Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.00715

RESEARCH ARTICLE

A Critical Analysis of the Use of Classical Test Theory (CTT) in Psychological Testing: A Comparison with Item Response Theory (IRT)

Samy Meguellati[1*], Adaika Samia[2], Ahmed Ferhat[3], Ahmed Djelloul[4], Zuoari Ahmed khalifa[5]

[1] University of Constantine 2(Algeria)

[2,3,4,5] University of El Oued (Algeria)

[3,4] Social Development and Community Service Laboratory, University of El Oued (Algeria)

ARTICLE INFO

Received: Aug 11, 2024

Accepted: Oct 23, 2024

Keywords

Classical Test Theory

Item Response Theory

Reliability

Validity

Psychological Assessment

Psychometrics

Educational Evaluation

***Corresponding Author:**

samy.meguellati@univ-contantine

ABSTRACT

This article reviews Classical Test Theory (CTT), one of the earliest and simplest models used in constructing and evaluating psychological and educational tests. CTT, which emerged in the early 20th century, focuses on the relationship between true and observed scores to explain performance. Despite its effectiveness, CTT faces criticism, mainly due to its reliance on the sample and its assumption of equal item weights, limiting its accuracy. Conversely, Item Response Theory (IRT) offers a more accurate and comprehensive alternative by analyzing individuals' responses to different items independently, providing more precise estimates of abilities. The article discusses the advantages and disadvantages of both theories and highlights practical applications for each, along with a detailed comparison. It also touches on future directions in psychometric assessments, emphasizing hybrid models and innovative techniques like game-based assessments and simulations.

1. INTRODUCTION

The Classical Test Theory (CTT) is considered one of the most important and simplest theories used in constructing and evaluating psychological and educational tests. This theory emerged in the early 20th century as a theoretical framework aimed at explaining test results and assessments in a straightforward manner, relying on two main foundations: the true score and the observed score. CTT focuses on attempting to explain individual performance based on the differences between these two scores. It assumes that the individual's true score represents their actual level of abilities or knowledge, while the observed score may contain a portion of errors resulting from random factors that could affect the measurement, such as the individual's surrounding conditions during the test or difficulties in understanding the instructions.

Although CTT is effective and easy to use, it has several limitations that affect its accuracy and effectiveness, the most prominent being its reliance on the sample used in the test and the nature of the items. These limitations have led to the emergence of more advanced models, such as the **Item Response Theory (IRT)**, which was developed to overcome these limitations and provide a more accurate model for measuring psychological and educational abilities.

The classical test theory (CTT)

The Classical Test Theory (CTT) is based on the idea that all individuals are subjected to the same assessments and items, and their performance is interpreted based on observed results, which are analyzed using relatively simple mathematical equations. In contrast, Item Response Theory (IRT) follows a different approach, analyzing each item independently and relying on complex mathematical models to interpret individuals' responses based on their actual abilities and the difficulty of the items.

This article aims to provide a critical analysis of the use of CTT in constructing psychological tests, highlighting its advantages and disadvantages, in addition to offering a detailed comparison with IRT to clarify how the measurement of abilities and psychological traits can be improved using modern models. Practical applications of both CTT and IRT will also be reviewed, along with the challenges associated with them in psychological and educational contexts.

2. THE HISTORICAL PERSPECTIVE OF CLASSICAL TEST THEORY

The Classical Test Theory (CTT) emerged in the early 20th century and is considered one of the earliest theories developed to provide a systematic framework for constructing and evaluating psychological and educational tests. Its primary goal was to provide accurate methods for measuring individuals' mental and psychological performance. With the growing need to assess mental abilities, psychologists focused on developing reliable and objective assessment tools that could conduct evaluations free from personal biases (Muñiz, 2010, p. 57).

