360 items was calibrated using the **Bilog, Bilog-MG, PIC** software, and the **Root Mean Square Difference (RMSD)** criterion was used to compare the accuracy of parameter estimates. The study concluded that estimation errors for **difficulty and discrimination parameters** were larger when the sample size was smaller.

The study conducted by (Sinharay & Monroe, 2024) aimed to apply the **three-parameter model** to estimate individuals' ability and item parameters for a multiple-choice test, with variations in the number of response alternatives. The study was applied to a sample of 1200 students using the **BILOG** software to estimate the **theta coefficient** for individuals and item parameters in light of modern theory. The results showed no statistically significant differences in individuals' ability on the multiple-choice test. However, statistically significant differences were found in the **difficulty parameter** depending on the number of alternatives, favoring the test with three alternatives. The results also indicated statistically significant differences in the **discrimination parameter** of the items on the multiple-choice test, with the five-alternative test being preferred. Furthermore, statistically significant differences were found in the **guessing parameter** of the items, favoring the three-alternative test.

The study conducted by (Shi et al., 2024)aimed to examine the impact of sample size, selection method, number of items, and selection method on the accuracy of item parameters and ability estimates according to the **three-parameter model**. A test

consisting of four sub-tests with 71 items was developed, and the sample size was 1000 individuals. The **Bilog 3.11** software was used to estimate the **theta coefficient** for individuals, item parameters, standard errors of estimation, and the data's fit to the **three-parameter logistic model**. The results showed a direct relationship between sample size and accuracy of item parameters. Individual ability estimates were stable when using large calibration samples. The accuracy of the **theta coefficient** was influenced by the data's fit to the model, and it was uncertain whether increasing the sample size beyond a certain threshold would result in greater accuracy. The results also indicated that the accuracy of the **discrimination parameter** increased with greater variability in the ability of the examinees. The accuracy of the **difficulty** and **ability parameters** improved when the range of ability levels in the examinees matched the difficulty range of the items. Moreover, the accuracy of the **guessing parameter** increased when using a sample of low-ability examinees for item calibration. The results also showed an increase in the accuracy of the **ability parameter** when the number of test items or their proportion to the total test increased.

(Liang et al., 2024)study aimed to examine the suitability of multiple-choice tests with three alternatives compared to those with different numbers of alternatives. The study analyzed a collection of research and studies conducted over 80 years to investigate the impact of alternatives on the psychometric properties of tests. Rodriguez compiled results from 27 studies conducted between 1920 and 1999, including 56 tests. The results showed that the three-alternative test was the best in most studies. The study found that changing the number of alternatives from five to four resulted in a decrease in the **difficulty parameter** by an average of 0.02, the **discrimination parameter** by 0.04, and the **reliability coefficient** by 0.035. Reducing the number of alternatives from five to three by removing the least attractive alternative led to a reduction in the **difficulty parameter** by 0.07, without affecting the **discrimination** and **reliability coefficients**. In cases where alternatives were randomly deleted to arrive at three, the **reliability coefficient** decreased by 0.06. When the number of alternatives was reduced from four to three, the **difficulty parameter** decreased by 0.04, while both the **discrimination** and **reliability coefficients** increased by 0.03 and 0.02, respectively. Furthermore, reducing the number of alternatives to two, whether from five or four, made the test items easier and reduced the **discrimination coefficients**.

The study conducted by(Khatri et al., 2024) aimed to examine the impact of test length and sample size on the **theta coefficient** for individuals using the **Bayesian method**. The study generated **dichotomous data** for tests with two levels of item numbers (200 and 440 items) and three sample sizes (500, 1000, and 2000 individuals). The results showed a relationship between test length and estimation accuracy. As the

sample size increased, the accuracy of the **theta coefficient** for individuals and item parameters improved, while the **standard errors** decreased.

The study by(Gikaro et al., 2024) aimed to investigate the effect of reducing the number of alternatives in a multiple-choice test on its **psychometric properties**. Two test models were prepared: the first consisted of 38 items with four alternatives each, applied to 1000 students; the second consisted of 38 items, with 10 items having four alternatives and 28 items with three alternatives (after deleting the least attractive alternative), applied to 192 students. The results revealed no statistically significant differences in **difficulty** and **discrimination parameters** between the two tests. It was found that a test with three alternatives served its purpose as effectively as one with four alternatives, as increasing the number of alternatives increased the likelihood of **guessing**, due to limited response time.

(Gibbons et al., 2024) study aimed to examine the stability of item parameters by comparing computerized testing programs with paper-based tests of varying lengths. The study results showed a positive correlation between the number of items and the stability of item parameters. Specifically, as the length of the test increased with more items, the stability of the item parameters also increased.

The study conducted by (Gershon et al., 2024)aimed to investigate the impact of the number and attractiveness of alternatives in multiple-choice test items on their compatibility with the **three-parameter model**. A test was designed with three versions, each containing 50 items. The first version had five alternatives per item, the second had three alternatives after randomly removing two alternatives from the first version, and the third had three alternatives after removing the least discriminative alternatives from the first version. The test was applied to 1656 students. The results showed that the items in all three versions of the test were compatible with the **three-parameter model**. The test with five alternatives was found to be the best among the three versions, regardless of the deletion method. There was no impact of the **discrimination** of alternatives in multiple-choice test items, even after removing the least discriminative alternatives from the third version or removing them randomly from the second version.

(Doğan & Atar, 2024) study aimed to examine the effect of the number of test items on item parameters according to the **one-parameter model**. The test contained different numbers of items (5, 10, and 15 items). The results showed that to achieve greater stability in item parameters, tests should contain more than 15 items.

The study by (Cook & Wind, 2024)aimed to investigate the impact of sample size on the accuracy of item parameters and ability estimates in tests developed according to

the **Item Response Theory (IRT)** models. Data was generated for sample sizes of (500 and 1000) individuals and test lengths of (10 and 20) items. The results indicated that both sample size and the number of test items affect item parameters when the sample size is 1000 individuals and the test length is 20 items. The results also showed that the ability parameter was not affected by sample size but was influenced by test length.

(Baghaei & Effatpanah, 2024) study aimed to examine the results of the ability parameter and item difficulty using five different **IRT models**, with varying levels of guessing, sample sizes, and test lengths. Data for 50 different scenarios were generated using varying conditions. The results indicated variability in the accuracy of item and individual parameters based on the level of guessing in the test, sample size, and test length. It was also found that the results for the ability parameter and item parameters depend on the accuracy criterion in each of the IRT models.

(Zhong et al., 2023) study aimed to explore the impact of sample size on item parameters using **Item Response Theory**. The sample sizes varied between (200 and 11,292) individuals, and a test consisting of 80 items was administered. The **BILOG-MG** software was used to estimate the parameters. The results showed that the difficulty parameter increased with larger sample sizes, with an average of 0.31 at a sample size of 200 individuals and a higher value at 11,292 individuals. It was also shown that the standard error for the difficulty parameter decreased as the sample size increased, with an error of 0.32 for the 200-person sample and 0.07 for the 11,292-person sample.

(Zhao et al., 2023) study aimed to examine the effect of the number of alternatives and the position of the strong distractor on the psychometric properties of the test and item parameters according to the **Item Response Theory**. A test consisting of 54 items was created and applied to 2,123 individuals. The test included four models: the first model had five alternatives with the strong distractor near the correct answer, the second model also had five alternatives but with the strong distractor far from the correct answer. In the third and fourth models, the two weakest alternatives were deleted: the third model had three alternatives with the strong distractor near the correct answer, and the fourth had three alternatives with the strong distractor far from the correct answer. The **three-parameter logistic model** and software **SPSS** and **BILOG-MG3** were used for analysis. The results showed no differences in the **difficulty** and **guessing** parameters for the items, despite changes in the number of alternatives and the position of the strong distractor. However, differences were observed in the **discrimination** parameter, attributed to the position of the strong distractor and its interaction with the number of alternatives.

In the study by(Wortham et al., 2023), which aimed to investigate the effect of item parameter estimation methods and individuals' abilities on the psychometric properties of the test, with consideration of sample size using Bayesian and maximum likelihood estimation methods, a multiple-choice test consisting of 33 items with four alternatives for each was constructed. The sample consisted of 1,000 students, and the analysis was conducted using the BILOG-MG software according to the three-parameter model. The results showed statistically significant differences at the 0.05 level in the mean standard errors of item parameter estimates due to the interaction between the estimation method and sample size. No statistically significant differences were found due to sample size and estimation method. The results also revealed statistically significant differences in the mean standard errors of ability estimates for individuals attributed to sample size and the interaction between the estimation method and sample size, but no significant differences were found due to the estimation method. Additionally, no significant differences were observed in the estimated reliability coefficients across different sample sizes (100, 500, 1000) participants.

The study by (Uto, Tomikawa, et al., 2023) aimed to investigate the effect of sample size on the information function of the test and its standard error estimates using Item Response Theory. The study used response data from 7,500 participants selected randomly and distributed into five different sample sizes (500, 1000, 1500, 2000, 3000), for a test consisting of 40 items of the multiple-choice type. Data analysis was carried out using the BILOG-MG3 software according to the three-parameter model. The results showed that the estimates of the information function were positively correlated with the sample size, while the standard error of the information function was inversely related to the sample size.

