

The Historical Development, Significance, and Challenges of Testing and Measurement in Educational Settings: A Literature Review

Wasantha Geekiyanage¹ Sabeeth Gunasekera²

Department of Western Music, Faculty of Music, University of the Visual and Performing Arts,
Colombo, Sri Lanka¹

School of Medicine and Dentistry University of Lancashire, Preston, Lancashire, UK²

DOI: <https://doi.org/10.47772/IJRISS.2026.1026EDU0604>

Received: 14 September 2026; Accepted: 19 September 2026; Published: 02 October 2026

ABSTRACT

Testing and measurement have become a vital part of our education system today, as they are important tools for measuring student learning, making education decisions, and holding educational systems accountable. Due to social, political, psychological, and pedagogical forces, educational testing and measurement have evolved over many centuries. This study focused on investigating relevant literature to understand the historical development and evolution of key theories and concepts related to educational testing and measurement, as well as assessing their significance in educational contexts. This study employed a narrative literature review to analyze existing scholarly studies. The review explains how formal testing came into being in ancient China, how oral and written exams came about in medieval Europe, and the scientific measurement of the nineteenth and twentieth centuries as a result of the progress in psychometrics and intelligence testing. The study has elucidated the contributions of innovators such as Francis Galton, Alfred Binet, and Theodore Simon in the development of modern educational assessments. The development of theories in the field of educational measurement has been explained, and the importance of using testing and measurements in learning and assessing curriculum, educational planning, and decisions has been highlighted. Finally, this paper looks into some of the challenges related to educational testing, like issues of validity, fairness, and culture. It was discovered that, despite being an ongoing process, testing and measurements play a vital role in enhancing the quality of learning and the achievements of students in various settings.

Keywords: Educational measurement, Educational testing, Assessments, Standardized testing, Educational evaluation

INTRODUCTION

Testing and measurement occupy a central position within educational practice and research (Sivakumar & Thirumoorthy, 2019). The words testing, measurement, assessment and evaluation are frequently used interchangeably, but they are different and related. Measurement is the assignment of numbers to the characteristics or performance of objects following a set procedure (Nitko & Brookhart, 2014). Testing is a particular method of gathering information about the knowledge, skills, abilities or attitudes of the person by means of a task or question that is laid out in a structured manner (Adom et al., 2020). In the terminology of Brown and Abeywickrama (2019), assessment is a wider process of collecting, interpreting and utilizing information on student learning and evaluation, making judgments about the quality or effectiveness of learning programmes, learning methods and student learning. Evaluation is an ongoing process of analyzing, assessing and judging on the basis of evidence a programme, intervention or education practice to determine its value, quality, effectiveness or appropriateness (Stufflebeam & Coryn, 2014). In this sense, testing and measurement are important aspects of educational assessment, and assessment goes beyond producing data or test scores to interpreting and utilizing evidence of learning (Adom et al., 2020)

Testing and measuring is an ancient practice which has been carried out by humans since the beginning of time to evaluate knowledge, skills, and abilities. Societies have looked for ways to discover who is capable, to find

ways to provide education, and to measure performance throughout history. Initially straightforward tests, have turned into sophisticated educational measurement systems involving statistical analysis, psychometric theory, and technical innovations (Wise & Rutstein, 2024; Clauser & Margolis, 2023). The significance of testing and measurement goes much deeper than just the awarding of grades (Jd, 1977; Burdett, 2016). The field of educational measurement has changed in the last few decades, thanks to recent developments in educational technology, cognitive psychology, and psychometrics (Clauser & Margolis, 2023; Educational Testing Service, 2025). New tools that come with computer-based testing, adaptive assessment, learning analytics and data-driven decision-making have widened the potential avenues for more efficient and meaningful learning assessments. Simultaneously, issues of fairness, cultural bias, over-standardization, and high-stakes testing have raised continuing controversies about the field of educators and researchers (Burdett, 2016; Clauser & Margolis, 2023). This literature review will examine the history of testing and measurement, discuss how educational measurement theory has evolved and discuss the significance of testing and measurement in educational environments. Knowing the history and current use of educational assessment can help teachers gain insights that assist in effective instruction and learning. (Clauser & Margolis, 2023; Sivakumar & Thirumoorthy, 2019).

Objectives of the Study

This study aimed to review literature and analyze the history of testing and measuring, and to assess their importance in educational settings in general. Accordingly, the following specific objectives were pursued:

1. To investigate how testing and measurement have changed over time in the educational context.
2. To explore the evolution of major theories and concepts related to educational testing and measurement.
3. To assess the value of testing and measurement in the assessment of student learning and in educational decision making.

METHODOLOGY

In this study, the narrative review was chosen due to the ability to synthesize and interpret knowledge from many different scholarly sources to gain a broad overview of the development, use and importance of educational testing and measurement (Sukhera, 2022). The literature was located using electronic databases such as Google Scholar, ERIC, Scopus, Web of Science, and JSTOR. The following were used as keywords in the process of searching: educational testing, educational measurement, educational assessment history, educational assessment, standardized testing and psychometrics. The following table presents the inclusion and exclusion criteria used to select relevant studies.

