International Journal of Assessment Tools in Education

2026, Vol. 13, No. 1, 205–223

https://doi.org/10.21449/ijate.1661896

journal homepage: https://dergipark.org.tr/en/pub/ijate

Review Article

From measurement to meaning: Reconsidering test-based accountability and educational quality

Zeynep Teke[1], Özen Yıldırım[1*], Zeliha Selçuk[1]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

ARTICLE HISTORY

Received: Mar. 20, 2025

Accepted: Nov. 21, 2025

KEYWORDS

Test based accountability,

Educational accountability,

Large-scale assessment,

PISA.

Abstract: This paper systematically reviews and synthesizes the existing literature on the effects of test-based accountability systems on educational quality, student achievement, and instructional practices. The study defines the theoretical framework of educational accountability, explains the functioning of various accountability models, and discusses the role of international large-scale assessments, particularly PISA, within these processes. Findings from the literature indicate that test-based accountability has contributed to greater transparency and measurability in education systems and has stimulated student achievement. However, it has also raised significant concerns, including curriculum narrowing, teaching to the test, increased pressure on teachers, and the exacerbation of educational inequalities. Moreover, many studies emphasize that test results are strongly influenced by socioeconomic factors, which limits their validity as measures of school and teacher effectiveness. In this regard, adopting value-added accountability models may offer a more balanced approach by supporting intellectual development, creativity, and student-centered learning. Equally important is the active involvement of teachers in decision-making processes, which can reduce workload-related stress while enhancing instructional quality. In conclusion, this review consolidates fragmented research to provide a comprehensive understanding of test-based accountability. It highlights both the strengths and limitations of such systems and offers recommendations for policymakers and practitioners seeking to design more balanced, inclusive, and sustainable accountability frameworks in education.

INTRODUCTION

The concept of accountability, which encompasses the responsibilities toward oneself and the institution one works for to meet quality standards and restructure policies, has become widely used internationally in recent years and has started to be implemented within education systems (Ambrosio, 2013; Hanushek & Raymond, 2005). As in every field of education, accountability has come to the forefront in the assessment and evaluation process. This is because education policies and related practices are shaped by the assessment and evaluation approaches accepted in education, and it is expected that these regulations will enhance learning, which is the most important outcome of education, qualitatively. In the measurement and evaluation process, accountability refers to assessing whether stakeholders in the education system (teachers,

*Corresponding author at: Özen YILDIRIM ozenyildirim@pau.edu.tr Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

e-ISSN: 2148-7456

administrators, schools, educational institutions, parents, and the government) fulfill their responsibilities in achieving educational goals and creating a transparent process for this purpose.

Addressing educational inequalities and making student achievement tangible and measurable became a prominent focus in educational policies in the early 2000s. With these developments, large-scale assessments (PISA, TIMSS, PIRLS, etc.) began to spread, and test-based accountability emerged as a reflection of learning achievements. Test-based accountability systems, which typically link the evaluation of schools, teachers, and the education system's performance to scores obtained from standardized achievement tests, offer features such as ensuring transparency in education, improving student achievement, and making schools more effective (Bozkurt & Karabıyık, 2003). This system has advantages as well as certain limitations, such as the pressure placed on schools, curriculum narrowing, and the neglect of important skills (e.g., creative, critical thinking) (Hanushek & Raymond, 2005). This system implies expecting all students to learn the same outcomes within the same period, standardizing school experiences, and groups. Given that a number of factors affect the learning and teaching process, standardized test results should not be the sole basis for accountability. Although international large-scale assessments are used for accountability in several countries, it is expected that systematic accountability mechanisms be established based on these results to evaluate schools, teachers, or education administrators. In accountability practices, national-level or teacher-made classroom assessments, which primarily measure individual student achievement but are insufficient in evaluating the entire system, are commonly used. Despite the growing interest in international large-scale assessments and national testing practices, there remains a gap in understanding how these measures function to evaluate not only individual student achievement but also systemic performance and accountability. Research on educational accountability indicates that without carefully designed mechanisms, test-based systems may inadvertently reinforce inequalities and fail to reflect broader educational goals (Torres, 2021). Without careful design, these systems may unintentionally reinforce existing educational inequalities, as socio-economic background and contextual factors strongly influence student performance (Tansel, 2015). This underscores the importance of developing accountability mechanisms that not only measure learning outcomes but also ensure fairness, transparency, and support for all stakeholders in the education system. The current study addresses this gap by examining the types, purposes, and effects of test-based accountability systems, evaluating their positive and negative contributions, and analyzing the implementation of existing practices. This research contributes to the broader discourse on educational assessment and informs policymakers, practitioners, and researchers about the potential and limitations of test-based accountability in promoting educational quality.

METHOD

The study employs a qualitative research approach designed as a comprehensive literature review. Review studies aim to synthesize existing ideas, approaches, and findings within a specific research domain rather than directly testing new research hypotheses (Aydoğdu *et al.*, 2017). Such studies contribute to the field by summarizing core arguments from existing research and generating meaningful syntheses (Herdman, 2006). Study systematically examines information, perspectives, and approaches from diverse sources concerning accountability, educational accountability, and test-based accountability, organizing them into a coherent framework with relevant examples.

The literature review was conducted between September 2024 and December 2024 using Google Scholar, EBSCO, Web of Science, ERIC, and SCOPUS databases. The search included sources in both Turkish and English, incorporating only openly accessible academic articles, review studies, books, book chapters, national and international reports, legislative documents, and educational curricula. Boolean operators were utilized in keyword selection, with terms such as "large-scale testing" OR "large-scale assessment," "educational accountability","geniş ölçekli testler," OR "geniş ölçekli değerlendirmeler," OR "geniş ölçekli uygulamalar", “accountability” AND large-scale assessment OR large-scale testing”. To provide comparative examples of accountability models, additional searches were performed using keywords like "state assessment" and "state accountability" to examine practices implemented across U.S. states. A total of 79 sources were examined in detail for this research, including 44 articles (peer-reviewed journal: review or research), 10 books, 3 book chapters, 20 reports published by an institution or organization, 1 internet-based publication, and 1 doctoral dissertation. (for details, see the references).

For the analysis of English-language sources, translation support was obtained from Google Translate and AI tools such as ChatGPT and DeepSeek.AI-assisted tools were used for translation and summarization purposes; however, all interpretations and analyses were conducted by the authors. Retrieved studies were categorized and synthesized using a descriptive analytical approach. In the development of the article, the concept of accountability was first defined, followed by an examination of educational accountability and test-based accountability models, discussing their impacts on educational policies and particularly on curricula.

RESULTS

In this section, the findings derived from the literature on test-based accountability systems are presented within a comprehensive framework. The functioning of various accountability models, their implications for educational policy and school practices, as well as the influence of national and international large-scale assessments on these processes, are examined under the relevant subheadings. The findings reveal both the strengths and limitations of current practices and contribute to the ongoing scholarly discussion on accountability in education.

Accountability

The concept of accountability has been defined differently by various researchers. According to the Organization for Economic Co-operation and Development (OECD, 2010), accountability is a structure that requires comparing performance outcomes against clearly defined goals at different levels. Acar (2013) defines accountability as “the obligation to provide information and justification to the public and society regarding decisions, actions, and processes which concern and affect the public.” (pp. 388). Similarly, Bilgin (2007) describes accountability as the responsibility of an individual or institution to provide necessary or requested explanations to the target audience they serve regarding their decisions and actions.

Accountability can be categorized as internal or external. External accountability involves obligations toward elected officials and judicial authorities, while internal accountability pertains to reporting and justification within organizational hierarchies (Linn, 2005). Regardless of type, any accountability system must address two central questions: what is being accounted for and who is responsible. Effective systems require comprehensive, multidimensional evaluations that consider both contextual and process factors, ensuring a balanced focus across objectives, accurate interpretation of outcomes, and identification of areas for improvement (Linn, 2005).