Alfred Binet is considered one of the pioneers in the field of psychological measurement. In the early 20th century, he developed the intelligence tests known as the "Stanford-Binet." This test was the first standardized tool for assessing children's mental abilities and served as a launching point for the development of other psychological measurement tools (Hardin & Leong, 2004, p. 25). During that period, there was an increasing need to measure intelligence and mental abilities in a way that could be generalized to different population groups.

During **World War I**, there was an urgent need to develop group tests to quickly and effectively assess the mental and military abilities of American soldiers. As a result, group tests such as the **Army Alpha Test**, which was built on the foundations of CTT, were used. This experience demonstrated the effectiveness of CTT in large-scale practical contexts that required fast and reliable assessments of large groups of individuals (Jimam et al., 2019, p. 189). These developments helped solidify CTT as a primary framework in test and assessment design in military, psychological, and educational fields.

Over time, the applications of CTT expanded to include not only intelligence tests but also personality and various cognitive ability tests. This theory played a significant role in the development of tests like the **Wechsler Adult Intelligence Scale (WAIS)**, which has become one of the most famous and widely used psychological tests for evaluating mental abilities (Muñiz, 2010, p. 58).

The Classical Test Theory (CTT) offered a simple solution based on analyzing the relationship between an individual's observed score and the assumed true score, taking into account the proportion of random error (Muñiz, 2010, p. 59). This approach was suitable for its time, particularly because of its ability to provide reliable estimates of human performance using simple mathematical tools such as the reliability coefficient.

Although CTT was a significant breakthrough in the field of psychological measurement when it emerged, rapid advancements in science and technology have revealed some important limitations. Among these limitations is its heavy reliance on the sample used in the test, as the accuracy of estimates can vary greatly depending on the characteristics of the sample. Additionally, CTT lacks flexibility in dealing with items of varying difficulty or discrimination, which led researchers to seek more advanced models, such as **Item Response Theory (IRT),** which has replaced CTT in many modern psychological and educational contexts (Hussein & Gasmalla, 2023, p. 113).

Overall, the Classical Test Theory remains an important part of the history of psychological measurement and has significantly contributed to the development of many tests and tools we use

today. However, the challenges faced by this theory have prompted researchers to pursue more accurate and comprehensive models for assessing individuals' mental and psychological performance in line with contemporary demands.

3. COMPONENTS OF CLASSICAL TEST THEORY

3.1 The true score and the observed score

According to CTT, an individual's observed score consists of the sum of the true score plus random errors (E) (Hussein & Gasmalla, 2023, p. 112). This concept can be represented by the equation:

$$X=T+EX = T + EX=T+E$$

Where XXX represents the observed score, TTT represents the true score, and EEE represents random error.

The Classical Test Theory provides a clear way to understand human performance by distinguishing between the true score and the observed score. This equation highlights the simplicity of the model, but at the same time, it reveals the limitations of the theory in dealing with variables that may affect results. For instance, environmental or psychological factors may increase the proportion of random error, thus reducing the accuracy of the estimates (Muñiz, 2010, p. 57). Despite this, CTT remains a solid foundation for understanding psychological measurement and is widely used in psychological and educational research. However, it remains limited in its ability to account for indirect factors that may influence an individual's results.

3.2 Reliability

The principle of reliability in CTT is based on the ability of a test to produce consistent results when repeated on the same group or individual under the same conditions. Reliability is calculated using coefficients such as Cronbach's Alpha, which is used to determine the internal consistency of items within the test (Muñiz et al., 2010, p. 58).

Reliability is a crucial element in any psychological test and is considered a key indicator of the test's quality. In CTT, Cronbach's Alpha provides an effective method to assess the consistency of items and evaluate the internal reliability of a test. However, researchers may face challenges in maintaining a high level of reliability when applying tests in different environments or on diverse samples. For example, the reliability level may change when a test is administered to populations that differ in culture or age, which may require redevelopment or modification of the items (Jimam et al., 2019, p. 189). This challenge highlights the limitations of CTT in adapting to changing environments.

