Physical Education, Health and Social Sciences

https://journal-of-social-education.org E-ISSN: 2958-5996

P-ISSN: 2958-5988

Measuring What Matters? Rethinking Standardized Testing Practices in Public Sector Institutions

Dr. Muhammad Javed Aftab[1], Anees Ur Rehman*[2], Maha Altamash[3]

[1] Assistant Professor (Special Education), Department of Special Education (DSE), Division of Education (DoE), University of Education (UoE), Lahore, Punjab, Pakistan.
Email: drmjavedaftab@ue.edu.pk

[2] M.Phil. Scholar (Special Education), Department of Special Education, Division of Education University of Education, Lahore, Punjab, Pakistan.
*Corresponding Author Email: aneesurrehman.sped@gmail.com

[3] M.Phil. Scholar (Special Education), Department of Special Education, Division of Education University of Education, Lahore, Punjab, Pakistan. Email: mahaaltamash786@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i1.1209

Abstract

This research focuses on how government school teachers view standardized tests and whether they think these tests are in line with the curriculum standards and classroom assessments. Accountability pressure, curriculum test alignment, test, preparation intensity, formative assessment practices, and perceived fairness were measured in a cross, sectional survey of 300 teachers. Descriptive statistics, t, tests for independent samples, one, way ANOVA, and Pearson correlations were used. Teachers indicated moderate accountability and alignment and slightly stronger formative practices. No significant differences between genders were found. The comparisons of urban rural revealed small but statistically significant differences in test preparation, formative practice, and fairness, while school level differences were only for fairness. Teaching experience was a significant predictor of formative assessment, where teachers with 11+ years scored higher than those with 05 years. The correlations between constructs were weak and negative, which indicates trade, offs between test preparation and formative routines. The internal consistency estimates were very low, suggesting item, coding problems and the necessity for further psychometric validation. The findings encourage the transition to balanced assessment systems and the development of teacher's assessment literacy as a means of better safeguarding learning and equity.

Keywords: Standardized Testing; Teacher Perceptions; Curriculum–Assessment Alignment; Accountability Pressure; Formative Assessment; Fairness; Public-Sector Schools

Introduction

Over the past twenty years or so, there has been a heavy reliance on standardized tests in the public sector education systems to validate learning, determine the distribution of resources, and hold schools accountable. While these tests deliver standardization, credibility, and cost, effectiveness at large scale, they also generate challenging questions about our definition of learning and if narrow measures really represent the outcomes of such complex skills as critical thinking, civic competence, collaboration, and wellbeing. After the COVID, 19 pandemic, new world, wide data illustrates an uneven recovery and expansion of disparities, which adds more importance to

questioning whether the existing testing systems are assessing the right things for learners and societies (OECD, 2023; NCES, 2024; Brookings, 2024, 2025).

Data from international trend studies bring the issue to a higher level. PISA 2022 which is the first major international study to compare pre, pandemic and post, pandemic performance shows that the average level of achievement dropped in numerous educational systems, and the inequality gaps especially those based on socio, economic status, continued to exist. Such trends lead to the suspicions that educational systems might be fine, tuning for the testable parts of the curriculum without guaranteeing the acquisition of more comprehensive and long, lasting capabilities (OECD, 2023).

Educational assessments reflect this situation as well. For instance, the National Assessment of Educational Progress (NAEP) in the United States reveals that although Grade 4 math scores slightly bounced back in 2024 compared to 2022, both math and reading scores are still below the 2019 levels, and the drop is more significant for lower, performing students. Such findings indicate that repeatedly measuring alone is not enough; educational systems must link assessment data to the provision of appropriate support and teaching methods. (NCES, 2024; NCES, 2025; Brookings, 2024, 2025).

Moreover, the issue is even bigger in low, and middle, income countries. The World Banks learning poverty measure suggests that the proportion of 10, year, olds who cannot read and understand a simple text rose to approximately 70% after the pandemic. UNICEFs monitoring of foundational learning also points to the fact that basic reading and math are still lagging the SDG 4 targets. These facts call into question whether high, stakes standardized testing is being used to initiate changes or if it is limiting the use of the assessment for formative purposes that actually lead to learning (World Bank, 2022, 2024; UNICEF, 2023).

Meanwhile, testing debates have become a lot more heated when it comes to admissions in higher education.

Several top, notch universities (e.g. Yale, Harvard) have reinstated standardized test requirements and stated that judicious use of scores supports a holistic, equitable selection process. This is an example of the overarching problem that holding tests. the tests can be the informative components of a well, balanced evidence system, but over, reliance can lead to bias, teaching to the test, and misalignment with the institution's mission in the public sector. (Reuters, 2024; Axios, 2024). Practitioners cannot help but feel pressure. Surveys of educators indicate a really strong perceived pressure to increase test scores, whereas the research shows that when accountability is present, it can limit courses to only the tested materials. A number of papers call for "balanced assessment systems" that use high, quality formative assessment, performance tasks, and periodic external checks so the emphasis is no longer on "how much did students score? " but rather on "what evidence do we have that students learner transfer and apply learning? " (Education Week, 2023; Zakharov, 2021; Marion et al., 2024).