Educational Accountability

Educational accountability, which plays a role in presenting the quality of education services in educational institutions for public and societal approval (Balcı, 2014), creates an area of responsibility for everyone, from policymakers to teachers (Hesapçıoğlu *et al*., 2001). The concept of educational accountability, which includes all developments that can be implemented to improve the quality of teaching, initially focused on physical aspects such as teachers, textbooks, and class size in the early 1970s, but by the 1980s, the concept was expanded to include teacher competencies and their regulation. The educational paradigm has undergone a fundamental transformation since the 1970s, shifting from behaviorist approaches (Skinner, 1953) toward constructivism. The cognitive and sociocultural theories developed by Piaget (1970) and Vygotsky (1978) reconceptualized learning as an active process of meaning construction rather than passive knowledge acquisition. This transformation was further reinforced in the 1980s by Gardner's (1983) Theory of Multiple Intelligences, which emphasized the need for individualized instructional methods, while Freire's (1970) critical pedagogy approach highlighted education's role in social transformation. Consequently, there has been a transition from teacher-centered, standardized models to student-centered, contextual, and pluralistic pedagogies. Today, under the framework of educational accountability, student learning has been emphasized (Watts *et al*., 1998). The use of accountability practices in education systems has increasingly been adopted, finding a greater place not only in developed countries but also in developing ones (Phelps, 2000). This widespread adoption occurs primarily through two mechanisms: 1) the implementation of national high-stakes testing linked to performance sanctions/rewards (Linn, 2005; Nichols & Berliner, 2007), and 2) the increasing influence of international large-scale assessments, such as PISA, which drive national policy agendas by facilitating cross-country comparisons (Froese-Germain, 2010; OECD, 2014).

Educational accountability is a system in which input, process, and outcome data are systematically collected and used to make decisions about the effectiveness of schools, regions, or countries (Daigle & Cuocco, 2002). The goals of educational accountability are to measure progress toward general and local objectives, highlight performance differences between student groups, and identify schools in need of support and additional resources (Erwin & Bloomquist, 2024).

Content and student achievement standards are the two most significant factors in a good school accountability system. While content standards define what students need to learn, achievement standards define how much students need to learn. Educational accountability should be explanatory and guide what students need to learn and how much throughout their educational life (Watts *et al*., 1998). The necessity of explicitly defined standards stems from a clear policy lever for improvement and establishing the legitimacy of measurement. Content standards serve as a fundamental blueprint, ensuring that large-scale assessments align directly with instructional practices and curricular goals (Anderson, 2005; Popham, 2007). Furthermore, the achievement standards translate into tangible, measurable benchmarks, enabling stakeholders to evaluate school effectiveness fairly (Linn, 2005). Thus, standards are not only directional but are integral to the validity and utility of the entire accountability.

Three primary accountability system types in education are most widely accepted (Curran, 1999, as cited in Anderson, 2005):

1. Compliance with regulations,
2. Adherence to professional norms,
3. Results-driven.

In the first system, compliance with regulations, educators focus on adherence to laws and bureaucracy. While it is not a flexible structure, unrealistic goals can be challenging for teachers, students, and parents. One of the first examples of educational accountability, the No Child Left Behind (NCLB) Act, was created in 2001 after the results of the Trends in International Mathematics and Science Study (TIMSS) exam, administered to 3rd graders in the United States in 1995, were lower than expected. It was decided that students from grades 3 to 8 would be tested annually and at least once in high school. The goal was to bring all students to a predetermined proficiency level by the 2013-14 school year, improve low-performing schools, and reward high-performing schools. However, it directed teachers and schools towards a test-based system (Dee & Jacob, 2011; Hanushek & Raymond, 2005). Some of the issues discussed regarding this law include the demotivating effect of a punitive approach towards schools that did not show improvement, the tendency of schools to ignore disadvantaged students due to the fear of low performance, and the resulting increase in dropout rates (Anderson, 2005).

The second system, adherence to professional norms, focuses on certain principles and practices that view education as a profession. In this system, educators and institutions are responsible for adhering to standards, but these standards can be set by themselves. To diminish the negative effects of the NCLB Act, the Every Student Succeeds Act (ESSA), which is more flexible, came into effect in 2015 (Cook-Harvey *et al*., 2016). According to this act, states were able to create their own accountability frameworks, focusing more on process achievement rather than a test-based system. In Türkiye, with the Regulation on the Appointment and Relocation of Teachers by the Ministry of National Education 2015, the authority to inspect lessons was transferred from inspectors to school principals, which exemplifies the flexibility of the accountability structure. However, the possibility of a lack of objectivity among school administrators and the reflection of personal and political views in their evaluations of teacher performance has been criticized. As a result of individual and union objections, the performance scores set by school principals were disregarded (Özdemir *et al*., 2018), and this process can be seen as an example of legal accountability.

The third system focuses on the outcomes and performance of students during the learning process. The goal is not to reduce administrative burdens but to make improvements based on analyzing student performance results. The communication of results to parents and the public is emphasized (Anderson, 2005). The three main principles of this system are: 1) Setting achievement standards and ensuring adherence to them, 2) Ensuring these standards guide all educational practices, and 3) Recognizing that all students can meet these standards. Schools are held responsible for student performance, but the responsibility here is to provide support to help students do their best. One of the key foundations of this system is the necessity of providing life skills that will be needed after education (Furman, 1994).

Currently, accountability systems focus more on results-driven (the third system) rather than compliance with regulations and adherence to professional norms. That is to say, accountability systems are focusing less on compliance and more on outcomes, and the focus is on 'student learning' (Anderson, 2005; Watts *et al*., 1998). As a result, governments and policymakers have increasingly begun to pay more attention to measuring and evaluating students, teachers, schools, and education systems. The results of these evaluations serve to identify students' learning progress, provide information to parents and society about the current situation, and improve the performance of schools and systems (Erwin & Bloomquist, 2024). According to Reeves (2004), evaluations created with student-centered accountability systems provide more precise and accurate information while playing a role in increasing student motivation, and these evaluations are critical for improving student performance. However, a point of criticism is that these systems focus on outcomes rather than the process (Hanushek & Raymond, 2005). An example of the results driven system is the Australian National Education Performance Monitoring Taskforce (Australian NEPMT). In this system, educators are required to be accountable for student learning to the public (Anderson, 2005). Similarly, countries such as the United Kingdom, France, China, and Japan have used national exam results to measure student and school progress (Leckie & Goldstein, 2018). However, one issue they overlooked is that these test scores are not only dependent on the educational programs that children are exposed to but also on students' socio-economic backgrounds (Anderson, 2005; Popham, 2007). For instance, Leckie and Goldstein (2019) argue that accountability systems that ignore pupil background may unfairly reward or penalize schools, since socio-economic status (SES) strongly predicts achievement even when prior attainment is controlled. OECD research shows that higher accountability is associated with greater performance gains but also with increased inequality, as disadvantaged students do not benefit equally from accountability mechanisms (Torres, 2021). Furthermore, Lane (2020) cautions that test-based accountability systems can produce unintended negative consequences for low-SES student groups if fairness and validity of use are not explicitly addressed.

Test-Based Accountability

In recent years, one of the significant indicators for countries to achieve their development-oriented strategic goals has been education. It is possible to predict a country's future based on the reforms in education and actual student achievement rates (Mallick *et al*.,2016). For this reason, there is a need for measurement and evaluation systems to assess the quality of learning experiences and provide feedback to stakeholders based on this assessment (Ülkü, 2023).