In the study by(Toledano-Toledano et al., 2023), which aimed to investigate the effect of sample size and test length on the theta coefficient for individuals and item parameters based on the three-parameter model, data generated from binary responses were used for samples of different sizes (100, 250, 500, 1000, 2000, 4000) and test lengths (10, 25, 50, 75, 100, 300). The results indicated that the accuracy of the theta coefficient for individuals and item parameters increased as the test length exceeded 50 items, and as the sample size exceeded 2000 participants.

In the study by (Tang et al., 2023), which aimed to evaluate the results of the Rasch model analysis using small sample sizes, a test consisting of 10 items was used for samples of sizes (30, 50, 100, 250), and data were analyzed using Mplus software. The results showed that the standard errors for item difficulty and ability parameters for sample sizes (30, 50) were higher than those for sample sizes (100, 250).

In the study by(Siraji et al., 2023), which aimed to compare the accuracy of item parameters based on the multidimensional graded response model with different sample sizes using maximum likelihood estimation, data were generated for samples of sizes (500, 1000, 1500, 2000) for tests consisting of (30, 90, 240) items. After analyzing the results using the flexMIRT software, the results showed that a sample size of 500 participants provided accurate item parameter estimates when using tests of lengths (30, 90) items, whereas for a test consisting of 240 items, a sample size of at least 1000 was necessary. Increasing the sample size beyond 1000 participants did not improve the accuracy of the item parameters.

In the study by(Silveira et al., 2023), which aimed to investigate the effect of sample size and the number of test items on the accuracy of item parameters in Item Response Theory, using the marginal maximum likelihood method for estimating item parameters through Xcalibre 4.1 software, three tests of lengths (10, 20, 30) items were constructed and applied to samples of sizes (150, 250, 350, 500, 750, 1000, 2000, 3000, 5000) participants. The results showed that a sample size of at least 150 could be used for tests with (10, 20, 30) items to estimate item difficulty parameters accurately according to the Rasch model. The mean standard errors of the difficulty parameter decreased as the sample size increased.

In the study by(Silva et al., 2023), which aimed to compare item parameter estimation methods in Item Response Theory models based on different sample sizes, data were generated for sample sizes (25, 50, 100, 250, 500, 1000) participants, with tests consisting of (10, 20, 30, 40, 50) items. Data analysis was carried out using R software. The results indicated that the pairing method provided more accurate estimates of item difficulty compared to the maximum likelihood and Bayesian methods. Additionally, the standard errors in item difficulty decreased with larger sample sizes, and increasing the number of test items did not necessarily lead to better accuracy in item difficulty.

In the study by(Shim et al., 2023), which aimed to examine the results of the item difficulty parameter for the Rasch model in light of changes in sample size and test length, binary response data were generated for four test models with different item lengths (20, 30, 40, 70 items) and five sample sizes (50, 100, 250, 500, 1000) participants. R software was used to estimate the item difficulty parameter using the pairing method and individual ability using the weighted likelihood method. The results showed statistically significant differences at the 0.05 level in the mean standard errors of the item difficulty parameter due to sample size, favoring the sample size of 1000 participants. The results also indicated small standard errors in the item difficulty parameter for all tests in the study, with values ranging from 0.002 to 0.05. Statistically

significant differences were also found at the 0.05 level in the mean standard errors of ability estimates due to sample size, test length, and the interaction between them.

In the study by(Santos et al., 2023), which aimed to investigate the effect of the number of multiple-choice test items on the accuracy of item parameters and individual abilities according to the three-parameter model, two test models were applied with (30, 60) items. BILOG-MG3 software was used to analyze and estimate the item parameters and individual abilities. The results showed that the means of the difficulty and discrimination parameters for the 60-item test were higher than the means of the difficulty and discrimination parameters for the 30-item test. No statistically significant differences were found between the mean guessing coefficients and the mean standard error in estimating them across the two test models. The study results also showed statistically significant differences in the means of individual ability and the means of standard errors for the theta coefficient, with the 60-item test showing higher means than the 30-item test, indicating that the number of items (60) had a greater impact on the theta coefficient.

The previous studies, including those by(Ayanwale et al., 2024; Ayasse & Coon, 2024; Baghaei & Effatpanah, 2024; Cook & Wind, 2024; Doğan & Atar, 2024; Frick et al., 2024; Gao et al., 2024; Gershon et al., 2024; Ma et al., 2024; Mangold, 2024; Santos et al., 2023; Shim et al., 2023; Silva et al., 2023; Silveira et al., 2023; Siraji et al., 2023; Tang et al., 2023; Toledano-Toledano et al., 2023; Uto, Aomi, et al., 2023; Uto, Tomikawa, et al., 2023; Wang et al., 2023; Wortham et al., 2023; Zhao et al., 2023; Zhong et al., 2023), shared a focus on the accuracy of item parameters and individual abilities in Item Response Theory models. Many of these studies were simulation studies, with data generated using software such as those mentioned above.

The experimental variables used in previous studies mostly focused on one to two variables, including sample size, test length, and the number of alternatives. The models used to compare the accuracy of item parameters and individual abilities were also varied. The three-parameter model was employed in studies by(Gewily et al., 2024; Guo et al., 2024; Howe et al., 2024; Hu & Valdivia, 2024; Jones et al., 2024). Various software programs were used, such as WinGen for data generation, BILOG, Xcalibre, and R software for estimating item parameters and individual abilities. Most of the estimation measures in the previous studies included correlation coefficients, efficiency indices, means of deviations, bias, and effect size. The current study serves as a continuation of the recommendations from previous studies to conduct future research involving simulation data to investigate the effects of sample size, test length, and the number of response alternatives on the accuracy of item parameters and individual ability using the three-parameter model. It is distinguished by having three sample sizes

(500, 1000, 2000), three levels of test item numbers (30, 40, 50), and also three levels of response alternatives (2, 3, 5).

Study Methodology and Procedures

Study Method:

The researchers in the current study employed a quasi-experimental design to answer its research questions. The reason for using this method is that it is the most appropriate for studying the effects of experimental variables that were controlled and adjusted through the simulation design on the dependent variable. It also provides an understanding of the relationship between independent and dependent variables, which is a directed causal relationship. The use of simulation-based data with Monte Carlo methods (MCM) achieves the highest level of experimental control, as the data is selected randomly from samples generated by the WinGen program (Han, 2007).In this study, all variables that could affect the results, such as the characteristics of individuals represented by their abilities, were controlled. A specific range with a defined mean and standard deviation was set for individual abilities. The item characteristics were also controlled, with all items being unidimensional, binary (0, 1), and having specific distributions within a range with a set mean and standard deviation for each parameter of difficulty, discrimination, and guessing. Thus, the experimental variables in the current study—test length (30, 40, 50 items), number of response alternatives (three, four, and five alternatives per item), and sample size (500, 1000, 2000 participants)—are the only ones that will affect the dependent variables, which are the theta coefficient (θ) for individuals and item parameters, isolated from any other variables.

Study Population:

The study population consists of all individuals who possess the specified ability level in the design, and all unidimensional binary response (0, 1) tests with the same distribution of parameters as defined in the study design.

Study Sample:

A random sample consisting of individuals with sample sizes of (500, 1000, 2000) participants, and random samples of tests with lengths of (30, 40, 50) items, with three, four, and five response alternatives per item. These samples are selected from data generated by the WinGen program according to the steps of the program and the data that align with the study design.

Test Conditions:

The data for the current study, which aims to compare the accuracy of individual ability estimates and item parameters under twenty-seven test conditions, were generated

based on the three-parameter logistic model. Each test condition was defined by a combination of three factors: the number of alternatives, the sample sizes, and the test lengths. Thus, data were generated for three sample sizes (500, 1000, 2000 participants), test lengths of (30, 40, 50 items), and response alternatives (three, four, and five alternatives per item). The experimental design simulates the test conditions outlined in the following Table (1):

Table (1) Summary of Experimental Conditions

Condition	Test Length	Number of Alternatives	Sample Size
1	30	3	500
2	30	3	1000
3	30	3	2000
4	30	4	500
5	30	4	1000
6	30	4	2000
7	30	5	500
8	30	5	1000
9	30	5	2000
10	40	3	500
11	40	3	1000
12	40	3	2000
13	40	4	500
14	40	4	1000
15	40	4	2000
16	40	5	500
17	40	5	1000
18	40	5	2000
19	50	3	500
20	50	3	1000
21	50	3	2000
22	50	4	500
23	50	4	1000
24	50	4	2000
25	50	5	500
26	50	5	1000
27	50	5	2000
Total Sample Size			**32000**

Table (1) above illustrates all the experimental conditions for the test. Virtual responses were generated to simulate the likelihood of responses of virtual individuals to the generated tests. This resulted in 27 response matrices, with a total of 1080 items and a total of 32,000 participants.