Table 01: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
• Peer-reviewed journal articles • Books and book chapters • Conference proceedings • Publications directly related to the research topic • Seminal publications were included irrespective of publication year	• Duplicate publications • Non-scholarly sources • Studies unrelated to the research objectives • Articles without full-text access

Sources were chosen because of their relevance to the historical development of testing and measurement, their significant contributions to the theory of testing and measurement, their educational applications, and problems of testing and measurement today. The most important and often-quoted publications were chosen. Publications already in print were also considered for the review, even if they were published after 1900, because the review focused on the history of testing and measurement.

Thematic analysis was employed to review, compare and synthesize the collected literature. Themes relating to the evolution of testing over time, theories of measurement, the meaning of these tests in education and issues pertaining to the use of testing practices were identified and clustered into thematic areas. These findings were then analyzed to gain a coherent picture of the nature of testing and measurement, and its development and function in educational contexts. This study is an information analysis including only published source materials and does not involve human subjects, for which ethical approval is not required. The study acknowledges and cites all sources with academic integrity.

Historical Development of Testing and Measurement

Throughout history, the testing and measurement of knowledge, skills, and human performance have been an ongoing process for societies seeking to assess them. To gain the most benefit from studying the history of testing and measurement, it is important to understand how testing and measurement have developed over time and how they continue to shape the field of education and education policy today. This section examines the major milestones and developments that have shaped the evolution of testing and measurement in educational settings in relation to assessment.

Early Forms of Educational Examination and Assessment

The roots of educational testing go back to ancient civilizations, and various techniques existed to assess knowledge, skills, and suitability for social and governmental functions (Kellaghan & Greaney, 2019). These were not standardized testing systems like those used today, but they were different types of assessment that helped pave the way for the development of educational measurement. Ancient societies closely linked assessment to practical purposes, ethical principles, and social structures, rather than formal education goals. Education in ancient Greece focused on intellectual development, philosophical inquiry, and cultivating moral character (Yousaf et al., 2025; Murphy, 2015; Madaus and O'Dwyer, 1999). The ideas of Plato and Aristotle, key philosophers, were that learning was about acquiring wisdom and virtue. Discussion and debate, or oral dialogue, were heavily used in the educational system, for example, at Plato's Academy, there was little evidence of any formal written examinations or systematic assessment procedures. Student knowledge was expressed in the form of participation in philosophical discussion rather than in the form of a test (Bowen, 1981, pp. 97-99). The assessment in this period was, therefore, informal and mostly done through observation and interaction between the teacher and learners. In the same way, ancient Roman education was centered on rhetoric and public speaking, as well as skills for civic engagement. Students were assessed primarily by recitation, oral presentation, and rhetorical exercises. The ability to demonstrate knowledge and be convincing in oral communication was a major determinant of success in education rather than the achievement of formal examinations (Boyd & King, 1975). They were in line with the philosophy of education of their time and sought to develop character and hands-on skills.

The earliest and most important instance of systematic testing was from ancient China. One of the earliest and largest efforts to evaluate people, in a standardized way, occurred with the Chinese imperial examination system, which was established during the Han Dynasty and was completely institutionalized during the Sui and Tang dynasties (Zhao, 2023; Kamens, 2015). The mere purpose of these examinations was to place government officials based on their merit instead of their hereditary rights. The candidates had to pass a series of exams at the local, provincial, and national levels to test their knowledge of the Confucian classics, literature, and administrative principles (Han & Yang, 2001; Madaus & Kellaghan, 1992). The system continued for more than a thousand years and had a profound influence on educational and governmental practices in many parts of the world. The Chinese examination system is known to be one of the oldest merit-based assessment systems. Elman (2013) states that the civil service examinations produced a professional administrative class and laid a foundation for the idea that education could be a criterion for social mobility.

Later in the period, Europe and North America adopted the Chinese model of the exam, which significantly influenced today's test-taking systems. Despite some criticisms regarding the exam system's overemphasis on memorization and underemphasis on creativity, it represented a significant milestone in the history of educational measurement. The early history of educational testing generally indicates that testing practices changed in response to social, political, and educational needs. From oral exams of ancient Greece and Rome to the well-established system of exams in imperial China, early assessment systems shared the values and priorities of the societies through which they passed. The historical developments have led to the modern systems of testing and measurement, which occupy a prominent place in present-day education.