3.3 Validity

Validity refers to the extent to which a test measures what it is supposed to measure. In CTT, validity is assessed through various methods such as content validity and criterion validity (Bijlsma et al., 2021, p. 42). While CTT provides strong tools for measuring validity, there are some limitations to these methods.

Validity is one of the most important elements in the design and development of psychological and educational tests. Without validity, a test loses its value as a measurement tool. In CTT, traditional methods for assessing validity rely on comparing results to external criteria or analyzing the extent to which the test covers all aspects of the content it aims to measure. Despite the effectiveness of these methods, there are limitations in their application, particularly in tests that need to predict individual performance in multiple contexts (Muñiz, 2010, p. 60). Additionally, it can sometimes be challenging to measure validity accurately in tests dealing with abstract or multidimensional concepts. Therefore, while CTT provides a useful framework, it may be insufficient at times for analyzing validity in complex tests.

4. CRITICAL ANALYSIS OF THE APPLICATIONS OF CLASSICAL TEST THEORY (CTT)

Despite the widespread use of Classical Test Theory (CTT) in various psychological and educational fields, it is not without criticism. There are several drawbacks and issues associated with the application of this theory in measuring psychological and cognitive abilities. Below are the main criticisms directed at CTT.

4.1 Dependence on the sample

CTT heavily relies on the characteristics of the sample used in the construction and evaluation of the test. This means that test results and its psychological indicators (such as reliability and validity) may be limited to the population on which the test was applied. For instance, if an intelligence test is developed using a small sample of students from a specific city, the accuracy of the test may be affected if applied to different populations from other regions or different age groups (DiCerbo, 2019, p. 67).

This criticism is based on the idea that any test developed using a particular sample may not be generalizable on a larger scale. Since CTT does not adequately account for individual differences, test results may vary significantly when applied to other groups. This implies that test design using CTT requires careful selection and diversity of samples to ensure the generalizability of the results. Although this presents a challenge, many researchers rely on increasing the sample size and its diversity to minimize this issue.

4.2 Fixed measurement

CTT assumes that all items in a test are equally weighted and that each item contributes equally to the total test score. However, these assumptions may be inaccurate, especially when dealing with multidimensional tests or items that vary significantly in difficulty or their ability to differentiate between individuals' abilities (Gannon et al., 2008, p. 319).

This is one of the main criticisms directed at CTT, where all items are treated equally, regardless of their difficulty or ability to distinguish between individuals. In complex tests, this can lead to inaccurate results, as some items may have a greater impact on individuals' overall performance. For example, a single very difficult item may disproportionately determine an individual's overall score, making the estimation of their abilities inaccurate. This criticism highlights the need for more advanced models, such as Item Response Theory (IRT), which accounts for differences in item difficulty.

4.3 Influence of external conditions

CTT relies on the assumption that the test is conducted under ideal and consistent conditions. However, in reality, external factors such as the environment, time, and location may affect individuals' test results. For example, inappropriate surrounding conditions like noise or stress could impact a person's performance, leading to inaccurate results (Muñiz et al., 2010, p. 62).

External conditions represent a factor that is difficult to control in psychological measurement. CTT largely overlooks these influences, meaning that any environmental or psychological impact can result in inaccurate or misleading outcomes. Test design within CTT needs to be more sensitive to these factors to ensure accurate results. However, CTT does not provide sufficient tools to address this type of variability, highlighting the need for more advanced models such as Item Response Theory (IRT), which can adapt to changing conditions.

4.4 Limitations in estimating true abilities

One of the major issues with CTT is its limitations in estimating an individual's true abilities. Since CTT relies on comparing the observed score with the true score, any measurement error directly affects the accuracy of these estimates. In reality, CTT does not sufficiently account for individual

differences among participants, meaning that the estimation of true abilities may sometimes be misleading (Muñiz, 2010, p. 63).