Citizen, led and system, diagnostic assessments provide additional perspectives. As an illustration, ASER Pakistan a non, formal educational environment periodically measures the country's Main stream Literacy and Numeracy Status through household surveys. The data generated is used by policymakers to go beyond exam pass rates and focus on understanding the actual proficiency of students in reading and arithmetic.

These additional complementary data sources can provide a basis for the improvement of the general outcomes of education and equity, focused interventions, which are often claimed to be the results of standardized testing but the tests themselves are not a guarantee of such outcomes (ASER Pakistan, 2022).

Standardized testing is a term used for evaluations given and graded in a standard way to facilitate the direct comparison of students, schools, or regions. These tests are popular because people

believe they are fair, and they are cheap and easy to compare; however, the opponents say that the test has a very limited scope, it is culturally biased, and it encourages teachers to prepare students for the test only. Data on post, pandemic performance along with widening equity gaps have brought back the debate on whether these tools, in their current form, genuinely contribute to public sector goals of quality, equity, and inclusion (EBSCO Research Starters, n. d.; OECD, 2023; NCES, 2024).

Research results regarding accountability's impact are diverse. A cross, country study of the PISA data did not find evidence of significant achievement gains as a result of the implementation of accountability practices in high, income systems and only limited evidence in some low, and middle, income settings. At the same time, it is documented by the studies that test, focusing can lead to the narrowing of instruction and there will be no sustaining of the changes related to the improvement of deeper learning (OECD, 2021; Zakharov, 2021).

Balanced assessment frameworks propose a recalibration: combine classroom formative assessment, curriculum-embedded performance tasks, and strategically timed external assessments to support learning while preserving system comparability. Recent syntheses outline how such systems can be designed, though implementation remains difficult (Marion et al., 2024; Marion, 2018).

Despite abundant test score data, relatively few studies examine *alignment* between what public sector institutions aim to deliver (foundational skills, equity, wellbeing, and transferable competencies) and what high-stakes standardized tests actually measure. There is also limited empirical work in LMIC contexts on how assessment design and use shape classroom practice at scale and whether integrating performance assessments with external measures yields better equity and learning outcomes than test-centric regimes (OECD, 2021; UNESCO GEM, 2023; World Bank, 2024; Morgan, 2025).

Public sector education systems face a dual challenge: (1) learning levels have stalled or declined and equity gaps persist, and (2) prevailing high-stakes standardized testing practices may be misaligned with the broader competencies societies need. Without rethinking what and how we measure by integrating more formative, authentic, and equity-sensitive approaches systems risk optimizing for test scores rather than meaningful, transferable learning (OECD, 2023; NCES, 2024; Brookings, 2025; UNICEF, 2023).

1. Evaluate how well current standardized tests reflect public-sector goals foundational skills, equity, and transferable competencies.
2. Quantify relationships between accountability pressures, classroom practices (e.g., test prep vs. formative use), and student outcomes across subgroups.
3. Compare learning gains and equity effects of balanced assessment models (formative + performance tasks + periodic externals) versus high-stakes-only regimes.
4. Design and pilot a scalable “Measuring What Matters” framework indicators, governance, reporting cycles, data-use protocols assessing feasibility, costs, capacity needs, and equity safeguards.

The research provides ministries, assessment bodies, and school leaders with policy, relevant advice they can use to raise learning quality without compromising equity and other broader outcomes. It achieves this by distilling worldwide evidence and plotting tangible design decisions for balanced systems, thus helping to shift the focus of measurement from a mere check, up to a facilitator of teaching improvement and student development.

To LMICs and financially challenged systems, the framework introduces the idea of complementing (instead of replacing) system, level exams with low, cost diagnostic tools (Marion et al., 2024; World Bank, 2024; ASER Pakistan, 2022).