Medieval and Early Modern Examinations

The evaluation of education in medieval times gradually evolved in connection with the development of formal educational institutions in Europe. Medieval education was linked with universities and religious institutions supplemented by little dialogue or demonstration of skills (Wang, n.d.; Gamisonia, 2024; Bevis, 2019). The assessment systems of ancient civilizations were mostly conversational and demonstrative. More systematic and formal methods of assessment were being sought, but formal written examinations were not yet a common practice (Romanelli et al., 2025). This was a precursor of the more formalized systems of examination developed in the early modern period. In medieval Europe, education was mainly based in monasteries and cathedral schools and, subsequently, universities (Rashdall, 2010, pp. 7–16; Bowen, 1981). Higher learning flourished and institutions like the Universities of Bologna, Paris and Oxford drew students from far and wide (Haskins, 1923; European universities emerge, 2021). In these institutions, the assessment methods were mainly oral examination, disputation, and public demonstration of knowledge. Students were supposed to argue and hold scholarly discussions with Masters or Professors. Being able to reason with logic and communicate effectively as well as have knowledge was important for success (Chang, 2021, pp. 10–15; Bowen, 1981). Disputation was one of the features of medieval assessment. In this approach, students were engaged in carefully controlled debates with the ability to answer questions and to back up their understandings of texts that were theological, philosophical, or legal in nature. During this time, oral exams became a significant tool for assessing academic performance and were deemed an integral part of higher education (Boyd & King, 1975). The teachers could assess students individually because their classes were small and they had few classes.

The universities grew in the later Middle Ages, and educational authorities started looking for more reliable ways of certifying academic achievements (Wilbrink, 1997). The Bachelor of Arts and Master of Arts degrees started to represent scholarly proficiency, and tests began to be used more and more to bestow these degrees. These examinations, however, were still largely oral and subjective, with the emphasis being placed on the judgment of individual examiners rather than on set criteria (Stobart, 2008). Education was affected by intellectual and cultural changes that occurred during the Renaissance (14th century to the 16th century). Humanist scholars gave pride to the study of classical languages, literature, history, and rhetoric. Educational institutions grew significantly after the invention of the printing press, and books were created, making it easier for many to learn to read. These developments led to a growing need for systemized teaching and assessment (Cuban, 1988).

Written examinations became increasingly common in the early modern period. In schools and universities, knowledge of academic disciplines in students was increasingly tested by writing assignments, essays, translations and compositions (Kruse, 2006; Harrow, 2001). Written examinations had several benefits over oral examinations. These tools helped teachers evaluate more students, maintain a permanent record of their work, and provide greater consistency in student evaluations. As a result, written assessments started to substitute in some instances, replacing traditional oral assessments (Stobart, 2008). Formal education continued to increase in the seventeenth and eighteenth centuries, and the state began to play a larger role in the administration of education. Governments and educational institutions appreciated the role of assessment in maintaining standards and selecting individuals for academic and professional purposes (Wilbrink, 1997). The growth in social and administrative requirements stimulated the need to address more formal examination procedures. Gradually, assessment became more widespread and individual, developing into a certification, selection, and accountability system (Madaus and O'Dwyer, 1999).

At the end of the eighteenth century, examinations had become a part of the educational systems of many parts of Europe. These tests were not statistically based or standardized like contemporary tests but were a significant step in the development of tests used in education. This shift from oral to written questioning was part of a wider social transformation, such as greater access to education, greater bureaucracy, and an increased demand for objective measures of achievement (Harrow, 2001; Stobart, 2008; Wilbrink, 1997). The medieval and early modern eras were important in the history of educational testing in general. As the universities grew and the educational institutions looked to more accurate ways of certifying knowledge and competence, the assessment practices were increasingly formalized. The gradual transition from oral exams to written exams laid the foundation for much of the theory later used for the development of standardized testing and scientific measurement during the 19th and 20th centuries.

The Rise of Scientific Measurement

Educational testing and measurement turned a corner in the 19th century; assessment shifted from examinations to the scientific measurement of human capacities. The new development of psychology, statistics, biology, and experimental science inspired a group of scholars to look for objective methods to study individual differences and measure mental characteristics (Fancher & Rutherford, 2017). Francis Galton was one of the most important pioneers in this movement. Galton's interest in the study of heredity and individual differences derived from his inspiration from the theory of evolution by Charles Darwin. He wanted to find ways to measure human characteristics scientifically (Goodwin, 2015). He set up an anthropometric laboratory in London, and gathered a huge amount of data about the physical and sensory characteristics of thousands of people. It was among the first to use quantitative methods to study human abilities (Fancher & Rutherford, 2017). Galton's work on intelligence was only a part of his contributions. He also devised statistical methods that proved essential, later in measuring the effectiveness of education. His research contributed to the development of ideas that later found application in psychometrics and educational research such as correlation and regression that would become fundamental tools in many areas of this research (Miller, 1962). These statistical advances enabled scientists to consider the relationships between variables and to ascribe scientific meaning to the study of individual differences. Galton's work thus influenced the development of educational and psychological measurement (Fancher & Rutherford, 2017). The phrase 'mental tests' were first used in 1890 by the American psychologist James McKeen Cattell, following Galton's work. An advocate of experimental methods in psychology, Cattell developed a way to test sensory and cognitive abilities (Goodwin, 2015). His tests measured reaction time, sensory discrimination, memory and muscle strength. Critics later criticized some of these initial tests for their limited validity, but they still represented major advances in systematic approaches to human assessment. Many of these initial tests were subsequently criticized as lacking in validity but nevertheless were steps forward towards systematic approaches to human assessment (Fancher & Rutherford, 2017).