Providing an accurate estimate of individuals' true abilities is a fundamental goal in psychological measurement. However, CTT heavily relies on simplified assumptions that overlook the complexities distinguishing individual differences. As a result, fully relying on CTT can lead to inaccurate or ineffective tests in assessing true abilities, thus diminishing the value of the test results. Designers need to be aware of these limitations when using CTT to develop measurement tools.

5. A COMPREHENSIVE OVERVIEW OF ITEM RESPONSE THEORY (IRT)

Item Response Theory (IRT) offers an advanced alternative to Classical Test Theory (CTT), characterized by its ability to provide more accurate analysis and account for individual complexities. While CTT focuses on analyzing overall test performance based on the observed score, IRT emphasizes the analysis of each item separately in terms of difficulty and discrimination, offering precise estimates of individuals' abilities based on their responses to different items. IRT utilizes mathematical models to analyze individual responses to each item independently, which helps in better estimating individuals' true abilities compared to traditional models (Muñiz, 2010, p. 61).

5.1 Advantages of IRT

- **Measurement accuracy**

  IRT is distinguished by its ability to provide more accurate estimates of individuals' abilities compared to CTT. By focusing on analyzing each item individually, IRT is able to offer a clearer picture of an individual's performance and true abilities. One of the key advantages of IRT is its ability to analyze the difficulty level of each item, thus estimating actual performance based on the individual's responses to items with varying difficulty (Jimam et al., 2019, p. 190).

  IRT provides greater flexibility compared to CTT, as it can handle multidimensional tests or tests that include significant differences in item difficulty. This feature allows researchers to obtain precise estimates of individuals' abilities, regardless of the statistical distribution of the sample. For example, in a test that contains items ranging from easy to difficult, IRT can accurately determine individual abilities based on how a person responds to the more difficult items compared to the simpler ones.

- **Overcoming sample issues**

  One of the most important advantages of IRT is that it does not heavily rely on the sample characteristics, unlike CTT. In IRT, items are analyzed based on individuals' responses independently of the sample used, making the results more generalizable to other populations (Muñiz & Bartram, 2007, p. 59).

  This feature makes IRT a robust model, especially when applying tests in diverse environments. Unlike CTT, which requires a large representative sample to generalize results, IRT allows for the use of tests across diverse samples without significantly affecting the results. This makes IRT a more accurate tool for measurement when dealing with different populations or multiple cultures, thus contributing to increased accuracy in psychological and educational assessments.

5.2 Disadvantages of IRT

- **Complexity in analysis**

  Despite the many advantages that IRT offers, it is more complex in analysis and implementation compared to CTT. IRT requires the use of advanced mathematical models for data analysis, meaning that researchers need specialized software and a deep understanding of statistical mathematics to conduct the analyses (DiCerbo, 2019, p. 64).

Although IRT provides greater accuracy, the complexity associated with it presents a barrier for many researchers and institutions that may not have the resources or expertise required to use this model. In many cases, the cost of training researchers to use IRT or purchasing specialized software can be high, making CTT the preferred option in environments that require simple and quick analyses. This complexity limits the widespread adoption of IRT, despite its theoretical superiority.

- **Need for large samples**

  One of the main limitations of IRT is its need for large samples to ensure the accuracy of estimates. IRT analysis relies on providing precise information about the difficulty and discrimination of each item, which requires a sufficient sample size to generate reliable data (Bijlsma et al., 2021, p. 45).

  IRT requires large samples to ensure accuracy in estimating individual abilities and item difficulties, making it unsuitable for tests conducted on small groups. This presents a challenge, especially in studies with limited samples or in practical testing contexts such as individualized educational tests. In such cases, CTT may be more suitable as it does not require such strict sample size requirements.

  Despite IRT's superiority in providing more accurate estimates of individuals' abilities and overcoming some of the limitations present in CTT, the complexities related to its implementation and the need for large samples make it a tool that is difficult to use on a broad scale. However, in academic and research environments where sufficient resources are available, IRT offers excellent solutions for measuring psychological and educational abilities.