Test Design and Response Alternatives:

Tests were created with lengths of 30, 40, and 50 items, where the item characteristics for the 30-item test were identical to those of the 40-item and 50-item tests. Additionally, response alternatives were used for each item in the test models with three, four, and five response alternatives, for three sample sizes (500, 1000, 2000 participants). This design was informed by a review of relevant previous studies.

The individuals were generated with a normally distributed ability, with a mean of 0 and a standard deviation of 1 for each test condition. The item discrimination parameters were generated using a uniform distribution with an initial value of 0.5 and a maximum value of 1.5, following a logarithmic normal distribution. The item difficulty parameters were generated similarly to the ability parameters, following a normal distribution with a mean of 0 and a standard deviation of 1. The guessing parameter was generated according to the experimental conditions, using a uniform distribution with initial and maximum values of (0.33, 0.25, 0.2) when the number of alternatives was three, four, and five, respectively.

Statistical Treatment of Data:

To answer the research questions, the researchers determined the statistical methods and software in light of the study's problem and objectives, as follows:

- Calculation of some descriptive statistics, including mean values, standard deviations, factor analysis, analysis of variance, and the Scheffé test for post hoc comparisons using the **SPSS** software.
- Analysis of individual responses to estimate individual abilities and item parameters according to Item Response Theory (IRT) using the software programs **R[1], Bilog-Mg3[2], and LDID**[3].

Study Results and Interpretation

This section presents the study results, discusses, and interprets them based on the six research questions, which primarily aim to examine the effect of test length (30, 40, 50

[1] https://www.r-project.org/
[2] https://bilog-mg-3-for-windows-shared-components.software.informer.com/
[3] shkim@uga.edu

items), the number of response alternatives (three, four, five), and sample size (500, 1000, 2000) on the estimation of individual abilities and item parameters.

Below are the results obtained:

Verification of Item Response Theory Assumptions and Fit:

The assumptions of Item Response Theory (IRT) were verified, including:

1. **Unidimensionality Check**:

   An exploratory factor analysis was conducted on the simulated responses, and the ratio of the first eigenvalue to the second eigenvalue was calculated. It is expected that this value should exceed 2.

Table (2): Exploratory Factor Analysis to Verify Unidimensionality

Eigenvalue	Eigenvalue Value	First Eigenvalue / Second Eigenvalue Ratio	Criterion
First	7.134	5.351	> 2
Second	1.333		

The eigenvalues were also graphically represented using a **scree plot** to verify unidimensionality by observing the curvature of the **scree plot** as shown in **Figure 7**.

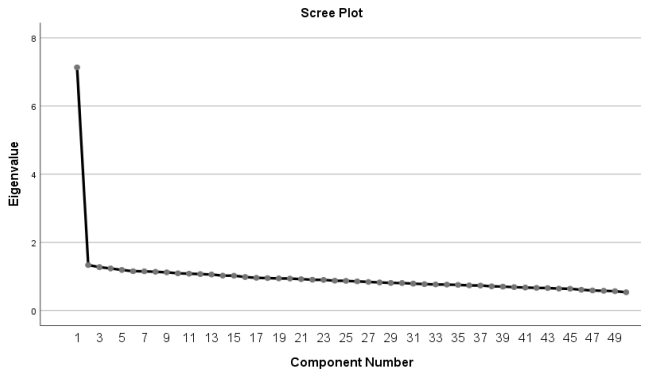
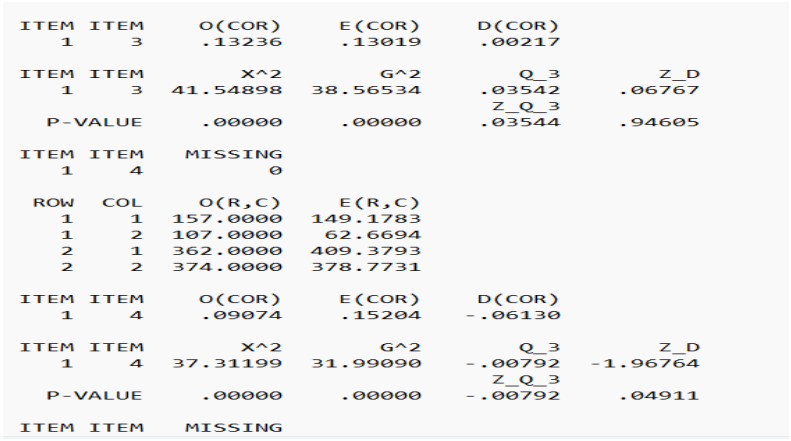


Figure 7: Graphical Representation of Eigenvalues to Verify Unidimensionality.

Second: Verification of Local Independence Assumption:

Software (Bilog-MG3) and (R) were used to analyze the simulated individual responses, and then the specialized software (LDID) was used for local independence. The correlation coefficients between the residuals and some statistical indicators of local independence, such as **X^2**, **G^2**, **ZQ**, and **ZQ3**, were computed. The data confirmed this assumption, as the ratio of independent pairs was at least 10 times greater than the

ratio of dependent pairs. The following figure shows a part of the software output for one of the experimental conditions.



```
ITEM ITEM     O(COR)     E(COR)     D(COR)
   1    3     .13236     .13019     .00217

ITEM ITEM        X^2        G^2        Q_3        Z_D
   1    3  41.54898  38.56534     .03542     .06767
                                   Z_Q_3
  P-VALUE     .00000     .00000     .03544     .94605

ITEM ITEM    MISSING
   1    4          0

 ROW  COL     O(R,C)     E(R,C)
   1    1  157.0000  149.1783
   1    2  107.0000   62.6694
   2    1  362.0000  409.3793
   2    2  374.0000  378.7731

ITEM ITEM     O(COR)     E(COR)     D(COR)
   1    4     .09074     .15204    -.06130

ITEM ITEM        X^2        G^2        Q_3        Z_D
   1    4  37.31199  31.99090    -.00792   -1.96764
                                   Z_Q_3
  P-VALUE     .00000     .00000    -.00792     .04911

ITEM ITEM    MISSING
```

Figure 8: Local Independence Indicators from LDID Software Outputs.

Third: Individual Fit to the Model: The responses were analyzed using **BilogMG3** software, and based on the **marginal likelihood index**, it was found that all individuals conformed to the model, and their number remained the same as during the generation phase.

Fourth: Item Fit to the Model: The **Chi-squared index** was used to assess the fit of items to the model, based on outputs from **BilogMG3** software. It was found that **81 items** were not a good fit to the model from a total of **1080 items**. These items were excluded, and the analysis was conducted again on the **999 items** that fit the model.

Fifth: Re-analysis of the Responses for Items that Fit the Model: The analysis was re-conducted on the responses to the items that fit the model to answer the research questions.

Answering the Main Research Question: To address the main research question "What is the effect of varying test length, number of response alternatives, and sample size on the accuracy of the ability parameter (θ) for individuals and item parameters according to the three-parameter model?", the following sub-questions were answered:

Results related to the first question:

"Is there an effect of test length on the ability parameter (θ) for individuals according to the three-parameter model?"

To answer this question, the responses of the simulated individuals were analyzed based on the generated tests designed to simulate the study's conditions. The ability of each individual in the sample was estimated according to the three-parameter model, and the accuracy of the **θ parameter** was measured by the standard error of the

estimation. Additionally, the mean **standard error (SE)** of the **θ parameter** estimation was calculated under the simulation conditions, specifically with the three levels of test length (**30**, **40**, and **50** items). These results are shown in **Table 3**.

Table 3: Accuracy of Theta (θ) Parameter by Test Length

Test Length	Sample Size	Mean Value	Standard Deviation
30	12000	0.4522	0.05312
40	12000	0.4101	0.06864
50	12000	0.3694	0.08719
Total	**36000**	**0.4106**	**0.07868**

From **Table 3**, it is clear that there are noticeable differences in the mean errors of the theta (θ) parameter's accuracy based on test length. The mean error for **standard error (SE)** of theta was **0.4522**, **0.4101**, and **0.3694** for test lengths of **30**, **40**, and **50** items, respectively. This means the accuracy of the theta parameter was higher for the **50-item** test. The standard error of the estimate acts as an inverse indicator of estimation accuracy.

Table 4: Analysis of Variance (ANOVA) in Standard Error of Theta (θ) by Test Length

Source of Variance	Sum of Squares	df	Mean Square	F-value	Sig
Between Groups	0.007	2	0.004	0.527	0.590
Within Groups	6.675	996	0.007		
Total	6.682	998			

The results of the **ANOVA** indicate that there are **no statistically significant differences** in the accuracy of the theta parameter based on test length. In other words, the accuracy of the theta parameter is not affected by the length of the test.

The result presented in **Table 3** shows the accuracy of the theta (θ) parameter with varying test lengths (30, 40, and 50 items). The **mean standard error (SE)** for the theta parameter decreases as the test length increases, with values of **0.4522**, **0.4101**, and **0.3694** for test lengths of **30**, **40**, and **50** items, respectively. This suggests that longer tests tend to yield more accurate estimates of theta. As the test length increases, the measurement becomes more precise, reflected by the decrease in the mean error.