Scientific measurement also occurred at the same time as the beginnings of the academic discipline of psychology. In 1879, Wilhelm Wundt established his experimental laboratory at the University of Leipzig that focused on observation, experiment, and measurement as key aspects of science (Goodwin, 2015). This conviction led to the belief that empirical methods could objectively study mental processes and human attributes. Consequently, more and more principles of reliability, objectivity, and standardization became influential in educational assessment (Fancher & Rutherford, 2017). By the end of the nineteenth century, the roots of psychometrics had clearly taken shape, while some early methods of measurement were limited in theory and method. These were all an improvement from the subjective methods of evaluation to more scientific and quantitative methods (Fancher & Rutherford, 2017). These developments eventually gave rise to the intelligence tests, achievement tests, and standardized assessment systems that typified the twentieth century (Goodwin, 2015).

Intelligence Testing and the Binet-Simon Scale

One of the great landmarks in the history of educational measurement was in the early 1900's when the Government of France requested Alfred Binet and Theodore Simon to develop a means for identifying students requiring extra educational support. The first practical intelligence test was published in 1905 (Binet-Simon Scale). The Binet-Simon Scale, unlike earlier tests, focused on higher-order cognitive abilities like

memory, reasoning, judgment, and problem-solving. Binet believed intelligence was something that could be changed by education, not a fixed trait to be labelled for life (Binet & Simon, 1916). The success of Binet-Simon Scale increased the interest in testing intelligence and the development of many educational and psychological tests. The use of intelligence testing soon became a major component of educational placement, guidance, and intervention programs throughout Europe and North America (Anastasi & Urbina, 1997; Fancher & Rutherford, 2017).

Standardized Testing in the Twentieth Century

The 20th century was a watershed in the history of educational measurement, and a tremendous expansion of standardized testing occurred. The need for objective and efficient means of measuring large populations led to the development and widespread use of standardized assessments (Anastasi & Urbina, 1997; Fancher & Rutherford, 2017). The development of psychology, statistics, and psychometrics has made the design and implementation of testing programs more scientific, and hence, educational assessment has become more systematic and reliable (Crocker & Algina, 2008). An important breakthrough during the First World War was the development of the Army Alpha and Army Beta tests by the United States Army for the efficient classification of soldiers. These tests were administered to groups and showed that it was feasible to assess large numbers of individuals and that standardized procedures were useful in assessing a diverse group of students who have differing educational backgrounds (Anastasi & Urbina, 1997). The results of their success had a significant influence on educational testing and led to an increase in the use of group intelligence tests in schools and colleges (Fancher & Rutherford 2017). Since the war, standardized testing in educational institutions began to play an important role. The development of achievement tests, aptitude tests, and intelligence tests aimed to measure various aspects of achievements and potential of the students. The use of standardized tests for educational selection and placement purposes grew, and it became part of the institutionalization of testing in Secondary and Higher education (Kaplan & Saccuzzo, 2018). The scientific quality of educational measurement was further enhanced through advances in item analysis, estimation of reliability, and statistics (Crocker & Algina, 2008).

In the mid-twentieth century, there were major advancements in technology that impacted assessment practice. Quick and efficient scoring of multiple-choice questions became possible with the advent of machine-scoring and optical mark recognition systems. The innovations enabled the management of large-scale evaluations and helped to increase the use of standardized exams in schools and universities (Stobart, 2008). This resulted in tests becoming cheaper and more students being tested. During the latter half of the 20th century, educational testing began to be incorporated into the school and educational system as a whole in many countries throughout the world. Assessment data were increasingly used by governments to assess educational performance, monitor school effectiveness, and make policy decisions. However, there was growing concern about accountability and about evidence-based educational reform, which led to the widespread development of international assessment programs and national testing systems (Kamens & McNeely, 2010). The major concerns related to standardized testing were those of education quality, national competitiveness and equity. At the same time, psychometricians developed new strategies to enhance the validity, reliability, and fairness of educational assessment. Increased development and standards for testing measurement theory resulted in increased precision and consistency in assessment practices. These developments helped establish the credibility of standardized testing and the principles that underlie measurement in education today (American Educational Research Association et al., 2014; Crocker & Algina, 2008). In general, standardized testing underwent tremendous development and innovation during the 20th century. The developments of science and technology and the needs of education have widened the scope of testing from a small portion to a bigger one in modern education. This progress has resulted in the contemporary techniques in measuring and assessing education (Stobart, 2008; Fancher & Rutherford, 2017).