Measurement and evaluation are inseparable parts of educational systems. To determine the quality and quantity of the educational services provided and to guide educational policies, measurement and evaluation processes are crucial. Measurement and evaluation are defined as comparing the goal to be achieved with student outcomes (Özçelik, 2016) and making an evaluation by expressing the results in quantitative and qualitative terms (Turgut & Baykul, 2010).

In education systems, tests that are developed through test development processes to measure students' achievement levels are called Standardized Achievement Tests, while tests prepared by teachers to measure specific subjects within the classroom are referred to as Teacher-Made Tests. Teacher-made tests help determine a student's status in the classroom, while standardized achievement tests provide the opportunity to determine the student's status across the country (Çağlar & Kılıç, 2017). Tests that are more comprehensive than standardized tests and are used to measure the performance of education systems and shape educational policies are called large-scale assessments. These tests are conducted by national and international organizations to provide feedback on the development and improvement of countries' education systems. The most well-known ones include PISA (Programme for International Student Assessment), PIAAC (Programme for the International Assessment of Adult Competencies), TIMSS (Trends in International Mathematics and Science Study), and PIRLS (Progress in International Reading Literacy Study).

Standardized achievement tests, which have started to be seen as a concrete indicator of the return on investments in education (Watts *et al*., 1998), are frequently used to show the results of the development of the education system to officials and the public. Based on these results, necessary teaching methods and the allocation of resources can be planned to improve student achievement. For instance, if more than 90% of the fourth-grade students in one region of the country have achieved the reading comprehension achievement standard and another region falls short of the standard, accountability based on test results can help plan the resources that need to be directed to that region. In other words, accountability based on test results reflects whether students in schools meet the learning standards, and schools and students who fall behind these standards are held accountable (Haertel & Herman, 2005). Evaluating and interpreting the results of such tests in terms of accountability is relatively easy and can be seen as the only way to determine if a student has learned the information (Hamilton *et al*., 2002).

Test-based accountability has gained widespread support as a method for improving education systems and is based on the belief that by requiring all students to take standardized achievement tests, rewards can be given when test scores increase, and sanctions can be applied when expected progress is not achieved (Rosenblatt, 2017). For years, these tests have been used solely to measure existing performance, but in recent years, there has been a significant increase in the scope of these tests and the number of purposes they serve. In addition to their measurement function, they have become a cornerstone of efforts to restructure education (Hamilton *et al*., 2002).

Test-based accountability systems are implemented in most OECD countries (Verger *et al*., 2021). Advocates of these systems believe that the use of standardized achievement tests will bring about positive changes in schools and classrooms (Ambrosio, 2013). In countries using this system to improve the quality of education, data on student achievement, which is collected annually, is compared to set goals. Schools that meet these goals receive prestige and financial rewards while schools that do not meet the goals face severe sanctions and enter a reconstruction process (Kalman & Gedikoğlu, 2014; Leckie & Prior, 2022). Although test-based accountability may seem simple, it actually has a more complex structure. These systems do not directly measure performance in non-tested subjects or grade levels, nor do they measure the development of affective domains like interest and curiosity. Furthermore, they fall short in measuring performance in the subjects and grade levels tested. In addition, they can only measure a subset of the content of a subject. Kalman and Gedikoğlu (2014) criticize the traditional accountability system and suggest that the results of standardized achievement tests do not truly reflect a student's performance and should express more. He argues that an accountability system, which includes a holistic view of student performance and transparency, may be emphasized.

Nichols and Berliner (2007) argue that the negative effects of standardized achievement tests narrow and corrupt education as a whole. These researchers call for a review of our fundamental understandings of education, measurement, and evaluation. In this context, an ideal test-based accountability system should be one that encourages development, supports educators, and empowers students. Test-based accountability should focus on effective assessments, reduce sanctions and pressures, promote equity, and emphasize continuous improvement, turning it into a tool for strengthening teaching rather than a source of anxiety, encouraging collaboration and innovation, and ensuring that every student reaches their potential. These aspects represent a balanced approach to the design and implementation of test-based accountability systems. A carefully planned accountability system, adhering to the principle of equity, can help create effective and inclusive educational environments.

Across the world, education systems hold schools accountable for student performance using different accountability models based on their objectives. These models aim to measure the impact of schools on student learning. To illustrate, in the United States, the ESSA, which was reapproved in 2001, made test-based accountability mandatory for all states. The law aims to ensure that all students have an equal opportunity to receive a quality education and meet at least the minimum standards of state academic achievement (Adequate Yearly Progress [AYP]) (Anderson, 2005; Hamilton *et al*., 2002). Through the tests administered, the performance of students, teachers, and schools is determined, and the tests have shifted from being a formative assessment tool (Stiggins, 2005). This is because teachers have focused on state achievement standards, setting aside creative lessons, techniques, and methods (Gallagher, 2007). In Australia, the national accountability system was created with NAP (National Assessment Program) and NAPLAN (Australian National Assessment Program Literacy and Numeracy) exams administered by ACARA (The Australian Curriculum, Assessment, and Reporting Authority). These exams aim to measure the basic skills of students in reading, writing, grammar, and mathematics at the 3rd, 5th, 7th, and 9th grades (Rudd & Gillard, 2008). The federal government views this program as an element supporting quality education by promoting accountability and transparency, with NAPLAN results published on the MySchool website. This allows comparisons between schools, and by comparing NAPLAN results over the years, parents can choose schools in the same region (Thompson, 2013). While providing transparency for families and enabling school choice, has this system had the same positive effect on schools? Publishing student achievement scores on a website for parents and the public may be a high-risk practice for schools, creating pressure on them.

The importance given to test-based accountability raises several important questions (Hamilton *et al*., 2002):

- Do these tests accurately assess student achievement?

- How can policymakers and educators select the right tests?
- How should test results be properly evaluated?
- How can the results of these tests be utilized for the accountability systems to work as intended?
- What are the costs of developing and implementing these tests?
- And what compromises do these policies bring?

Test-based assessment results are inadequate in predicting student achievement in several respects. A test only covers a small sample of the broad subject area it aims to measure and is limited to multiple-choice or short-answer questions. In other words, only a small representation of the area being measured is included in the tests. Additionally, only critical goals considered essential in education are covered by the tests. For instance, while mathematics and science are measured, music and computer skills are not included (National Research Council [NRC], 2010). Furthermore, affective domains such as curiosity and motivation cannot be measured within the scope of these tests.

The careful planning of cost-effective and content-standard-compliant assessments can strengthen the support of stakeholders, such as policymakers and the public. Policymakers should be careful to choose tests that are fair and non-biased while considering the reliability and validity values of the tests they use for accountability (Watts *et al*., 1998). The validity of a test refers to whether the test measures the information and skills intended, and its reliability depends on the consistency of the scores it produces. The fairness or bias of a test can be explained as ensuring that it does not disadvantage variables like race, ethnic group, or gender.

In the U.S. education system, mathematics achievement as measured by standardized achievement tests is regarded as one of the strongest predictors of educational attainment (Joensen & Nielsen, 2009). Success during high school is associated not only with university admission and graduation (Gaertner *et al*., 2014; Kim *et al*., 2015) but also with broader life outcomes (Cole *et al*., 2016; Montez *et al*., 2012; Woods *et al*., 2018). Consequently, educational policies have primarily focused on raising test scores in standardized assessments and enhancing mathematics performance (Harland *et al*., 2000; Parsad, 2012; Sabol, 2013). These policy initiatives have substantially reduced instructional time allocated to music and art classes (Dee & Jacob, 2011), reallocating greater time to language, science, and mathematics (Harland *et al*., 2000; Parsad, 2012; Sabol, 2013).