However, when we examine **Table 4** (the ANOVA results), we observe that there are **no statistically significant differences** in the accuracy of the theta parameter between the three test lengths. The **F-value** of **0.527** and the **p-value** of **0.590** indicate that the observed differences in standard error are not significant enough to conclude that test length has an impact on the accuracy of theta estimates. This means that, despite the

trend in Table 3 suggesting that longer tests provide more accurate estimates, the statistical analysis suggests that **test length does not significantly influence the precision** of theta estimation.

This result could be interpreted in several ways. One possible explanation is that other factors, such as the number of response alternatives or sample size, may play a more crucial role in determining the accuracy of the estimates than the length of the test itself. Additionally, the study suggests that the accuracy of theta estimation might be more influenced by the **quality** and **distribution of items** in the test rather than the sheer **length** of the test. Therefore, while longer tests might intuitively seem more accurate, this particular study finds no significant effect of test length on theta estimation accuracy.

Results related to the second question, which asks, "Is there an effect of test length on item parameters according to the three-parameter model?"

To answer this question, the researchers analyzed the responses of virtual individuals to the generated tests that simulate the study conditions. Item parameters were estimated according to the three-parameter model, and the accuracy of each item was represented by the standard error of the estimation. Additionally, the mean standard error (SE) for the estimation of each item parameter was extracted for the virtual test items, which simulated the experimental conditions, specifically the test length with three levels (30, 40, and 50), as shown in **Table 5**.

Table 5: Accuracy of Item Parameters with Varying Test Lengths

Test Length		Discrimination Standard Error	Difficulty Standard Error	Guessing Standard Error
30	Mean:	0.1889	0.2538	0.0670
	Std. Dev.:	0.07912	0.11734	0.01854
	N:	228	228	228
40	Mean:	0.1869	0.2322	0.0635
	Std. Dev.:	0.07774	0.10338	0.02046
	N:	340	340	340
50	Mean:	0.1825	0.2352	0.0619
	Std. Dev.:	0.08634	0.11246	0.02211
	N:	431	Count: 431	431
Total	Mean:	0.1855	0.2384	0.0636
	Std. Dev.:	0.08183	0.11083	0.02085
	N:	999	999	999

From **Table 5**, significant differences are observed in the mean standard errors of the item parameter estimates with varying test lengths. The mean standard error for the discrimination parameter was (0.1889, 0.1869, 0.1825) for test lengths of (30, 40, 50) respectively, indicating that the discrimination parameter was more accurate when the

test length was 50, as the standard error is inversely related to the accuracy of the estimate. The mean standard error for the difficulty parameter was (0.2538, 0.2322, 0.2352) for test lengths (30, 40, 50), showing similar results. Furthermore, the mean standard error for the guessing parameter was (0.0670, 0.0635, 0.0619) for test lengths (30, 40, 50), indicating that the guessing parameter's estimate was least accurate when the test length was 50, again reflecting the inverse relationship of standard error and estimation accuracy.

Table 6: Analysis of Variance for the Standard Error of Estimates of Item Parameters by Test Length

Parameter	Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value	Sig
Discrimination	Between Groups	0.007	2	0.004	0.527	0.590
	Within Groups	6.675	996	0.007		
	Total	6.682	998			
Difficulty	Between Groups	0.072	2	0.036	2.926	0.054
	Within Groups	12.187	996	0.012		
	Total	12.259	998			
Guessing	Between Groups	0.004	2	0.002	4.541	0.011
	Within Groups	0.430	996	0.000		
	Total	0.434	998			

From the results of the analysis of variance (ANOVA), it is evident that there are no statistically significant differences in the estimates of the discrimination and difficulty parameters due to test length. This means that the estimates for these two parameters are unaffected by the test length. However, for the guessing parameter, the differences due to test length are statistically significant, indicating that the guessing parameter is influenced by the test length.

To further understand the nature of these differences, post-hoc comparisons using the Scheffé method were conducted, as shown in **Table 7**.

Table 7: Post-hoc Comparisons for Differences in Guessing Parameter Results Due to Test Length Using the Scheffé Method

Test Length	30	40	50
30	-	0.00351	0.00512*
40	0.143	-	0.00161
50	0.011	0.564	-

Note: Values above the diagonal indicate differences, and values below the diagonal indicate the significance of these differences.

From the post-hoc comparisons, it is evident that the differences in estimation accuracy favored the longer test (50 items) compared to the shorter test (30 items).

Results related to Question 3: "Is there an effect of the number of response alternatives in the test items on the ability parameter for individuals according to the three-parameter model?"

To answer this question, the responses of the virtual individuals were analyzed on the generated tests that simulate the study's conditions. The ability of each sample individual was estimated using the three-parameter model, and the standard error of estimation for the ability parameter (Theta) was calculated. The mean standard error of estimation (SE) for Theta was then computed for different levels of response alternatives (three, four, and five) as shown in ***Table 8***.

Table 8: Theta Estimation Accuracy Based on the Number of Response Alternatives in the Test Items

Number of Response Alternatives	Count	Mean	Standard Deviation
3	10669	0.4179	0.07655
4	10669	0.4133	0.07561
5	10669	0.4004	0.08263
Total	32000	0.4106	0.07868

From **Table 8**, it is clear that there are noticeable differences in the mean standard error of the Theta estimate based on the number of response alternatives in the test items. The mean standard error of estimation for Theta was (0.4179, 0.4133, 0.4004) when the number of response alternatives for the test items was (three, four, five), respectively. This indicates that the accuracy of Theta was better when the number of alternatives was greater (five), as the standard error of estimation is inversely related to the accuracy of the estimation. To determine the significance of these differences, analysis of variance (ANOVA) was conducted, as shown in **Table 9**.

Table 9: Analysis of Variance for the Standard Error of Theta Estimation Based on the Number of Response Alternatives

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	F-value	Sign
Between Groups	1.981	2	0.990	161.417	0.000
Within Groups	220.846	31997	0.006		
Total	222.827	31999			

The results of the analysis of variance indicate significant differences in the accuracy of Theta estimation attributable to the number of response alternatives, and these differences are statistically significant at the level of significance ($\alpha < 0.05$). To identify

which conditions these differences favor, post-hoc comparisons using the Scheffé method were conducted, as shown in **Table 10**.

Table 10: Post-hoc Comparisons for Differences in Theta Estimation Accuracy Based on the Number of Response Alternatives in the Test Items Using the Scheffé Method

Number of Alternatives	**3**	**4**	**5**
3	-	0.00462*	0.01753*
4	0.000	-	0.01291*
5	0.000	0.000	-

*: The values above the diagonal indicate the differences, while those below indicate the significance of the differences.

From the post-hoc comparison results, it can be concluded that the differences in the accuracy of Theta estimation were in favor of the larger number of response alternatives when comparisons were made between any of the alternatives.

Results for Question 4: "Is there an effect of the number of response alternatives in the test items on the item parameters according to the three-parameter model?"

To answer this question, responses of simulated individuals were analyzed on the generated tests that mimic the study's conditions. Item parameters were estimated according to the three-parameter model, and the accuracy of each parameter was represented by the standard error of the estimation. The mean standard error (se) for each item parameter was extracted based on the number of response alternatives in the test items, which had three levels (three, four, and five), as shown in **Table 11**.

Table 11: Accuracy of Test Item Parameters Based on the Number of Response Alternatives in the Test Items

Number of Alternatives		**Mean Standard Error of Parameter Accuracy**		
		Discrimination	Difficulty	Guessing
3	Mean:	0.1884	0.2563	0.0659
	SD:	0.08245	0.12452	0.02119
	N:	324	324	324
4	Mean:	0.1826	0.2351	0.0641
	SD:	0.08094	0.10393	0.01980
	N:	339	339	339
5	Mean:	0.1856	0.2247	0.0607
	SD:	0.08225	0.10117	0.02129
	N:	336	336	336
Total	Mean:	0.1855	0.2384	0.0636

	SD:	0.08183	0.11083	0.02085
	N:	999	999	999

From **Table 11**, it is observed that there are notable differences in the mean standard errors of parameter accuracy across different numbers of response alternatives. Specifically, the mean standard error for the discrimination parameter was (0.1884, 0.1826, 0.1856) when the number of response alternatives was (three, four, and five) respectively. For the difficulty parameter, the mean standard error was (0.2563, 0.2351, 0.2247) across the three conditions. For the guessing parameter, the mean standard error was (0.0659, 0.0641, 0.0607) when the number of response alternatives was (three, four, and five) respectively. This indicates that the guessing parameter's standard error decreased with the increase in the number of response alternatives, with the largest being when there were five alternatives. The standard error serves as an inverse indicator of estimation accuracy.

To assess the significance of these differences, an analysis of variance (ANOVA) was conducted, as shown in **Table 12**.