6. CASE STUDIES: PRACTICAL APPLICATIONS OF THE TWO THEORIES

In this section, we will look at how the Classical Test Theory (CTT) and Item Response Theory (IRT) are applied in various practical fields, and how each contributes to improving measurement and evaluation in the domains of education and psychological abilities.

6.1 The use of CTT in educational assessment

CTT is widely used in educational assessment systems around the world, particularly in school achievement tests. Many educational institutions rely on this theory to evaluate students' performance and analyze their scores. By measuring the observed score and determining how close it is to the true score, CTT provides an effective tool for assessing students' academic progress. For example, CTT is used in school achievement tests to measure students' understanding of educational content and to provide accurate reports on their progress compared to national or international standards (Hardin & Leong, 2004, p. 28).

Although CTT has proven its effectiveness in educational assessment over the decades, the heavy reliance on average student performance makes it difficult to identify fine individual differences in abilities. For example, if a test contains a set of items with varying difficulty levels, CTT does not account for the impact of these differences on the final results, which may lead to inaccurate outcomes for some students. Nevertheless, CTT remains highly suitable for assessing academic performance in environments that do not require complex analysis of individual abilities, particularly in group tests that require quick and reliable evaluation.

6.2 The Use of IRT in the assessment of psychological abilities

IRT is particularly used in fields that require precise evaluation of individuals' abilities, such as personality and intelligence tests. In these tests, each item is analyzed separately based on the individuals' responses, providing more accurate estimates of their true abilities. Since IRT can determine the difficulty level of each item and how well it discriminates between respondents, it is used in advanced intelligence tests and personality assessments that require accuracy in estimating psychological abilities (Muñiz, 2010, p. 63).

IRT is an extremely powerful tool in the field of psychological ability assessment, as it allows for detailed and precise analysis of individuals' responses to various items. This can be especially useful in evaluating complex psychological abilities, such as intelligence testing, where IRT can identify items that contribute to measuring true differences between individuals. For example, IRT can be used in a test that includes a set of items with varying levels of difficulty, where an individual's ability is estimated based on their performance on each item with high accuracy. This enhances the effectiveness of psychological assessments, making them more responsive to individual differences compared to CTT.

However, the use of IRT requires significant resources, including specialized software and large samples for accurate data analysis. This may make it difficult to apply IRT in environments where these resources are limited. Nevertheless, its ability to provide precise estimates means that its use in psychological fields can offer highly valuable and accurate information about individuals' abilities.

While CTT is widely used in educational assessments aimed at providing a general picture of academic performance, IRT offers more complex and accurate solutions for analyzing psychological abilities. In both cases, each theory can have its appropriate place in the context in which the test is used, but IRT remains the better choice when there is a need for in-depth analysis of individual performance at the item level.

7. COMPARISON BETWEEN CTT AND IRT

The Classical Test Theory (CTT) and the Item Response Theory (IRT) represent two different models for analyzing and evaluating performance in psychological and educational tests. While CTT focuses on averages and reliability and relies on the analysis of overall observed scores, IRT focuses on a detailed analysis of each item individually, taking into account item difficulty and individual ability. Below is a detailed comparison of the various aspects of both models:

Aspect	**CTT**	**IRT**
Sample Dependence	High, test results are influenced by the sample	Low, less dependent on the sample
Accuracy	Less accurate in predicting individual performance	More accurate in individual assessment
Required Resources	Low, does not require complex techniques	High, requires advanced software and more resources
Ease of Application	Relatively easy	Complex, requires advanced mathematical analysis

While CTT may be more common and easier to apply, IRT provides much greater accuracy in estimating individual abilities and analyzing individuals' performance on different items. In terms of precision, IRT allows for more detailed estimation of each item individually, enabling researchers to identify fine individual differences between respondents. However, in terms of required resources, IRT demands additional resources and more complex analysis compared to CTT.