Despite these reforms, the U.S. continues to lag behind many countries in mathematics performance (DeSilver, 2017). According to the National Assessment Governing Board, reading scores for 4th and 8th grades have remained stagnant for nearly a decade, while mathematics scores have shown promising gains, particularly among students in the lowest decile (Popham, 2016). Similarly, an analysis of NAPLAN results revealed statistically significant increases in 3rd and 5th-grade reading skills and 5th-grade mathematics skills, though no significant improvements were observed in other areas (Thompson, 2013). Moreover, socioeconomic status (SES)-based disparities persist (Chiu, 2010; Gonzalez & Kuenzi, 2012).

Given such modest improvements, should the cultivation of 21st-century skills fostered through artistic domains such as visual arts, music, and drama be compromised? Research demonstrates that arts education enhances students' cognitive process skills (Hetland *et al*., 2007), contributes to their academic achievement (Catterall, 2012; Hetland & Winner, 2001; Sheridan *et al*., 2022; Winner *et al*., 2020), and reduces dropout rates (Thomas, Singh, & Klopfenstein, 2015). Arts education not only develops problem-solving, observational, and critical thinking skills that support learning in other disciplines but also strengthens students' cognitive processing and meaning-making capacity (Eisner, 2002). Art, fundamentally, is a medium for self-expression, identity formation, and interpretation of the world. Through visual culture, metaphors, and narratives, the mind is continuously reconstructed (Eisner, 2002; Rolling, 2006).

Just as “physical education” is not mandated because of its contribution to science but because it is essential for healthy individuals, arts education should also be recognized as a fundamental component of schooling. It ensures students’ holistic development while encompassing cultural, cognitive, and human values (Hetland & Winner, 2001).

Test-Based Accountability Models in Education

Although there are various definitions of educational accountability in the literature, the goals of improving quality in education and converting educational investments into academic achievement are common. Different accountability models serve these purposes. For test-based educational accountability, models such as the Status Model, Cohort Model, Growth Model, and Value-Added Model are used, each designed to answer a different set of policy-related questions. The most commonly used models are listed below (NRC, 2010);

Status Model: This model is based on measurements taken at a specific time and is generally compared with a pre-determined goal. To illustrate, the average test score of a student group or a school is compared with the country's predefined goal. This model answers questions like, "What percentage of students in the country reached a sufficient level of achievement this year?" or "Did the mentioned school meet the country's proficiency goal this year?" The functioning of the No Child Left Behind (NCLB) Act in the United States is an example of this model. A standard is set for each state, and student performances are compared with these standards, leading to sanctions or rewards for schools (NRC, 2010). For instance, the accountability system in Texas is considered an example of this model. Assessments are used to determine whether state standards have been met, to rate schools, and to apply sanctions or rewards (Gong, 2002). The fear of sanctions can lead teachers and administrators to shift from the "learning for life" context to the "learning for tests" context (NRC, 2010).

Cohort Model: This model compares the test results of students in the same grade level at different times to determine teacher, school, or national achievement. For example, the percentage of fourth-grade students' reading proficiency this year can be compared with the proficiency percentage of the same grade level students from last year. This model answers the question, "Are students in a specific grade level performing better this year compared to students at the same grade level from last year?" (NRC 2010). States like Kentucky, Louisiana, and California have state accountability systems based on the cohort model. In schools, achievement is expected to increase over time within groups. For instance, the performance of fourth-grade students in the third period is expected to be better than the performance of fourth-grade students in the first period. School ratings are determined by comparing the expected improvement with the improvement actually demonstrated by schools. Schools showing the expected growth are rewarded, while those failing to show the expected growth are labeled as low-performing (Gong, 2002).

Growth Model: This model tracks the year-to-year changes in the achievement of the same students to determine the magnitude of their progress. Achievement scores can be determined by comparing the performance of current fourth-grade students with the performance of the same group of students when they were in third grade the previous year. This model is preferred by those who want to answer the question, "How much did students' performance change on average between Grade X and Grade Y?" There can also be a nationwide growth target that subgroups or school systems need to meet. Accountability systems based on this model credit teachers and schools for student growth, often ignoring students' starting levels. These systems usually do not account for background factors such as the students' or schools' histories, and therefore do not try to address which factors are responsible for student growth (NRC, 2010). North Carolina uses both the status model and the growth model in its school accountability system. For example, both whether students have reached state proficiency levels and whether a student who started with a reading score of 3 in Grade 5 and progressed to a score of 4 by the end of Grade 5 are evaluated, with the student gaining credit points. Schools are held responsible for achieving at least one grade level of growth per year for their students (Gong, 2002).

Value-Added Model: This model is more complex and statistical compared to others. It attempts to determine the contributions of the school, teachers, or educational programs to student achievement. These models answer questions like, "How did a school or teacher contribute to student development compared to the average impact of schools or teachers?" or "How much of the change in student performance can be attributed to a school or a teacher's classroom?" To control for school or teacher effects, at least two years of student test results, and sometimes other student and school variables such as socio-economic status, family background, or school management quality, are considered. In some models, the evaluation of a school or teacher is based on the difference between students' observed progress and the expected progress (NRC, 2010).

In educational systems, accountability models can be used either separately or together, depending on the questions being addressed. For instance, while status models are suitable for understanding the proficiency level of students in a specific school in a given year, cohort models may provide more meaningful results for tracking a school's progress over time (CCSSO, 2008). However, neither the status model nor the cohort model provides the necessary information for comparing the effectiveness of teachers or teaching practices within or between schools. For example, in a school with a high proportion of advantaged students, teachers' contributions to achieving the achievement standard may be minimal, whereas in a school with a high proportion of disadvantaged students that fails to meet the achievement standard, the teacher's impact might result in better progress. In other words, both models fail to provide insights into the true factors contributing to changes in school achievement (NRC, 2010). In value-added models, more advanced statistical methods are used to identify factors influencing performance. While the status model shows a student's proficiency level at a specific point in time, the value-added model reflects how much progress has been made compared to the previous year. In Ohio, the accountability system was changed to a value-added model in 2003, acknowledging that even schools and districts that failed to meet the achievement standard could show significant academic progress in their students' performance. Reading and math performance in grades 4 and 8 in Ohio were described according to the value-added model, and the results influenced the region's or school's rating (5-Excellent, 1-Immediate intervention). For example, when a district or school achieved greater growth than expected, its rating could increase by one category, while schools showing below-expectation growth could see their rating drop (Ohio Department of Education, 2009). In Türkiye, however, there is no official school rating system. The Ministry of National Education (MONE) conducts central exams to monitor student performance, but these exam results are not used for school rating purposes.

Large-Scale Assessment: PISA as an Illustrative Case and Its Effects on Accountability

Standardized achievement tests and large-scale assessments are now considered an essential part of the educational assessment process and widely used to determine students' progress in learning. The administration of large-scale assessment and using their results to monitor student achievement provides data on the effectiveness of educational reforms and schools in need of support (Yıldırım, 2021). It is obvious that national and international large-scale assessments, which have a wide range of applications from primary to higher education, are taken into consideration while making important decisions on behalf of countries' education systems (UNESCO Institute for Statistics, 2018). The widespread adoption of large-scale assessment in countries is attributed to the promise of accountability and the assumption that the data it provides on student achievement is more reliable and valid (Volante, 2006).