Table 12: Analysis of Variance (ANOVA) for the Standard Error of Parameter Estimation Based on the Number of Response Alternatives in the Test Items

Parameter	Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value	Sig
Discrimination	Between Groups	0.006	2	0.003	0.416	0.660
	Within Groups	6.676	996	0.007		
	Total	6.682	998			
Difficulty	Between Groups	0.171	2	0.085	7.031	0.001
	Within Groups	12.088	996	0.012		
	Total	12.259	998			
Guessing	Between Groups	0.005	2	0.002	5.373	0.005
	Within Groups	0.429	996	0.000		
	Total	0.434	998			

From the results of the **ANOVA**, it is observed that there are no statistically significant differences in the discrimination parameter based on the number of response alternatives in the test items, meaning that the discrimination parameter is not affected by the number of response alternatives. However, significant differences were found for the difficulty and guessing parameters, indicating that both of these parameters' accuracies are affected by the number of response alternatives.

To further understand which groups the differences are in favor of, post-hoc comparisons using the Scheffé method were conducted, as shown in **Table 13** and **Table 14**.

Table 13: Post-hoc Comparisons of Difficulty Parameter Differences Based on the Number of Response Alternatives Using Scheffé Method

Number of Alternatives	3	4	5
3	-	0.02119*	0.03161*
4	0.047	-	0.01042
5	0.001	0.470	-

The values above the diagonal indicate the differences, and the values below indicate statistical significance.

From the post-hoc comparisons, it is clear that the differences in the difficulty parameter were in favor of the groups with four and five alternatives, compared to the group with three alternatives.

Table 14: Post-hoc Comparisons of Guessing Parameter Differences Based on the Number of Response Alternatives Using Scheffé Method

Number of Alternatives	3	4	5
3	-	0.00182	0.00521*
4	0.530	-	0.00339
5	0.006	0.106	-

The values above the diagonal indicate the differences, and the values below indicate statistical significance.

From the post-hoc comparisons, it is observed that the differences in the guessing parameter were in favor of the group with five alternatives, compared to the group with three alternatives, These findings suggest that increasing the number of response alternatives can enhance the accuracy of the difficulty and guessing parameters, whereas the discrimination parameter remains unaffected by the number of alternatives.

The results related to the fifth question, which states:

"Is there an effect of sample size on the ability parameter of individuals according to the three-parameter model?"

To answer this question, the responses of virtual individuals were analyzed based on the generated tests that simulate the study's conditions. The ability of each individual in the sample was estimated according to the three-parameter model, and the accuracy of the theta parameter was extracted, represented by the standard error of estimation. The mean of the standard error of estimation (se) for the theta parameter was also calculated in light of the experimental condition's simulation, specifically for the three sample size levels (500, 1000, 2000) individuals, as shown in Table (15).

Table (15) Accuracy of Theta Parameter by Sample Size

Sample Size	Mean (SE)	Standard Deviation
500	0.3908	0.08619
1000	0.4115	0.07952
2000	0.4141	0.07615
Total	0.4106	0.07868

From Table (15), noticeable differences in the mean standard errors of accuracy for the theta parameter are observed across different sample sizes. The mean standard error of the theta parameter was (0.3908, 0.4115, 0.4141) for sample sizes of (500, 1000, 2000) individuals, respectively. This means that the accuracy of the theta parameter was greater when the sample size was smaller (500 individuals), as the standard error of estimation serves as an inverse indicator of accuracy. To determine the statistical significance of these differences, an analysis of variance was conducted, as shown in Table (16).

Table (16): Analysis of Variance in the Standard Error of Theta by Sample Size

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value	Sig
Between Groups	2.054	2	1.027	167.473	.000
Within Groups	220.772	31997	.006		
Total	222.827	31999			

The results of the analysis of variance indicate significant differences in the precision of the Theta coefficient due to sample size, with statistical significance at the significance level ($\alpha < 0.05$). To identify which groups the differences favor, post-hoc comparisons were conducted using the Scheffé method, as shown in Table (17).

Table (17): Post-hoc Comparisons of Differences in Theta Coefficient Precision by Sample Size Using Scheffé Method

Sample Size	3	4	5
500	-	-0.0207*	-0.0233*
1000	0.000	-	-0.0025*
2000	0.000	0.000	-

*: Values above the diagonal indicate the differences, while values below the diagonal indicate the significance of the differences.

The post-hoc comparisons indicate that the differences in the precision of the Theta coefficient favored the smaller sample sizes when comparisons were made between any of the groups.

The results related to the sixth question, which states, "Is there an effect of sample size on item parameters according to the three-parameter model?" To answer this question, the responses of virtual individuals to the generated tests simulating the study's conditions were analyzed. Item parameters were estimated according to the three-parameter model, and the precision of each was represented by the standard error of estimation. The mean value of the standard estimation error (SE) for the estimation of each item parameter was also calculated for the virtual test items, simulating experimental conditions, specifically sample size, which had three levels (500, 1000, 2000), as shown in Table (18).

Table (18): Item Parameter Precision by Sample Size

Number of Alternatives		**Mean Standard Error of Parameter Accuracy**		
		Discrimination	Difficulty	Guessing
3	Mean:	0.2400	0.2634	0.0693
	SD:	0.08792	0.10475	0.01829
	N:	352	352	352
4	Mean:	0.1790	0.2444	0.0644
	SD:	0.06469	0.11257	0.02011
	N:	337	337	337
5	Mean:	0.1306	0.2037	0.0562
	SD:	0.04295	0.10711	0.02220
	N:	310	310	310
Total	Mean:	0.1855	0.2384	0.0636
	SD:	0.08183	0.11083	0.02085
	N:	999	999	999

From Table (18), it is observed that there are differences in the mean standard errors of item parameter precision based on sample size. The mean standard error of the discrimination parameter was 0.2400, 0.1790, and 0.1306 for sample sizes of 500, 1000, and 2000, respectively. This indicates that the discrimination parameter was more precise when the sample size was larger (2000). The standard error of estimation is inversely related to estimation precision. Similarly, the mean standard error for the difficulty parameter was 0.2634, 0.2444, and 0.2037 for sample sizes of 500, 1000, and 2000, respectively. For the guessing parameter, the mean standard error was 0.0693, 0.0644, and 0.0562 for sample sizes of 500, 1000, and 2000, respectively. Therefore, the guessing parameter's precision improved as the sample size increased (2000). To assess the significance of these differences, an analysis of variance was conducted, as shown in Table (19).

Table (19): Analysis of Variance in Standard Error of Estimation for Each Item Parameter by Sample Size

Parameter	Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value	Sig
Discrimination	Between Groups	1.993	2	0.996	211.658	.000
	Within Groups	4.689	996	0.005		
	Total	6.682	998			
Difficulty	Between Groups	0.605	2	0.302	25.852	.000
	Within Groups	11.654	996	0.012		
	Total	12.259	998			
Guessing	Between Groups	0.028	2	0.014	34.900	.000
	Within Groups	0.406	996	0.000		
	Total	0.434	998			

The results of the analysis of variance indicate statistically significant differences in the precision of each item parameter due to sample size. This suggests that the precision of the item parameter estimates is affected by sample size. To identify which groups the differences favor, post-hoc comparisons were conducted using the Scheffé method, as shown in Tables (20), (21), and (22).

Table (20): Post-hoc Comparisons of Differences in Discrimination Parameter Results by Sample Size Using Scheffé Method

Sample Size	500	1000	2000
500	-	0.06102*	0.10936*
1000	0.000	-	0.04834*
2000	0.000	0.000	-

*: Values above the diagonal indicate differences, and values below the diagonal indicate the significance of the differences.

The post-hoc comparisons reveal that the differences in the discrimination parameter results favored the larger sample size when compared to the smaller sample size for each pair.

Table (21): Post-hoc Comparisons of Differences in Difficulty Parameter Results by Sample Size Using Scheffé Method

Sample Size	500	1000	2000
500	-	0.01900	0.05968*
1000	0.071	-	0.04068*
2000	0.000	0.000	-

*: Values above the diagonal indicate differences, and values below the diagonal indicate the significance of the differences.

The post-hoc comparisons show that the differences in the difficulty parameter results favored the sample size of 2000 when compared to the smaller sample sizes (500 and 1000).

Table (22): Post-hoc Comparisons of Differences in Guessing Parameter Results by Sample Size Using Scheffé Method

Sample Size	500	1000	2000
500	-	0.00485*	0.01305*
1000	0.007	-	0.00820*
2000	0.000	0.000	-

*: Values above the diagonal indicate differences, and values below the diagonal indicate the significance of the differences.

The post-hoc comparisons indicate that the differences in the guessing parameter results favored the larger sample size when compared to the smaller sample size for each pair.

Summary of Study Results, Recommendations, and Suggested Research

- **Interpretation of Results Related to the First Question: "Is there an effect of test length on the ability parameter for individuals according to the three-parameter model?"**

  The results of this question indicate that there are no differences in the precision of the Theta coefficient attributed to test length. This outcome can be explained by the mechanism of the Theta coefficient, as it relies on the marginal conditional probability, which compares the estimated score using the model with the probability of response to the items, regardless of the total score on the test. According to previous studies, such as Lord's, the sample sizes used in this study are sufficient for precise Theta coefficient estimation. This, in turn, nullifies any differences in estimation precision due to varying test lengths. These findings contradict those of(Fernandes et al., 2023; Shibata & Uto, 2022), who suggested that the precision of the Theta coefficient for individuals increases with the number of test items.