Evolution of Educational Measurement Theories

The development of educational measurement theories is a continuous process of striving to make more accurate and more fair assessments of students. The early history of educational measurement, the late nineteenth- and early twentieth-century psychometric movement, significantly influenced the early history of educational measurement. In this time, researchers tried to use scientific methods to evaluate human

capabilities, developing standardized testing procedures and quantitative measurements of ability (Anastasi & Urbina, 1997, Cook & Pitoniak, 2025, pp. 34-35). The earliest development in educational measurement was Classical Test Theory (CTT). Classical Test Theory suggested that a person's observed score is equal to the true score in addition to an error score. This framework provided the basis for understanding the concepts of reliability, validity, and measurement error in educational evaluation (Zanon et al., 2016; Lord & Novick, 1968, p. 55). CTT was easily accepted because of its simplicity and utility, but subsequently researchers found some drawbacks of sample dependence and item properties.

The development of more sophisticated assessments in education led to the development of an alternative measurement framework, namely 'Item Response Theory (IRT).' IRT has been used for many decades with consistent outcomes in educational testing (Ogunsakin & Shogbesan, 2018). It is more concerned with the link between a person's latent ability and their likelihood of getting a particular test item correct than classical test theory. The theory enables more accurate estimates the ability of students and can assist with the creation of systems of adaptive testing (Zanon et al., 2016; Hambleton et al., 1991). IRT is a superior framework that researchers can use to analyze non-cognitive data from sociological, psychological, and psychiatric examinations, as well as cognitive data in assessment and evaluation studies. For results to be valid, dependable, and credible, the test data must fit the IRT model, and all statistical assumptions must be satisfied. IRT is now the basis for a number of large-scale standardized tests used throughout the world today (Ogunsakin & Shogbesan, 2018).

Educational measurement continued to develop with the emergence of more sophisticated theoretical models after Classical Test Theory (CTT) and Item Response Theory (IRT) were created. One of the most important developments was Generalizability Theory (G-Theory), developed by Cronbach et al. (1972), which was an extension of traditional reliability theory and allowed researchers to account for and estimate more than one source of measurement error at a time. While Classical Test Theory usually conceals the nature of measurement error under a single umbrella of "error", Generalizability Theory examines the different facets of error, including rater, task, occasion, and testing condition (Brennan, 1992, p. 27; Cronbach et al., 1972). Based on these further developments, Diagnostic Classification Models (DCMs) or cognitive diagnostic models emerged, describing in detail the mastery states of a learner's knowledge, skills and cognitive attributes, rather than just reporting the overall result of the test. These models can be used as a diagnostic tool to pinpoint individual student learning strengths and weaknesses, which can then be used to guide more focused instructional interventions (Rupp et al., 2010). Another significant development was an increased focus on validity. Previous assessment systems tended to pay attention mainly to the consistency and reliability of scores. Nevertheless, in addition to statistical evidence, tests should be evaluated for validity in terms of the social implications and meanings of test scores as scholars such as Messick (1989) have claimed. It shifted the emphasis of assessment from technical activity to one that considers ethical, educational and social aspects.

Computerized Adaptive Testing (CAT) is another example of how the field of educational measurement has been revolutionized by computer technology. CAT generally relies on Item Response Theory (IRT) to dynamically select items that are indicative of an examinee's cognitive ability level (Van der Linden & Glas, 2010; Weiss, 2011), thus maximizing the precision of the score measurement while minimizing the testing time and burden on the examinee. In recent years, measurement has also been more widely used in e-learning settings, particularly in combination with Artificial Intelligence and automated item generation methods that enhance the validity, fairness and efficiency of assessments (Von Davier et al., 2019; Gierl & Lai, 2013).

In the last few decades, there has been an evolution of measurement theories that goes beyond the traditional testing paradigm in the field of education. The current testing systems recognize the importance of formative assessment, authentic assessment, performance-based assessment, and competency-based measurement. These shifts represent a shift in thinking and a new focus on learning as a multi-faceted process, instead of the gathering of facts. This has given rise to the contemporary education measurement conception, which focuses more on quantitative precision and at the same time aims for a more meaningful assessment of student learning.

Significance of Testing and Measurement in Educational Settings

Testing and measurement play a fundamental role in educational systems across the world. They provide educators with valuable information regarding student achievement, learning progress, instructional effectiveness, and curriculum outcomes. Through systematic assessment, teachers can identify strengths and weaknesses of students and make informed decisions regarding instructional strategies and educational interventions. One of the primary functions of testing is the assessment of student learning. Educational assessments help determine the extent to which learners have achieved intended learning outcomes and curriculum objectives. Such information enables educators to evaluate whether instructional goals have been met and whether modifications to teaching practices are necessary (Nitko & Brookhart, 2014). Testing and measurement also contribute significantly to educational accountability. Schools, educational institutions, and policymakers often rely on assessment data to evaluate educational quality and effectiveness. Standardized tests provide comparable information that can be used to monitor educational performances across different regions, schools, and student populations. This information supports evidence-based policy development and resource allocation (Supovitz, 2006; Linn, 2010).