One of the large-scale assessments guiding education policies is the PISA conducted by OECD, which measures the reading, mathematics, and science literacy skills of 15-year-old students while also collecting information on social, cultural, economic, and educational factors related to student performance. The results of PISA tests and surveys are analyzed at national and international levels, allowing the collection of evidence for reconstructing education policies and developing new ones. PISA directly shapes education policies by redefining national curriculum standards and learning outcomes (Dee & Jacob, 2011). Policymakers transform the "proficiency" levels identified in assessments like PISA into national benchmarks, and these data simultaneously form the basis of accountability systems used to monitor school and teacher performance. Moreover, through the concept of the assessed curriculum, such evaluations not only determine what is measured but also influence what is taught, thereby shifting the curricula of many countries from a focus on theoretical knowledge toward the development of higher-order skills such as critical thinking and problem solving (UNESCO Institute for Statistics, 2018).

The primary factor that led the OECD to develop PISA in the late 1990s was to help shape scientific education policies through scientific research. The PISA mechanism was intended to enable schools, teachers, and policymakers to assess their own performance rather than a hierarchical accountability system. PISA aimed to enable schools, teachers, and policymakers to assess their own performance. PISA evaluates the collected data within the context of social outcomes and shares it to help educators and policymakers make better decisions (Schleicher, 2021). PISA is designed to help governments develop policies and is included in the OECD indicators. It also examines the quality of learning outcomes, the policy instruments, and contextual factors that shape these outcomes, as well as the returns on investment in education. Covering nearly nine-tenths of the world economy, PISA includes all 30 OECD member countries and more than 60 participating nations, allowing them to systematically monitor their progress in achieving key learning goals. Thanks to the policy-oriented aspect of PISA, its connection to OECD education indicators, and its other features, it is a highly effective instrument in guiding education policies in OECD countries and others (UNESCO Institute for Statistics, 2018).

PISA has an important place in monitoring and assessing the educational performance of countries. It has an increasing impact on policy-making and development at both national and local levels, but its impact at the national level is greater than at the local level, while it is comparatively less on school practices and teaching. To enhance PISA's effectiveness at all levels, certain improvements can be made. One of these is sharing PISA results more strategically and in a coordinated manner. Participating countries should utilize these results more effectively and, in those participating countries, integrate them more clearly into their national assessment strategies. Additionally, increasing support for education stakeholders to design and implement improvement strategies based on these results is another critical factor.

OECD reports indicate that PISA data are utilized in many OECD member countries to evaluate the performance of education systems and to inform policy development processes (OECD, 2015). Countries that rank higher in PISA tend to use these results less frequently for policy initiatives, whereas those with lower rankings tend to undertake more policy interventions and reforms (OECD, 2014). The level of incorporation of PISA results into national assessment systems varies considerably across countries. In essence, the lower a country performs in PISA, the greater its awareness of educational issues, leading to more policy interventions and reforms.

Countries that take in policy initiatives based on PISA results advocate for their educational reforms by referring to the performance data obtained from PISA. Policymakers and the media play a significant role within the scope of PISA results. The more they emphasize these results, the greater the awareness is raised at other levels of the education system. Alongside policymakers and the media, academics and researchers are also stakeholder groups with a high awareness of PISA. Participating countries view PISA as a remarkable tool for evaluating school system performance, recognizing it as an accountability mechanism. It is highly reliable research in a legal sense as well. The accuracy, consistency, and scientific methodology underlying the collected data ensure that PISA results are reliable and influential, positioning PISA strategically in determining educational policies, reform processes, national assessment systems, and accountability systems.

Torres (2021) conducted a study using data from the 2006-2015 PISA cycles to examine to what extent test-based accountability practices impact academic achievement and equity in high-, middle-, and low-income countries. Although there was no definitive evidence of the impact of accountability practices on educational outcomes in high-income OECD countries, accountability systems conducted in low- and middle-income OECD countries were associated with increased achievement in math and science, along with increasing inequality. This indicates that test-based accountability systems may enhance overall school performance in these countries but may also be accompanied by inequality. In other words, some schools or students benefit more from these systems while others fall behind. The study's findings demonstrate that test-based accountability practices in high-income countries do not directly impact academic achievement, fail to reduce educational inequalities, and do not provide the expected improvements in schools. Moreover, it may lead to unintended negative consequences, such as competition between schools. Test-based accountability has not yielded remarkable results in providing the advantages expected and reducing inequalities in high-income countries. Test-based accountability has been found to have positive effects on mathematics and science in low- and middle-income countries while increasing educational inequality. Due to some limitations on sample size, broader research and case studies are necessary to explore this issue further.

Since the first cycle of PISA in 2000, it has triggered worldwide debates. Some countries that believed they had invested sufficiently in education found themselves in what is known as the PISA shock when confronted with PISA results. Germany was one of the first countries to experience this shock after the results, as it was widely accepted by the public and policymakers that schools were provided with the necessary resources and opportunities. In contrast to Finland, where school differences accounted for only 5% of the variance, and there was evidence of consistency across schools, in Germany, this figure was around 50%, revealing significant differences between schools through PISA results. Based on the PISA results, in Finland, the top-performing country, there was consistency across schools, and students were given equal opportunities no matter which school they enrolled in, whereas the same was not the case in German schools. In the German system, students are assessed for academic proficiency after the age of ten; students with academic deficiencies are directed to different professions. Generally, students with a higher socio-economic background were directed towards academic life, while those with a lower socio-economic background were directed towards lower professions, which was thought to be the main reason for the performance difference between schools. This caused a class-based difference and inequality of opportunity in the education system. PISA results highlighted the significance of socioeconomic background on student performance, and this prompted several countries to take steps toward reducing inequality. Although Germany was left behind in this respect, PISA findings ultimately led to significant educational reforms and transformations in the German education system (Schleicher, 2021).

Germany implemented comprehensive reforms by analyzing its unexpected performance in the PISA 2000 and 2003 results. Yalçın (2011) included these reforms in his study. In this regard, the reasons for the low performance were investigated with scientific methods, and the federal government, state governments, education unions, parent and teacher associations, scientists, and all other stakeholders collaborated to improve performance. The education systems of successful PISA countries such as Finland, Japan, and South Korea were reviewed. Each of the schools made evaluations on what kind of reforms could be made within their own teachers' council, and all schools shared their findings with each other.

On the one hand, in-service training programs were provided for teachers, and on the other hand, scientific committees were established by the Ministry of Education to solve the problem. In addition, experts were sent to high-performing countries like Finland to review their systems. With the contributions of universities, researchers, and published reports and studies, new teaching methods and approaches have been adopted. All stakeholders have been encouraged to participate in the effort by addressing education thoroughly. Teacher training programs have made some changes in their programs, and primary schools have switched to full-day education, with playgrounds being provided in these schools. The innovations in Germany have not been limited to local students; specific steps and practices have also been carried out for migrant students. For instance, language education has started to be provided in kindergartens, and to eliminate educational inequalities for migrant students, teachers who speak both the student's native language and German have been appointed in the schools where these students are enrolled. As a result of these reforms, Germany has taken PISA performance beyond the OECD average at the end of 2009. This is a sign that institutions can develop sustainable policies by applying set standards and relying on scientific research and solutions.

Türkiye has not attained its desired performance since first participating in PISA; however, the results may have played a role in shaping the country’s education policies over time. The critical issue is how these results are related to Türkiye’s accountability framework. Züftüoglu and Kocak (2017) suggest that Turkish policymakers have increasingly used PISA outcomes to justify reforms, but accountability mechanisms remain constrained by inequities. Tansel (2015) shows that socioeconomic background remains a strong determinant of students' PISA scores in Türkiye, indicating that accountability practices tied to these assessments are vulnerable to reinforcing existing inequities rather than alleviating them. There are also recommendations for Türkiye to consider the outcomes of this assessment when making decisions in its education policies. For instance, TÜSİAD and TÖDER (2012) released a policy document entitled “PISA 2012 Evaluation: Data Driven Education Reform Proposals for Türkiye”, which explicitly proposes reform measures based on PISA data.