- **Interpretation of Results Related to the Second Question: "Is there an effect of test length on item parameters according to the three-parameter model?"**

  The results of this question indicate that the length of the test only affects the guessing parameter, favoring longer tests. This can be explained by the lower

ability levels simulated by the guessing parameter, where an increase in the number of test items provides additional information that can more accurately estimate the guessing parameter. In contrast, the difficulty and discrimination parameters do not show similar behavior. The difficulty parameter requires items of varying levels of difficulty, which supports the observation of the test information function, the inverse of the estimation error. The test information function equations reveal that the guessing parameter has a different relationship than the difficulty and discrimination parameters and occupies distinct positions in the test information function. This result aligns with the findings of (Silva et al., 2023; Siraji et al., 2023; Wang et al., 2022).

- **Interpretation of Results Related to the Third Question: "Is there an effect of the number of response alternatives for test items on the ability parameter for individuals according to the three-parameter model?"** The results of this question indicate that the number of alternatives affects the precision of the Theta coefficient, favoring tests with a greater number of alternatives. This can be explained by the ability of response alternatives to better determine an individual's ability. The number of alternatives is inversely related to random correct responses. As the number of alternatives increases, the number of randomly correct responses decreases, enabling the model to more accurately determine the individual's ability. Figure (9) illustrates this relationship.

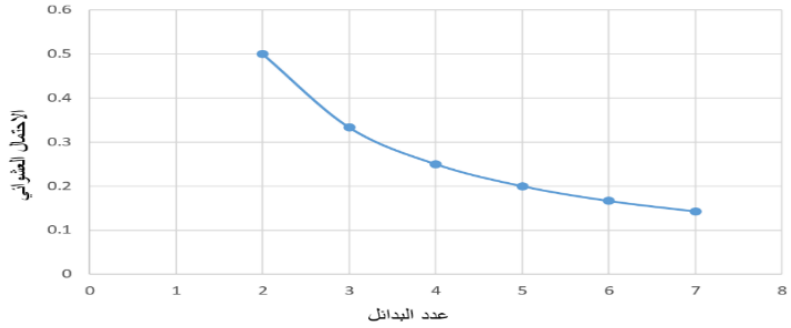


Figure (9): The Relationship Between the Number of Alternatives and the Probability of Random Correct Responses

It can be observed from Figure (9) that the level of random probability decreases as the number of alternatives increases, which enables a more accurate estimation of the individual's ability.

- Interpretation of Results Related to the Fourth Question: "Is there an effect of the number of response alternatives for test items on the item parameters according to the three-parameter model?"

The study found that the number of alternatives affects the precision of the difficulty and guessing parameters only, and this effect is positive. An increase in the number of alternatives improves the accuracy of both the difficulty and guessing coefficients. This can be explained by the increased ability of the model to predict as the number of alternatives increases, which is sometimes referred to as the model's probabilistic space. Figure (10) illustrates this.

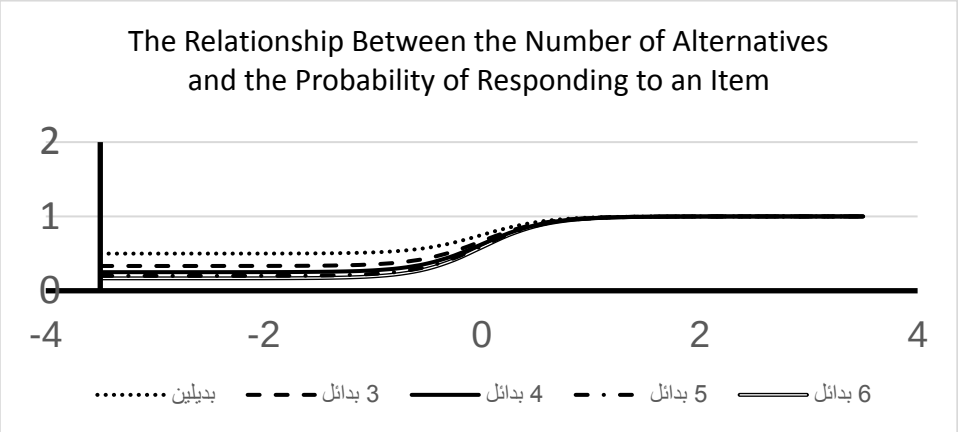


Figure (10): The Relationship Between the Number of Alternatives and the Probability of Responding to an Item

It can be observed from the figure that the number of response alternatives plays a crucial role in the guessing parameter, which in turn affects the probabilistic range of the model. In items with two alternatives, the probabilistic range is from 0.5 to 1, indicating a range of 0.5, which is considered low. Therefore, the three-parameter model is able to make predictions with a lower degree of certainty, which impacts its ability to estimate accurately. On the other hand, for items with five alternatives, the probabilistic range spans from 0.2 to 1, representing a range of 0.8, which is relatively high. As a result, the three-parameter model is more capable of making highly accurate predictions, leading to better estimation precision. This finding is consistent with the study by(Charamba et al., 2023; Cotter et al., 2023; Effatpanah & Baghaei, 2023), which showed that a test with three alternatives was the most effective in multiple-choice testing.

- **Interpretation of Results Related to the Fifth Question: "Is there an effect of sample size on the ability parameter for individuals according to the three-parameter model?"**

  The study found that sample size does affect the precision of the Theta coefficient, with smaller sample sizes being more beneficial. This result aligns with previous studies that identified optimal sample sizes based on test length. Specifically, Lord’s reference indicated that a sample size of 1000 is suitable

for a test with 50 items. In this study, the average test length was fewer than 40 items, suggesting that the ideal sample size should be smaller than 1000. The findings showed that a sample size of 500 yielded the best estimation accuracy compared to sample sizes of 1000 and 2000. Furthermore, an important consideration is that estimation relies on response patterns. A test with a large number of items requires a sufficient number of individuals to cover the possible response patterns for effective estimation. However, increasing the sample size may lead to repeated or conflicting response patterns, which can hinder optimal estimation. Increasing the sample size while keeping the number of items constant reduces the model's effectiveness for the Theta coefficient, but enhances its effectiveness for item parameters.

- **Interpretation of Results Related to the Sixth Question: "Is there an effect of sample size on item parameters according to the three-parameter model?"**
  The study found that sample size affects the precision of item parameters, favoring larger sample sizes. This outcome can be explained by the model's reliance on response patterns. With a sufficient number of items, a larger sample size is needed to cover the potential response patterns, enabling the model to estimate item parameters more effectively. As the sample size increases, the response patterns follow a more normal distribution around the item parameters, reinforcing the accuracy of the item parameters. Thus, a larger sample size with a constant number of items enhances the model's effectiveness for estimating item parameters while decreasing its effectiveness for estimating the Theta coefficient.

Recommendations

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

First: Practical Recommendations

- The test should be at least 30 items long when the goal is to estimate the Theta coefficient accurately.
- The test should consist of at least 50 items when the goal is to estimate item parameters according to the three-parameter model, with a specific focus on the guessing parameter.
- Increase the number of response alternatives to the maximum possible extent to ensure greater accuracy in estimating the Theta coefficient and item parameters.

- A suitable sample size should be selected that aligns with the length of the test when estimating the Theta coefficient. A larger sample size should be used when the goal is to estimate item parameters.
- To enhance the quality of Theta coefficient estimates, the sample size should not exceed the point at which the model can no longer accurately account for response patterns, especially for tests with shorter lengths.
- When constructing a test, researchers should carefully balance between test length and sample size to ensure optimal precision in both item parameter estimation and Theta coefficient accuracy.
- In situations where multiple item types are used (e.g., multiple-choice, true/false), it's important to adjust the test design to accommodate the unique characteristics of each item type for more reliable estimates.

Second: Theoretical Recommendations

- The researchers recommend replicating the study using multiple response item response theory (IRT) models to explore the effects of different response patterns on the accuracy of estimates.
- The study should be extended to compare the three-parameter model with other widely used models such as the unidimensional and bidimensional IRT models to assess the robustness of findings across various test designs.
- Further investigation should be carried out into the effect of varying item difficulty levels on Theta coefficient estimation, particularly in tests with a large number of items.
- Future research could focus on refining the estimation of item parameters in tests with small sample sizes to understand the boundary conditions where accurate estimation is still achievable.
- Researchers should also explore the influence of test-taking strategies (e.g., guessing behaviors) on the reliability of estimates within the three-parameter model framework.

Future Research Proposals:

1. "Investigating the Impact of Test Length on the Accuracy of Theta Coefficient Estimation in Educational Assessments."
2. "The Effect of Sample Size Variability on Item Parameter Estimation in Three-Parameter Models."