CTT remains suitable in situations that do not require in-depth analysis or when working in resource-limited environments, whereas IRT excels in academic or research settings that require high accuracy in assessments.

8. FUTURE TRENDS IN PSYCHOLOGICAL MEASUREMENT

With the advancement of technology and the increased reliance on digital tools, the field of psychological measurement is moving toward the use of innovative methods and techniques, such as game-based and simulation-based assessments. These methods offer new opportunities to improve the accuracy of psychological assessments by simulating real-life situations and evaluating individuals' responses in interactive environments. These modern approaches allow for deeper and more comprehensive evaluations of individuals' abilities and behaviors in ways that may not be possible using traditional tests (DiCerbo & Behrens, 2014, p. 85).

Game-based and simulation-based assessments represent a significant step forward in improving the quality and accuracy of psychological evaluations. These methods can provide data that are more relevant to real-world scenarios, enhancing the validity and effectiveness of the assessments. Additionally, another trend is the development of new models that integrate CTT and IRT, where the simplicity of CTT is combined with the precision of IRT to create more flexible and effective measurement tools. These hybrid models contribute to meeting the growing demands in the modern field of psychological measurement and ensure the provision of comprehensive assessment tools that are adaptable to various environments and contexts (Jimam et al., 2019, p. 191).

9. CONCLUSION

As psychological measurement continues to evolve, it is likely that IRT will become more widespread due to its high accuracy and flexibility in handling individual differences. However, the use of CTT will remain prevalent in environments that require simple and quick solutions. Ultimately, hybrid models and innovative techniques such as interactive assessments could bring about a significant shift in how psychological performance is analyzed and evaluated.

REFERENCES

Bijlsma, H., Rollett, W., & Spiel, C. (2021). A Reflection on Student Perceptions of Teaching Quality from Three Psychometric Perspectives: CCT, IRT and GT. In W. Rollett et al. (Eds.), Student Feedback on Teaching in Schools (pp. 42-45). Springer.

DiCerbo, K., & Behrens, J. (2014). Impacts of the Digital Ocean on Education. Technology, Knowledge and Learning, 19(1), 85-99. https://doi.org/10.1007/s10758-014-9211-8

DiCerbo, K. (2019). Psychometric Methods: Theory into Practice. Measurement: Interdisciplinary Research and Perspectives, 17(1), 60-64. https://doi.org/10.1080/15366367.2018.1521190

Gannon, T. A., Keown, K., & Rose, M. (2008). An Examination of Current Psychometric Assessments of Child Molesters' Offense-Supportive Beliefs. International Journal of Offender Therapy and Comparative Criminology, 53(3), 316-333. https://doi.org/10.1177/0306624X07312791

Hardin, E. E., & Leong, F. T. (2004). Decision-Making Theories and Career Assessment: A Psychometric Evaluation of the Decision Making Inventory. Journal of Career Assessment, 12(1), 22-41. https://doi.org/10.1177/1069072703257730

Hussein, A., & Gasmalla, H. E. (2023). Introduction to the Psychometric Analysis. In H. E. E. Gasmalla (Ed.), Written Assessment in Medical Education (pp. 112-115). Springer. https://doi.org/10.1007/978-3-031-11752-7_9

Jimam, N. S., Ahmad, S., & Ismail, N. E. (2019). Psychometric Classical Theory Test and Item Response Theory Validation of Patients' Knowledge, Attitudes and Practices. Journal of Young Pharmacists, 11(2), 186-191. https://doi.org/10.5530/jyp.2019.11.39

Muñiz, J. (2010). Test Theories: Classical Theory and Item Response Theory. Papeles del Psicólogo, 31(1), 57-66.

Muñiz, J., & Bartram, D. (2007). Improving Personnel Selection through the Use of IRT. International Journal of Selection and Assessment, 15(1), 58-63. https://doi.org/10.1111/j.1468-2389.2007.00369.x