DISCUSSION and CONCLUSION

Accountability refers to the responsibility of institutions or individuals to the public for the decisions they make and actions they take. An effective accountability system should be built on a balanced structure where multidimensional evaluations are conducted. Educational accountability is a multidimensional structure with educational policies and classroom practices. This system holds all education stakeholders accountable to the public for the quality of educational services. An effective school accountability system should have content and student achievement standards. The emphasis on evaluating education systems has increased, aiming to ensure equal opportunities and to continuously improve the system. However, adopting an only results-driven approach hinders reflecting the educational process as a whole.

Standardized tests are used to determine student achievement levels and guide educational policies. In this sense, test-based accountability aims to improve student achievement and the performance of education systems. While standardized achievement tests focus on measuring students’ individual academic progress, large-scale assessments evaluate the overall performance of education systems. International large-scale assessments contribute to countries’ ability to compare and develop their education systems. While these tests aim to enhance quality in education, the direct impact of their results on school accountability has triggered debates.

Test-based accountability evaluates education systems based on student achievement, applying rewards or sanctions to schools, as seen in the NCLB. This system creates pressure on teachers and students and leads to a narrowed curriculum (Duke *et al.*, 2003). While test-based accountability systems lead to improving education quality, making student achievement measurable, facilitating data-driven policy-making, and encouraging teachers to work systematically, they also have negative consequences such as excessive pressure on schools, restriction of creativity, curriculum narrowing, rote learning, validity and reliability issues, and educational inequalities. These results are consistent with previous findings indicating that accountability-oriented educational reforms often prioritize measurable outcomes while neglecting the genuine and more valuable aims of education (Au, 2011; Darling-Hammond, 2010; Linn, 2013). In parallel with this view, Carnoy and Loeb (2002) and Hanushek and Raymond (2005) emphasized that although test-based accountability systems promote transparency and performance monitoring, they frequently create pressures that hinder creativity and student-centered approaches in education. Therefore, the reviewed evidence supports the argument that educational quality cannot be evaluated solely on the basis of standardized test results. Although these systems contribute to improving education quality, using them as the sole assessment tool and achievement criterion hinders a holistic evaluation of student development by disregarding factors such as teaching processes, motivation, and learning environments. These findings align with the concerns expressed by Fullan (2011) and Ravitch (2016), who emphasized that accountability systems primarily focused on test scores tend to overlook the holistic aims of education. Tests should be utilized as a tool for shaping educational policies, and alternative assessments such as project and portfolio-based evaluations, student-centered approaches, and methods that foster critical thinking and problem-solving skills should be supported in education policies.

High-stakes testing has increased achievement differences (Darling-Hammond, 2004). Countries attempting to eliminate these achievement differences through standardized curricula and assessments risk creating uniform schools that produce individuals lacking intellectual thinking skills (Mulford, 2002). As a result, curricula are significantly narrowed, leading to a fundamental shift toward test-focused instruction (Duke *et al*., 2003). The U.S., which has pursued similar standardization programs, has experienced a significant decline in creativity over the past decade (Merryman & Bronson, 2010). This raises concerns that schools, due to incorrect attitudes and methods, may be killing creativity. Similarly, Sahlberg (2016) argued that test-oriented policies may narrow the curriculum and weaken teacher autonomy. In the Turkish context, as noted by Berberoğlu and Kalender (2005), centralized testing mechanisms often determine curriculum priorities and may potentially marginalize critical thinking and creativity. Therefore, educational policy reforms should aim to integrate accountability with developmental and formative assessment practices (Fullan, 2011; Ravitch, 2016). Additionally, students should not be labeled as inadequate or deprived of the opportunity to explore their interests simply because they do not perform well on standardized tests.

It has been argued that the accountability exams used today have low instructional sensitivity and that student performance is influenced more by socioeconomic factors than by teacher effectiveness. Holding only schools or teachers accountable for student achievement without considering external factors is not a fair approach (Muñoz-Chereau & Thomas, 2016). Teachers and school leaders, in their effort to be accountable through test-based evaluations, attempt to teach the same information to all students within the same time frame. This, in turn, results in highly standardized learning experiences and increasingly homogenized classroom environments. Education should foster creativity, intellectual thinking, and free classroom environments. Instead of accountability systems that place pressure on teachers, school administrators, students, and parents, value-added accountability models that support student learning, contribute to life success, and consider not only test scores but also student performance should be preferred. Moreover, accountability systems developed without including teachers have been reported to increase teachers' workload and stress (Zahm, 2020). In summary, it is emphasized that improving educational performance is impossible without directly supporting the work of teachers and students in the classroom (Black & Wiliam, 2010).

LIMITATIONS

Our focus was primarily on test-based accountability systems, particularly those relying on standardized achievement scores. Future studies could examine how accountability pressures influence teachers' instructional strategies, students' long-term learning outcomes, and school climate. Comparative studies across different educational systems could provide greater insight into how cultural and systemic factors influence test-based accountability. (Darling-Hammond, 2010; Sahlberg, 2016). Mixed-methods studies that integrate teacher perceptions, classroom practices, and student outcomes could contribute to a more holistic perspective on how accountability impacts educational quality across contexts.

A significant limitation of this research is the exclusion of non-English and non-Turkish language literature, which may introduce a linguistic bias into the findings. As this study employs a descriptive review methodology, it is unable to perform a quantitative synthesis of effect sizes or meta-analytic comparisons. Therefore, given these methodological constraints and the inherent contextual differences in accountability practices globally, the conclusions drawn should be interpreted judiciously.

Since this study was developed according to a general review methodology rather than a systematic review, guidelines such as PRISMA were not followed.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. This study does not involve humans and/or animals; for this reason, no ethics approval was involved in the writing of this paper.

Data Availability Statement

No datasets were generated or analyzed during the current study.

Contribution of Authors

Zeynep Teke, Özen Yıldırım, and **Zeliha Selçuk** performed the same contribution for all the processes of the research from the beginning to the end.

Orcid

Zeynep Teke https://orcid.org/0009-0008-5281-5902
Özen Yıldırım https://orcid.org/0000-0003-2098-285X
Zeliha Selçuk https://orcid.org/0009-0007-4760-9170

REFERENCES

Acar, M. (2013). Eğitim yönetiminde kuram ve uygulama [Theory and practice in educational administration]. In S. Özdemir, F. Sezgin, & S. Koşar (Eds.), *Eğitimde hesap verebilirlik* [Accountability in education] (pp. 385-413). Pegem Akademi.

Ambrosio, J. (2013). Changing the subject: Neoliberalism and accountability in public education. *Educational Studies*, *49*(4), 316-333. https://doi.org/10.1080/00131946.2013.783835

Anderson, J.A. (2005). *Accountability in education*. International Institute for Educational Planning.

Au, W. (2022). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge. https://doi.org/10.4324/9781003005179

Aydoğdu, Ü.R., Karamustafaoğlu, O., & Bülbül, M.Ş. (2017). Akademik araştırmalarda araştırma yöntemleri ile örneklem ilişkisi: Doğrulayıcı doküman analizi örneği [The relationship between research methods and sampling in academic research: An example of confirmatory document analysis]. *Dicle University Journal of Ziya Gökalp Education Faculty*, *30*(1), 556-565. https://doi.org/10.14582/DUZGEF.1803

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi [Examination of student achievement by years, school types, and regions: ÖSS and PISA analysis]. *Eğitim Bilimleri ve Uygulama*, *4*(7), 21-35.