3. "Exploring the Relationship Between Response Alternatives and Item Parameter Accuracy in Item Response Theory Models."
4. "Comparing the Performance of Three-Parameter and Two-Parameter Item Response Models in Large-Scale Educational Testing."
5. "Examining the Role of Item Difficulty Variations in Enhancing the Accuracy of Theta Coefficient Estimates."
6. "Evaluating the Effectiveness of Multiple-Response Item Response Theory Models in Predicting Student Ability."
7. "An Investigation of Optimal Sample Size for Accurate Theta Coefficient Estimation in Shortened Tests."
8. "The Influence of Test Format (Multiple Choice vs. True/False) on Item Parameter Estimation in the Three-Parameter Model."
9. "Assessing the Impact of Test-Taking Behaviors on the Accuracy of Item Parameter and Theta Coefficient Estimates."
10. "A Comparative Study of Unidimensional and Bidimensional Models in Estimating Item Parameters and Theta Coefficients."
11. "Exploring the Limitations and Boundaries of Item Response Theory Models in Tests with Small Sample Sizes."
12. "The Role of Test Design and Response Distribution in Improving the Accuracy of Ability Estimates in Educational Assessment."
13. "Effects of Varying Response Option Quantities on the Precision of Estimating Item Parameters in Large-Scale Tests."
14. "Developing a Framework for Optimizing Test Design Based on Response Pattern Analysis and Model Fit."
15. "An Investigation into the Use of Simulation-Based Approaches for Enhancing the Calibration of Item Response Theory Models."

Reference:

Adetutu, O., & Lawal, H. (2022). APPLICATIONS OF ITEM RESPONSE THEORY MODELS TO ASSESS ITEM PROPERTIES AND STUDENTS'ABILITIES IN DICHOTOMOUS RESPONSES ITEMS. *Open Journal of Educational …*. https://www.openjournalsnigeria.org.ng/journals/index.php/ojed/article/view/304

Ali, A. (2022). *The measurement of subjective well-being: item response theory, classical test theory, and multidimensional item response theory*. gupea.ub.gu.se. https://gupea.ub.gu.se/handle/2077/71426

Ali, A., & Istiyono, E. (2022). An analysis of item response theory using program R. *Al-Jabar: Jurnal Pendidikan Matematika*. https://ejournal.radenintan.ac.id/index.php/al-jabar/article/view/11252

Apriyani, D., Susanto, H., & ... (2023). Analysis of Pre-Olympic Middle School Mathematics Test Instruments Based on Item Response Theory. *AlphaMath: Journal of ...*. https://jurnalnasional.ump.ac.id/index.php/alphamath/article/view/18021

Ayanwale, M., Amusa, J., Oladejo, A., & Ayedun, F. (2024). *Multidimensional item response theory calibration of dichotomous response structure using R language for statistical computing*. Springer. https://doi.org/10.1007/s10780-024-09517-y

Ayasse, N., & Coon, C. (2024). Investigating item response theory model performance in the context of evaluating clinical outcome assessments in clinical trials. *Quality of Life Research*. https://doi.org/10.1007/s11136-024-03873-z

Aybek, E. (2023). The relation of item difficulty between classical test theory and item response theory: Computerized adaptive test perspective. *Journal of Measurement and Evaluation in Education ...*. https://dergipark.org.tr/en/pub/epod/issue/78540/1209284

Baghaei, P., & Effatpanah, F. (2024). *Nonparametric kernel smoothing item response theory analysis of Likert items*. mdpi.com. https://www.mdpi.com/2624-8611/6/1/15

Bjorner, J., Terluin, B., Trigg, A., Hu, J., Brady, K., & ... (2023). *Establishing thresholds for meaningful within-individual change using longitudinal item response theory*. Springer. https://doi.org/10.1007/s11136-022-03172-5

Brucato, M., Frick, A., Pichelmann, S., & ... (2023). Measuring spatial perspective taking: Analysis of four measures using item response theory. *Topics in Cognitive ...*. https://doi.org/10.1111/tops.12597

Cai, L., Chung, S., & Lee, T. (2023). *Incremental model fit assessment in the case of categorical data: Tucker–Lewis index for item response theory modeling*. Springer. https://doi.org/10.1007/s11121-021-01253-4

Charamba, V., Kazembe, L., & Nickanor, N. (2023). *Application of item response theory modelling to measure an aggregate food security access score*. Springer. https://doi.org/10.1007/s12571-023-01388-y

Cook, R., & Wind, S. (2024). Item response theory: A modern measurement approach to reliability and precision for counseling researchers. *Measurement and Evaluation in Counseling ...*. https://doi.org/10.1080/07481756.2023.2301284

Cotter, K., Chen, D., Christensen, A., & ... (2023). Measuring art knowledge: Item response theory and differential item functioning analysis of the Aesthetic Fluency Scale. *... , Creativity, and the ...*. https://psycnet.apa.org/record/2021-45481-001

Doğan, Ö., & Atar, B. (2024). Comparing differential item functioning based on multilevel mixture item response theory, mixture item response theory and manifest groups. *Journal of Measurement and Evaluation in ...*. https://dergipark.org.tr/en/pub/epod/article/1457880

Donaldson, S., Donaldson, S., McQuaid, M., & ... (2023). *The PERMA+ 4 short scale: A cross-cultural empirical validation using item response theory*. Springer. https://doi.org/10.1007/s41042-023-00110-9

Effatpanah, F., & Baghaei, P. (2023). Kernel Smoothing Item Response Theory in R: A Didactic. *Practical Assessment, Research &Evaluation*. https://eric.ed.gov/?id=EJ1392871

Fernandes, R., Rocha, T., Coelho, J., & Andrade, D. (2023). *... of a measurement instrument to evaluate integrated management systems and differences in perception: an approach to item response theory and the quality ...*. SciELO Brasil. https://www.scielo.br/j/prod/a/MStKMyMdd8HFPg9bL5m5HkH/

Frick, S., Krivosija, A., & Munteanu, A. (2024). Scalable learning of item response theory models. *... Conference on Artificial ...*. https://proceedings.mlr.press/v238/frick24a.html

Gao, S., Ma, X., Tsui, H., Wang, J., & Zhang, X. (2024). *Item response theory analysis of the Chinese version compulsive shopping scale*. Elsevier. https://www.sciencedirect.com/science/article/pii/S0010440X24000865

Garcia, D., Kazemitabar, M., & Asgarabad, M. (2023). *The 18-item Swedish version of Ryff's psychological wellbeing scale: psychometric properties based on classical test theory and item response theory*. frontiersin.org. https://doi.org/10.3389/fpsyg.2023.1208300

Gershon, S., Anghel, E., & Alexandron, G. (2024). An evaluation of assessment stability in a massive open online course using item response theory. *Education and Information ...*. https://doi.org/10.1007/s10639-023-11925-z

Gewily, M., Plan, E., Yousefi, E., König, F., & ... (2024). Quantitative comparisons of progressive supranuclear palsy rating scale versions using item response theory. *Movement ...*. https://doi.org/10.1002/mds.30001

Gibbons, R., Lauderdale, D., Wilson, R., & ... (2024). Adaptive measurement of cognitive function based on multidimensional item response theory. *... Research &Clinical ...*. https://doi.org/10.1002/trc2.70018

Gikaro, J., Zi-Yan, Z., Hui-Hui, S., & ... (2024). *Simplified functioning assessment for low back pain: ICF-based item response theory modelling*. pmc.ncbi.nlm.nih.gov. https://pmc.ncbi.nlm.nih.gov/articles/PMC10792692/

Gilbert, J., Himmelsbach, Z., Soland, J., Joshi, M., & ... (2024). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *arXiv preprint arXiv ...*. https://arxiv.org/abs/2405.00161

Gilbert, J., Miratrix, L., Joshi, M., & ... (2025). Disentangling person-dependent and item-dependent causal effects: applications of item response theory to the estimation of treatment effect heterogeneity. *... of Educational and ...*. https://doi.org/10.3102/10769986241240085

Guo, Z., Wang, D., Cai, Y., & Tu, D. (2024). An Item Response Theory Model for incorporating response times in forced-choice measures. *Educational and Psychological ...*. https://doi.org/10.1177/00131644231171193

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459.