Another notable contribution of testing is its role in guiding educational placement and selection decisions. Test results are frequently used to determine student admission, promotion, certification, and placement into specialized educational programs. When appropriately designed and implemented, assessments can help ensure that such decisions are based on objective and consistent criteria. Wilson (2023) suggests that instruments designed based on item response modeling can provide more intricate perceptions of student performance. Through the accurate identification of students' ability and needs, educators can provide flexible teaching for students with diverse needs. This congruence of assessment and instruction can aid in improving learning outcomes, because teachers have the data, they need to help guide diverse learning pathways.

Regular assessment provides feedback to both teachers and students regarding current levels of achievement and areas requiring improvement. Research suggests that effective feedback can enhance motivation, self-regulation, and academic performance (Black & Wiliam, 1998). Consequently, assessment is increasingly viewed not merely as a tool for measuring learning but also as a mechanism for improving learning. Furthermore, educational measurement contributes to educational research by providing reliable data for investigating learning processes, instructional methods, and educational outcomes. Accurate measurement enables researchers to evaluate interventions, test theoretical models, and generate evidence that supports educational improvement.

Challenges and Criticisms

Testing and measurement are critical but have been the subject of a tremendous deal of criticism. One of the most significant fears is the overreliance on standardized testing. Critics, however, argue that overemphasis on test scores may lead to a narrowing of the curriculum and an emphasis on teaching strategies that focus on the outcomes of examinations rather than on genuine learning (Au, 2007). Another big criticism is the fairness and bias issues. Cultural, language, socioeconomic or context differences can place some groups of students at a disadvantage in educational assessments. Test items that are not responsive to diverse backgrounds may produce distorted measures of student ability and performance. Assessment design and implementation, therefore, continue to be of significant concern in fairness (Jonson et al., 2019; Walker et al., 2023). The phenomenon popularly known as “teaching to the test” has attracted criticism as well. If educational success is closely linked to assessment outcomes, teachers may give priority to test preparation activities over creativity, critical thinking and broader educational goals (Popham, 2001; Welsh et al., 2014). Such practices may hamper the opportunities for holistic development of students. Some experts say standardized tests may not be able to measure complex human skills such as creativity, collaboration, problem solving and emotional intelligence. Assessments can be effective in assessing some knowledge and skills but may not capture the full spectrum of competencies required in today’s world (Paramole & Adeoye, 2024; Upguage, n.d.).

The progress of technology presents both opportunities and challenges. Computer-based assessments are more efficient and flexible but problems like access to digital devices, skills of people to use technology, privacy of data and security still exist. Therefore, educational institutions must ensure that technological innovations do

not create new inequalities among learners (Pereira, 2023; Rajuroy, 2025; Wang et al., 2026). Such criticisms underscore the importance of balanced assessment systems that include standardized measurement as well as other testing approaches. Such systems can help to develop more comprehensive and equitable assessments of student learning while preserving the reliability and validity needed for educational decision making (Pereira, 2023; Paramole & Adeoye, 2024).

CONCLUSION

The history of testing and measurement shows that with it has come a constant, concerted drive to enhance assessment of learning and learning achievement. Educational measurement has changed significantly since the early times of the development of standardized tests and psychometric theories, to the current focus on validity, fairness and authentic assessment. Classical Test Theory (CTT) and Item Response Theory (IRT) have been developed to improve the scientific rigor of assessment practices in the field of education. Testing and measurement remain a crucial part of education systems for assessment, program improvement, accountability, and policy making. Concurrently, ongoing problems of fairness, bias, overemphasis on standardized tests, and the measurement of complex competencies highlight the need for continued development of assessment practices. With the changing face of society and the rapid pace of technology development, the nature of education is evolving rapidly, and so should testing and measurement in relation to assessment be valid, equitable, and meaningful. Going forward, it is important to achieve a balance between technical accuracy and a holistic understanding of student learning, which would help further advance inclusive and effective learning systems.