Bilgin, K.U. (2007). Kamuda ölçülebilir denetime hazırlık: Performans yönetimi [Preparing for measurable audit in the public sector: Performance management]. *Sayıştay Dergisi*, *65*(1), 53-87. https://dergi.sayistay.gov.tr/files/4392_der65m9.pdf

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81-90. https://doi.org/10.1177/003172171009200119

Bozkurt, E., & Karabıyık, İ. (2003). *Türk milli eğitiminde denetim sistemi sorunları ve çözüm önerileri: Türk milli eğitim teftiş sisteminde yapılanma sorunu* [Problems in the supervision system of Turkish national education and proposed solutions: Structural issues in the Turkish national education inspection system]. Temsen.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331. https://doi.org/10.3102/01623737024004305

Catterall, J.S. (2012). *The arts and achievement in at-risk youth: Findings from four longitudinal studies* (Research Report No. 55). National Endowment for the Arts. https://www.arts.gov/sites/default/files/Arts-At-Risk-Youth.pdf

Chiu, M.M. (2010). Effects of inequality, family and school on mathematics achievement: Country and student differences. *Social Forces*, *88*(4), 1645-1676. https://doi.org/10.1353/sof.2010.0019

Cole, S., Paulson, A., & Shastry, G.K. (2016). High school curriculum and financial outcomes: The impact of mandated personal finance and mathematics courses. *Journal of Human Resources*, *51*(3), 656-698. https://doi.org/10.3368/jhr.51.3.0113-5410R1

Cook-Harvey, C.M., Darling-Hammond, L., Lam, L., Mercer, C., & Roc, M. (2016). *Equity and ESSA: Leveraging educational opportunity through the Every Student Succeeds Act*. Learning Policy Institute. https://learningpolicyinstitute.org/media/63/download

Curran, B. (1999). *State accountability issues, challenges, and strategies*. National Governors Association.

Çağlar, M., & Kılıç, A. (2017). Standart testler ile öğretmen yapımı testlerinin karşılaştırılması [Comparison of standardized tests and teacher-made tests]. *Eğitime Bakış*, *13*(40), 60-65. https://www.ebs.org.tr/storage/publication/FXxs6OkKAHhn3F1IDDmbANjYFYyZOHObxGZ0PSIv.pdf

Daigle, S.L., & Cuocco, P. (2002). Public accountability and higher education: Soul mates or strange bedfellows?. *EDUCAUSE Center for Applied Research Bulletin*, *2002*(9), 1-14. https://er.educause.edu/~/media/files/articles/2004/1/erb0209.pdf

Darling-Hammond, L. (2006). No Child Left Behind and high school reform. *Harvard Educational Review*, *76*(4), 642-667. https://doi.org/10.17763/haer.76.4.d8277u8778245404

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. Teachers College Press.

Dee, T.S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418-446. https://doi.org/10.1002/pam.20586

Duke, D.L., Grogan, M., & Tucker, P.D. (2003). Educational leadership in an age of accountability. In W.F. Heinecke (Ed.), *Educational leadership in an age of accountability: The Virginia experience* (pp. 198-214). State University of New York Press.

Erwin, B., & Bloomquist, L. (2024). *50-state comparison: School accountability systems*. Education Commission of the States. https://www.ecs.org/50-state-comparison-school-accountability-systems-2024/

Froese-Germain, B. (2010). *The OECD, PISA and the impacts on educational policy*. Canadian Teachers' Federation. https://files.eric.ed.gov/fulltext/ED532562.pdf

Fullan, M. (2011). *Choosing the wrong drivers for whole system reform*. Centre for Strategic Education.

Furman, G.C. (1994). Outcome-based education and accountability. *Education and Urban Society*, *26*(4), 417-437. https://doi.org/10.1177/0013124594026004008

Gallagher, C.W. (2007). Reclaiming assessment: A better alternative to the accountability agenda. *ENCOUNTER: Education for Meaning and Social Justice*, *20*(4), 107-112.

Gaertner, M.N., Kim, J., DesJardins, S.L., & McClarty, K.L. (2014). Preparing students for college and careers: The causal role of Algebra II. *Research in Higher Education*, *55*(2), 143-165. https://doi.org/10.1007/s11162-013-9322-7

Gong, B. (2002). *Designing school accountability systems: Towards a framework and process*. Council of Chief State School Officers & Accountability Systems and Reporting. https://www.govinfo.gov/content/pkg/ERIC-ED464412/pdf/ERIC-ED464412.pdf

Haertel, E., & Herman, J. (2005). A historical perspective on validity: arguments for accountability testing. *Yearbook of the National Society for the Study of Education*, 104(2), 1-34. https://doi.org/10.1111/j.1744-7984.2005.00023.x

Hamilton, L.S., Stecher, B.M., & Klein, S.P. (Eds.). (2002). *Making sense of test-based accountability in education. RAND Education*. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf

Hanushek, E.A., & Raymond, M.E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327. https://doi.org/10.3386/w10591

Harland, J., Kinder, K., Lord, P., Stott, A., Schagen, I., Haynes, J., & Paola, R. (2000). *Arts education in secondary schools: Effects and effectiveness.* National Foundation for Educational Research.

Herdman, E.A. (2006). Derleme makale yazımında, konferans ve bildiri sunumu hazırlamada pratik bilgiler [Guidelines for conducting a literature review and presenting conference papers]. *Journal of Education and Research in Nursing, 3*(1), 2-4.

Hesapçıoğlu, M., Bakioğlu, A., & Baltacı, R. (2001). Öğretmen eğitiminde sorumluluk ve akreditasyon [Responsibility and accreditation in teacher education]. *Journal of Theory and Practice in Education, 1*(1), 143-160.

Hetland, L., & Winner, E. (2001). The arts and academic achievement: What the evidence shows. *Arts Education Policy Review, 102*(5), 3-6. https://doi.org/10.1080/10632910109600008

The Council of Chief State School Officers (CCSSO). (2008). *An implementer's guide to growth models.* https://www.isbe.net/Documents_GMWG/Implementers_guide_growth_models.pdf

Joensen, J.S., & Nielsen, H.S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, *44*(1), 171-198. https://doi.org/10.3368/jhr.44.1.171

Kalman, M., & Gedikoğlu, T. (2014). Okul yöneticilerinin hesap verebilirliği ile örgütsel adalet arasındaki ilişkinin incelenmesi [An investigation of the relationship between school administrators' accountability and organizational justice]. *Hacettepe University Journal of Education*, *29*(2), 115-128.

Kim, J., Kim, J., DesJardins, S. L., & McCall, B. P. (2015). Completing Algebra II in high school: Does it increase college access and success? *The Journal of Higher Education*, *86*(4), 628-662. https://doi.org/10.1080/00221546.2015.11777377

Lane, S. (2020). Test-based accountability systems: The importance of paying attention to consequences. *ETS Research Report Series, 2020*(1), 1-22. https://doi.org/10.1002/ets2.12283

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, 45(3), 515-539. https://doi.org/10.1002/berj.3511

Leckie, G., & Prior, L. (2022). A comparison of value-added models for school accountability. *School Effectiveness and School Improvement*, *33*(3), 431-455. https://doi.org/10.1080/09243453.2022.2032763

Linn, R.L. (2005). Issues in the design of accountability systems. *Yearbook of the National Society for the Study of Education*, *104*(2), 78-98. https://doi.org/10.1111/j.1744-7984.2005.00026.x

Merryman, A., & Bronson, P. (2010). *Nurtureshock: Why everything we thought about children is wrong.* Random House.