Hanzlová, R., & Lynn, P. (2023). *Item response theory-based psychometric analysis of the Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS) among adolescents in the UK*. Springer. https://doi.org/10.1186/s12955-023-02192-0

Harrison, C., Plessen, C., Liegl, G., Rodrigues, J., & ... (2023). *Item response theory assumptions were adequately met by the Oxford hip and knee scores*. Elsevier. https://www.sciencedirect.com/science/article/pii/S0895435623000938

Howe, M., Miller, S., Tran, S., Buscemi, J., & ... (2024). Examining the psychometric properties of the CEFIS-AYA using item response theory. *Journal of Pediatric …*. https://academic.oup.com/jpepsy/article-abstract/49/12/856/7817889

Hu, T., & Valdivia, D. (2024). *Assessing the Psychometric Properties of Quality Experience in Undergraduate Research Using Item Response Theory*. Springer. https://doi.org/10.1007/s11162-024-09814-6

Huang, H. (2023). Modeling rating order effects under item response theory models for rater-mediated assessments. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216231174566

Huang, J., Shu, T., Dong, Y., & Zhu, D. (2023). … a self-assessment scale for Chinese college English-major students' feedback knowledge repertoire in EFL academic writing: Item response theory and factor analysis …. *Assessing Writing*. https://www.sciencedirect.com/science/article/pii/S1075293523000247

Jiang, S., Xiao, J., & Wang, C. (2023). *On-the-fly parameter estimation based on item response theory in item-based adaptive learning systems*. Springer. https://doi.org/10.3758/s13428-022-01953-x

Jones, S., Ton, M., Malen, R., Newcomb, P., & ... (2024). Item response theory analysis of benefits and harms of cannabis use in cancer survivors. *JNCI …*. https://academic.oup.com/jncimono/article-abstract/2024/66/275/7728490

Kawakubo, S., Kamata, T., Arata, S., Murakami, S., & ... (2024). *Item response theory in building environment engineering: A novel approach to identifying key residential environment items*. Elsevier. https://www.sciencedirect.com/science/article/pii/S0360132324006061

Khatri, C., Harrison, C., MacDonald, D., Clement, N., & ... (2024). *Item response theory validation of the Oxford knee score and Activity and Participation Questionnaire: a step toward a common metric*. Elsevier. https://www.sciencedirect.com/science/article/pii/S0895435624002713

Kiliç, A., Koyuncu, I., & Uysal, I. (2023). Scale Development Based on Item Response Theory: A Systematic Review. ... *Journal of Psychology and Educational Studies*. https://eric.ed.gov/?id=EJ1378903

Liang, M., Yin, M., Guo, B., Pan, Y., Zhong, T., Wu, J., & ... (2024). Validation of the Barthel Index in Chinese nursing home residents: an item response theory analysis. *Frontiers in …*. https://doi.org/10.3389/fpsyg.2024.1352878

Ma, C., Ouyang, J., Wang, C., & Xu, G. (2024). A note on improving variational estimation for multidimensional item response theory. *Psychometrika*. https://www.cambridge.org/core/journals/psychometrika/article/note-on-improving-variational-estimation-for-multidimensional-item-response-theory/B28AC3707E22D13224E27DC7539ACAD1

Mangold, F. (2024). Improving media trust research through better measurement: An item response theory perspective. *Journal of Trust Research*. https://doi.org/10.1080/21515581.2023.2229791

Raykov, T. (2023). Item response theory and modeling with Stata. *Measurement: Interdisciplinary Research and …*. https://doi.org/10.1080/15366367.2022.2133528

Reise, S., & Moore, T. (2023). *Item response theory*. psycnet.apa.org. https://psycnet.apa.org/record/2023-76874-037

Santos, K., Arantes, F., Leite, M., & ... (2023). Supplier's performance evaluation in supply chains using item response theory. ... *Journal of Services …*. https://doi.org/10.1504/ijsom.2023.132858

Shi, H., Ren, Y., Xian, J., Ding, H., Liu, Y., & Wan, C. (2024). *Item analysis on the quality of life scale for anxiety disorders QLICD-AD (V2. 0) based on classical test theory and item response theory*. Springer. https://doi.org/10.1186/s12991-024-00504-2

Shibata, T., & Uto, M. (2022). Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. *... of the 29th International Conference on ...*. https://aclanthology.org/2022.coling-1.257/

Shim, H., Bonifay, W., & Wiedermann, W. (2023). Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01824-5

Silva, J. d., Silva, J. d., Bispo, L., & ... (2023). *Construction of a musculoskeletal discomfort scale for the lower limbs of workers: An analysis using the multigroup item response theory*. mdpi.com. https://www.mdpi.com/1660-4601/20/7/5307

Silveira, V., França, A., Campelo, C., & ... (2023). *Proposition of an Energy Intake Estimating Scale through Item Response Theory*. mdpi.com. https://www.mdpi.com/2072-6643/15/21/4511

Silvia, P. (2022). The self-reflection and insight scale: Applying item response theory to craft an efficient short form. *Current Psychology*. https://doi.org/10.1007/s12144-020-01299-7

Sinharay, S., & Monroe, S. (2024). Assessment of fit of item response theory models: A critical review of the status quo and some future directions. *British Journal of Mathematical and ...*. https://doi.org/10.1111/bmsp.12378

Siraji, M., & Haque, S. (2022). Psychometric evaluation of the Bangla-Translated Rotter's Internal-External Scale through classical test theory and item response theory. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2022.1023856

Siraji, M., Jahan, N., & Borak, Z. (2023). Validation of the Bangla Communication Scale among Bangladeshi adolescents: A classical test theory and item response theory approach. *Asian Journal of Psychiatry*. https://www.sciencedirect.com/science/article/pii/S1876201823001417

Soland, J. (2022). Evidence that selecting an appropriate item response theory–based approach to scoring surveys can help avoid biased treatment effect estimates. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211007551

Stavropoulos, V., Footitt, T., Zarate, D., & ... (2022). The online flow questionnaire: An item response theory examination. *... , Behavior, and Social ...*. https://doi.org/10.1089/cyber.2022.0031

Tang, X., Schalet, B., Peipert, J., & Cella, D. (2023). *... estimation of significant individual changes assessed by patient-reported outcome measures? Comparing classical test theory versus item response theory*. Elsevier. https://www.sciencedirect.com/science/article/pii/S1098301523030176

Toledano-Toledano, F., Jiménez, S., Rubia, J. M. d. l., & ... (2023). *Positive mental health scale (PMHS) in parents of children with cancer: a psychometric evaluation using item response theory*. mdpi.com. https://www.mdpi.com/2072-6694/15/10/2744

Tomikawa, Y., Suzuki, A., & Uto, M. (2024). Adaptive Question–Answer Generation with Difficulty Control Using Item Response Theory and Pre-trained Transformer Models. *IEEE Transactions on Learning ...*. https://ieeexplore.ieee.org/abstract/document/10742557/

Toraman, Ç., Karadağ, E., & Polat, M. (2022). *Validity and reliability evidence for the scale of distance education satisfaction of medical students based on item response theory (IRT)*. Springer. https://doi.org/10.1186/s12909-022-03153-9

Uto, M., Aomi, I., Tsutsumi, E., & ... (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on ...*. https://ieeexplore.ieee.org/abstract/document/10061235/

Uto, M., Tomikawa, Y., & Suzuki, A. (2023). Difficulty-controllable neural question generation for reading comprehension using item response theory. *... of the 18th workshop on innovative ...*. https://aclanthology.org/2023.bea-1.10/

Vaganian, L., Boecker, M., Bussmann, S., Kusch, M., & ... (2022). *Psychometric evaluation of the Positive Mental Health (PMH) scale using item response theory*. Springer. https://doi.org/10.1186/s12888-022-04162-0

Veldkamp, K., Grasman, R., & ... (2025). Handling missing data in variational autoencoder based item response theory. *British Journal of ...*. https://doi.org/10.1111/bmsp.12363

Wang, L., Wu, Y., Lin, Y., Wang, L., Zeng, Z., & ... (2022). Reliability and validity of the Pittsburgh Sleep Quality Index among frontline COVID-19 health care workers using classical test theory and item response theory. *Journal of clinical ...*. https://doi.org/10.5664/jcsm.9658

Wang, X., Cai, Y., & Tu, D. (2023). The application of item response theory in developing and validating a shortened version of the Rotterdam Emotional Intelligence Scale. *Current Psychology*. https://doi.org/10.1007/s12144-022-03329-y

Wolcott, M., Olsen, A., & Augustine, J. (2022). Item response theory in high-stakes pharmacy assessments. *Currents in Pharmacy Teaching and ...*. https://www.sciencedirect.com/science/article/pii/S1877129722001927

Wortham, S., Borowiec, K., & Kim, D. (2023). Measuring Faculty Engagement in Online Formative or Whole-Person Education: A Revised Instrument and Item Response Theory Model. *Online Learning*. https://eric.ed.gov/?id=EJ1412305

Yiğiter, M., & Boduroğlu, E. (2024). Item Response Theory Assumptions: A Comprehensive Review of Studies with Document Analysis. *... Journal of Educational Studies and Policy*. https://dergipark.org.tr/en/pub/ijesp/issue/90802/1659816

Young, R., Courtney, E., Kah, A., & ... (2025). Content and Item Response Theory Analysis of ChatGPT-4-Generated Multiple-Choice Items. *Teaching of ...*. https://doi.org/10.1177/00986283241311220

Zhang, J., Valdivia, D. S., & ... (2022). An item response theory analysis of residents' perceived sporting event impacts. *Journal of Global Sport ...*. https://doi.org/10.1080/24704067.2020.1731701

Zhao, N., Mason, J., Blum, A., Kim, E., & ... (2023). Using item-response theory to improve interpretation of the Trans Woman Voice Questionnaire. *The ...*. https://doi.org/10.1002/lary.30360

Zhong, S., Zhou, Y., Zhumajiang, W., Feng, L., Gu, J., & ... (2023). *A psychometric evaluation of Chinese chronic hepatitis B virus infection-related stigma scale using classical test theory and item response theory*. frontiersin.org. https://doi.org/10.3389/fpsyg.2023.1035071