REFERENCES

1. Adom, D., Adu-Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, 9(1), 109–119. <https://doi.org/10.11591/ijere.v9i1.20457>
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
3. Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
4. Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258–267.
5. Bevis, T. B. (2019). The notion of universities. In *A world history of higher education exchange: The legacy of American scholarship* (pp. 35–53). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-12434-2_2
6. Binet, A., & Simon, T. (1916). *The development of intelligence in children* (The Binet-Simon Scale). (E. S. Kite, Trans.). Williams & Wilkins Co. <https://doi.org/10.1037/11069-000>
7. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
8. Bowen, J. (1981). *A history of Western education: Volume 1. The ancient world: Orient and Mediterranean, 2000 B.C.–A.D. 1054* (3rd ed.). Methuen.
9. Boyd, W., & King, E. J. (1975). *The history of Western education* (10th ed.). Adam and Charles Black.
10. Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27–34. <https://doi.org/10.1111/j.1745-3992.1992.tb00260.x>
11. Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.
12. Burdett, N. (2016). The good, the bad, and the ugly: Testing as a key part of the education ecosystem (RISE Working Paper Series No. 16/010). *Research on Improving Systems of Education (RISE)*. https://doi.org/10.35489/BSG-RISE-WP_2016/010
13. Clauser, B. E., & Margolis, M. J. (2023). Past, present, and future of educational measurement. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (4th ed., pp. 1–14). Elsevier. <https://doi.org/10.1016/B978-0-12-818630-5.10001-6>

14. Cook, L. L., & Pitoniak, M. J. (Eds.). (2025). Educational measurement (5th ed.). Oxford University Press. National Council on Measurement in Education (NCME) PDF
15. Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage Learning.
16. Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Wiley.
17. Cuban, L. (1988). The managerial imperative and the practice of leadership in schools. State University of New York Press.
18. Educational Testing Service. (2025, February). Why standardized testing matters: A teacher's guide to fair and effective assessments. ETS Global. <https://www.etsglobal.org/na/en/blog/news/why-standardized-testing-matters>
19. Elman, B. A. (2013). Civil examinations and meritocracy in late imperial China. Harvard University Press.
20. European universities emerge. (2021). In EBSCO Research Starters. EBSCO Information Services. <https://www.ebsco.com/research-starters/history/european-universities-emerge#full-article>
21. Fancher, R. E., & Rutherford, A. (2017). Pioneers of psychology (5th ed.). W. W. Norton & Company.
22. Gamisonia, N. (2024). Some issues of education and upbringing in the Middle Ages. International Scientific Journal "The Caucasus and the World", (28), 74–83. <https://doi.org/10.52340/isj.2024.28.09>
23. Gierl, M. J., & Lai, H. (2013). Using automatic item generation to create test items. Educational Measurement: Issues and Practice, 32(3), 3–10. <https://doi.org/10.1111/emip.12018>
24. Goodwin, C. J. (2015). A history of modern psychology (4th ed.). John Wiley & Sons.
25. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.
26. Han, M., & Yang, X. (2001). Educational assessment in China: Lessons from history and future prospects. Assessment in Education: Principles, Policy & Practice, 8(1), 5–10. <https://doi.org/10.1080/09695940120033216>
27. Harrow, M. (2001). The shift from oral to written examination: Cambridge and Oxford, 1700–1900. Assessment in Education: Principles, Policy & Practice, 8(1), 33–50. <https://doi.org/10.1080/09695940120033243>
28. Haskins, C. H. (1923). The rise of universities. Henry Holt and Company. <https://www.gutenberg.org/files/63574/63574-h/63574-h.htm>
29. Jd, E. (1977). A comparison of diagnostic and certifying examinations. American Journal of Medical Technology, 43(5), 487–490.
30. Jonson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. Educational Measurement: Issues and Practice, 38(3), 6–19. <https://doi.org/10.1111/emip.12259>
31. Kaplan, R. M., & Saccuzzo, D. P. (2018). Psychological testing: Principles, applications, and issues (9th ed.). Cengage Learning.
32. Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. Comparative Education Review, 54(1), 5–25. <https://doi.org/10.1086/648471>
33. Kamens, D. H. (2015). A maturing global testing regime meets the world economy: Test scores and economic growth, 1960–2012. Comparative Education Review, 59(3), 420–446. <https://doi.org/10.1086/681989>
34. Kellaghan, T., & Greaney, V. (2019). A brief history of written examinations. In Public examinations examined (pp. 43–74). The World Bank. https://doi.org/10.1596/978-1-4648-1418-1_ch3
35. Kruse, O. (2006). The origins of writing in the disciplines: Traditions of seminar writing and the Humboldtian ideal of the research university. Written Communication, 23(3), 331–352. <https://doi.org/10.1177/0741088306289259>
36. Ku-ming (Kevin) Chang (2021). For the love of the truth: The dissertation as a genre of scholarly publication in early modern Europe. KNOW: A Journal on the Formation of Knowledge, 5(1), 7–70. <https://doi.org/10.1086/713251>
37. Linn, R. L. (2010). Educational measurement: Overview. In P. Peterson, E. Baker, & B. McGaw (Eds.), International encyclopedia of education (3rd ed., Vol. 4, pp. 45–49). Elsevier. <https://doi.org/10.1016/B978-0-08-044894-7.00243-8>

38. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
39. Madaus, G. F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 119–154). Macmillan.
40. Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, 80(9), 688–695.
41. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.
42. Miller, G. A. (1962). *Psychology: The science of mental life*. Harper & Row.
43. Murphy, M. M. (2015). Plato's philosophy of education and the common core debate. Paper presented at the Association for the Development of Philosophy Teaching (ADOPT) Spring Conference, De Paul University, Chicago, IL, United States. <https://files.eric.ed.gov/fulltext/ED559997.pdf>
44. Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.
45. Ogunsakin, I. B., & Shogbesan, Y. O. (2018). Item response theory (IRT): A modern statistical theory for solving measurement problem in 21st century. *International Journal of Scientific Research in Education*, 11(3B), 627–635
46. Paramole, O. C., & Adeoye, M. A. (2024). Reassessing standardized tests: Evaluating their effectiveness in school performance measurement. *Curricula: Journal of Curriculum Development*, 3(2), 217–234. <https://doi.org/10.17509/curricula.v3i2.74535>
47. Pereira, Á. (2023, December). Digital assessment. In *Digital Education Outlook*. OECD Publishing. <https://doi.org/10.1787/b5dc4b5a-en>
48. Popham, W. J. (2001). Teaching to the test? *Educational Leadership*, 58(6), 16–20.
49. Rajuroy, A. (2025). The role of digital tools in implementing continuous formative assessment in mathematics education [Unpublished manuscript]. ResearchGate. https://www.researchgate.net/profile/Adam-Rajuroy/publication/392796383_The_Role_of_Digital_Tools_in_Implementing_Continuous_Formative_Assessment_in_Mathematics_Education
50. Rashdall, H. (2010). *The universities of Europe in the Middle Ages: Volume 1: Salerno, Bologna, Paris*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511722301>
51. Romanelli, F., Black, E. P., & Cain, J. (2025). Testing by talking: Rethinking oral examinations. *American Journal of Pharmaceutical Education*, 89(12), Article 101892. <https://doi.org/10.1016/j.ajpe.2025.101892>
52. Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
53. Sivakumar, A., & Thirumoorthy, G. (2019). *Measurement and evaluation in education*. A.P.H. Publishing Corporation. Retrieved from https://www.researchgate.net/publication/369479186_Measurement_and_Evaluation_in_Education
54. Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Routledge.
55. Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). Jossey-Bass.
56. Sukhera, J. (2022). Narrative reviews: Flexible, rigorous, and practical. *Journal of Graduate Medical Education*, 14(4), 414–417. <https://doi.org/10.4300/JGME-D-22-00480.1>
57. Supovitz, J. A. (2006). Joan Herman and Edward Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* [Book review]. *Journal of Educational Change*, 7(4), 351–354. <https://doi.org/10.1007/s10833-006-9005-7>
58. Upguage. (n.d.). The future of assessments: From standardized tests to skills-based evaluation. https://www.upguage.com/the_future_of_assessments
59. Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer.
60. Von Davier, M., Zhu, M., & Kyllonen, P. C. (Eds.). (2019). *Innovative assessment of collaboration*. Springer.
61. Walker, M.E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D. and Gholson, M. (2023), *Culturally Responsive Assessment: Provisional Principles*. ETS Research Report Series, 2023: 1-24. <https://doi.org/10.1002/ets2.12374>
62. Wang, K. (n.d.). *Monasteries and Universities in Middle Age*. <https://doi.org/10.3969/j.issn.1000-5218.2007.06.002>

63. Wang, N., Gao, X., Yan, L., Peng, C., Yu, Z., & Ren, X. (2026). Integrating AI-assisted formative evaluation into a full course: A case study in university-level programming education. *Education and Information Technologies*. Advance online publication. <https://doi.org/10.1007/s10639-026-13933-1>
64. Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1–27. <https://doi.org/10.2458/v2i1.12366>
65. Welsh, M. E., Eastwood, M., & D'Agostino, J. V. (2014). Conceptualizing teaching to the test under standards-based reform. *Applied Measurement in Education*, 27(2), 98–114. <https://doi.org/10.1080/08957347.2014.880439>
66. Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation*, 23(1), 31–48. [https://doi.org/10.1016/S0191-491X\(96\)00024-3](https://doi.org/10.1016/S0191-491X(96)00024-3)
67. Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge. <https://doi.org/10.4324/9781003286929>
68. Wise, L. L., & Rutstein, D. (2024). *Educational Measurement: Models, Methods, and Theory*. *Educational Measurement: Issues and Practice*. <https://doi.org/10.1111/emip.12642>
69. Yousaf, M., Umber, S., & Imran, M. (2025). A comparative study of ancient and contemporary educational philosophies. *Social Sciences & Humanity Research Review*, 3(2), Article 016. <https://doi.org/10.63468/sshrr.016>
70. Zanon, C., Hutz, C., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29, Article 18. <https://doi.org/10.1186/s41155-016-0040-x>
71. Zhao, J. (2023). The influence of the imperial examination system reform in the Northern Song Dynasty on the scholars. *SHS Web of Conferences*, 179, Article 01012. <https://doi.org/10.1051/shsconf/202317901012>