Ministry of National Education. (2016). *PISA 2015 ulusal raporu* [PISA 2015 national report]. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2023_10/09155512_PISA_2015.pdf

Montez, J.K., Hummer, R.A., & Hayward, M.D. (2012). Educational attainment and adult mortality in the United States: A systematic analysis of functional form. *Demography*, *49*(1), 315-336. https://doi.org/10.1007/s13524-011-0082-8

Mulford, B. (2002). The global challenge: A matter of balance. *Educational Management & Administration*, *30*(2), 123-138. https://doi.org/10.1177/02611X02030002505

Muñoz-Chereau, B., & Thomas, S.M. (2016). Educational effectiveness in Chilean secondary education: Comparing different 'value added' approaches to evaluate schools. *Assessment in Education: Principles, Policy & Practice*, *23*(1), 26-52. https://doi.org/10.1080/0969594X.2015.1066307

National Research Council. (2010). *Getting value out of value-added: Report of a workshop*. The National Academies Press. https://doi.org/10.17226/12820

Nichols, S.L., & Berliner, D.C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Harvard Education Press. https://doi.org/10.1080/15434300701776393

OECD. (2010). *Accountability and transparency: A guide for state ownership*. https://doi.org/10.1787/9789264056640-en

OECD. (2014). *How has student performance evolved over time? (PISA in focus, No. 47)*. https://doi.org/10.1787/5js7q3kbkklt-en

OECD. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. https://doi.org/10.1787/9789264248373-en

Ohio Department of Education. (2009). *Guide to understanding Ohio's accountability system 2009-2010*. https://www.rrcs.org/Downloads/StateReportCardGuide0910.pdf

Özdemir, S., Acar, M., & Bülbül, M. (2018). *Eğitimde hesap verebilirlik* [Accountability in education]. Pegem Akademi.

Özçelik, D.A. (2016). *Ölçme ve değerlendirme* [Measurement and evaluation]. Pegem Akademi.

Parsad, B. (2012). *Arts education in public elementary and secondary schools: 1999-2000 and 2009-10*. DIANE Publishing.

Phelps, R. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice*, *19*(1), 11-21. https://doi.org/10.1111/j.1745-3992.2000.tb00018.x

Popham, W.J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-150. https://doi.org/10.1177/003172170708900211

Popham, W.J. (2016). *The ABCs of educational testing: Demystifying the tools that shape our schools*. Corwin Press. https://doi.org/10.4135/9781506351520

Ravitch, D., & Stoehr, L.A. (2017). The death and life of the great American school system: How testing and choice are undermining education. In E. S. Shaffer (Ed.), *Early learning* (pp. 125-128). Routledge. https://doi.org/10.4324/9781315429977

Reeves, D.B. (2004). *Accountability for learning: How teachers and school leaders can take charge*. Association for Supervision & Curriculum Development.

Rolling, J.H. (2006). The arts and the creation of mind. *Journal of Curriculum Studies*, *38*(1), 113-125. https://doi.org/10.1080/00220270500122604

Rudd, K., & Gillard, J. (2008). *Quality education: The case for an education revolution in our schools*. Australian Politics. https://australianpolitics.com/2008/08/27/education-revolution-rudd-npc.html

Sabol, R.F. (2013). Seismic shifts in the education landscape: What do they mean for arts education and arts education policy? *Arts Education Policy Review*, *114*(1), 33-45. https://doi.org/10.1080/10632913.2013.744250

Schleicher, A. (2021). PISA gibi büyük ölçekli değerlendirmeler nasıl bir fark yaratabilir? [How can large-scale assessments like PISA make a difference?] (Ç. Çetin & S. Aslan Mutlu, Trans.). *İstanbul Measurement and Evaluation Journal*, *2*(2), 10-19. https://istanbul.meb.gov.tr/dergi/istanbul_olcme_degerlendirme_2/mobile/index.html

Sahlberg, P. (2021). *Finnish lessons 3.0: What can the world learn from educational change in Finland?* Teachers College Press.

Sheridan, K.M., Veenema, S., Winner, E., & Hetland, L. (Eds.). (2022). *Studio thinking 3: The real benefits of visual arts education*. Teachers College Press.

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, *87*(4), 324-328. https://doi.org/10.1177/003172170508700414

Tansel, A. (2015). *Inequality of opportunities of educational achievement in Turkey over time (IZA Discussion Paper No. 9005)*. Institute for the Study of Labor. https://www.econstor.eu/handle/10419/110732

Thomas, K., Singh, P., & Klopfenstein, K. (2015). Arts education and the high school dropout problem. *Journal of Cultural Economics*, *39*(4), 327-339. https://doi.org/10.1007/s10824-014-9238-x

Thompson, G. (2013). NAPLAN, MySchool and accountability: Teacher perceptions of the effects of testing. *The International Education Journal: Comparative Perspectives*, *12*(2), 62-84. https://files.eric.ed.gov/fulltext/EJ1017709.pdf

Torres, R. (2021). *Does test-based school accountability have an impact on student achievement and equity in education?: A panel approach using PISA (OECD Education Working Paper No. 250)*. OECD Publishing. https://doi.org/10.1787/0798600f-en

Turgut, M.F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme* [Measurement and evaluation in education] (2nd ed.). Pegem Akademi.

TÜSİAD, & TÖDER (2012). *PISA 2012 Değerlendirmesi: Türkiye için veriye dayalı eğitim reformu önerileri* [PISA 2012 Assessment: Evidence-based education reform recommendations for Turkey]. TÜSİAD.

UNESCO Institute for Statistics. (2018). *The impact of large-scale assessments on education policy: A synthesis of evidence from 13 country case studies.* UNESCO. https://uis.unesco.org/sites/default/files/documents/impact-large-scale-assessments-2018-en.pdf

Ülkü, S., & Atar, B. (2023). Investigation of measurement invariance of Turkish subtest on ABIDE 2016 in relation to characteristics of teachers: Sub-sampling method. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(2), 154-170. https://doi.org/10.21031/epod.1084985

Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning*, *4*(1). https://doi.org/10.22329/jtl.v4i1.89

Winner, E., Hetland, L., Veenema, S., Sheridan, K., & Palmer, P. (2006). Studio thinking: How visual arts teaching can promote disciplined habits of mind. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 189-206). Routledge.

Woods, C.S., Park, T., Hu, S., & Bertrand Jones, T. (2018). How high school coursework predicts introductory college-level course success. *Community College Review*, *46*(2), 176-196. https://doi.org/10.1177/0091552118759419

Watts, J.A., Gaines, G.F., & Creech, J.D. (1998). *Getting results: A fresh look at school accountability*. Southern Regional Education Board.

Yalçın, K. (2011). Almanya ve Türkiye'de 2003-2006-2009 PISA araştırma sonuçlarına göre alınan önlemler ve uygulamalar [Precautions and applications according to 2003-2006-2009 PISA research results in Germany and Turkiye]. *Türk Kütüphaneciliği*, *25*(4), 494-508.

Yıldırım, Ö. (2021). Geniş ölçekli testlerde madde yapıları ve hazırlanmasında dikkat edilmesi gerekenler (PISA ve TIMSS) örnekleri [Considerations in the construction and preparation of item structures in large-scale assessments (examples from PISA and TIMSS)]. *İstanbul Ölçme Değerlendirme Dergisi*, 2, 50-59. https://istanbul.meb.gov.tr/dergi/istanbul_olcme_degerlendirme_2/mobile/index.html

Zahm, K. (2020). Teacher accountability and professional stress: A policy analysis. *Educational Policy Journal*, *34*(3), 245-263.

Zülfüoğlu, Ö., & Koçak, H. (2017). Need for education reforms in Turkey: An assessment of PISA results. In A.M. Çilingirtürk, A.A.-Słocińska, & B.B. Bali (Eds.), *Economics, management, and econometrics: Education systems and reform* (pp. 55-76). IJOPEC Publication.