Conceptual Framework

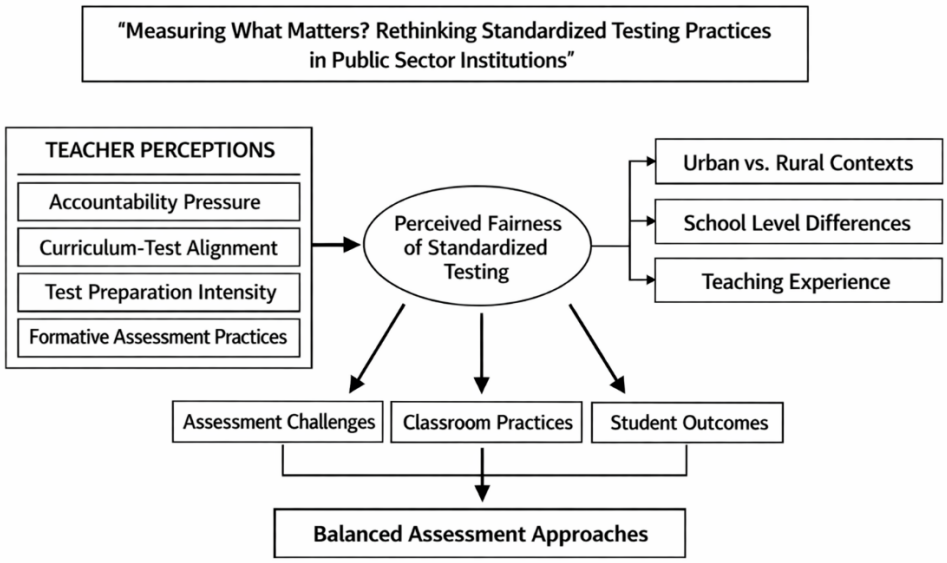

Figure: Conceptual Framework

Theoretical Framework

The study is grounded in a combined validity, and, consequences perspective that views standardized testing as a policy instrument and the emphasis is on the policy instrument as its meaning is resulting from how scores are interpreted, used and lived in schools. From this perspective, test, based accountability would bring about system, level washback, which would change curriculum coverage and teaching methods, thereby increasing test, preparation efforts and narrowing instruction only to what is tested. At the classroom level, the framework identifies curriculum test alignment and accountability pressure as upstream factors that firstly affect teachers decisions regarding their assessments, and secondly, their commitment to the use of formative assessment for learning (provision of feedback, diagnosis, and instructional adjustment). Furthermore, the model recognizes that testing cannot be evaluated solely from a technical standpoint; it also has to be assessed in terms of its fairness and social consequences, which include factors such as access, bias, conditions of administration, and the equity implications of score use across different contexts (e. g., urbanrural and school level). Last but not least, by pulling from argument, based validation, the framework sees validity as a well, established reason (interpretation, use argument) that should be able to explicitly state the intended claims, alternative explanations, and unintended impacts, thus turning "rethinking standardized testing" into a question of both measurement quality and educational justice (Alderson & Wall, 1993; Au, 2007; Black & Wiliam, 1998; Gerasimova, 2024; Kane, 2013; Kunnan, 2004; Messick, 1995; Nichols & Berliner, 2007).

Review of Related Literature

Standardized testing in public systems traditionally serves multiple purposes certification, selection, accountability, and system monitoring yet these purposes do not always align with stated policy goals such as equity, breadth of competencies, and instructional improvement. Recent policy syntheses emphasize that assessment design should be judged not only by technical quality but also by its consequences for teaching, learning, and equity what contemporary validity theorists call consequential validity. The shift toward "measuring what matters" therefore entails attending

to impacts on learning opportunities, not just score precision, and designing assessment ecosystems that include classroom, school, and system-level components in coherent ways. (Kinnear, 2024; OECD, 2021).

2) Global achievement trends after the pandemic: why stakes feel so high

Across countries, large-scale assessments show unprecedented declines in learning especially in mathematics raising the political salience of testing even as many systems question its design. PISA 2022 reported an average fall of ~15 points in mathematics and ~10 in reading across OECD systems relative to 2018 (roughly three-quarters and one-half of a school year, respectively), while science was largely flat. In the United States, the NAEP 2024 cycle showed grade-4 reading still below 2019 and mathematics only partially rebounding; lower-performing percentiles remain most affected. Synthesis work tracking multiple interim assessments indicates that math recovery is inching forward but full recovery could take many years at current rates, underscoring the need for assessments that both monitor and improve learning. (OECD, 2023; NCES, 2024; Sawchuk & Kane, 2025; Brookings, 2025). At the same time, learning poverty the share of 10-year-olds unable to read and understand a simple text remains stubbornly high in many contexts, with a refreshed methodology and database released in April 2024 to improve monitoring. These global trends amplify pressures on public systems to demonstrate progress through test score gains, even when such gains may reflect test prep rather than deeper learning. (World Bank, 2024).

3) Accountability and alignment: what high-stakes tests capture and what they miss

Evidence on the effects of test-based accountability is mixed and context-dependent. Using PISA panel data (2006–2015), an OECD working paper found no conclusive evidence that accountability practices improved outcomes in high-income systems; in some low- and middle-income systems, accountability correlated with higher performance but also with increased inequality suggesting trade-offs when accountability pressure is not paired with capacity and autonomy for curriculum and assessment decisions. The implication is that alignment between assessment content, instructional goals, and supports determines whether accountability steers systems toward or away from "what matters." (Torres, 2021, OECD Working Paper No. 250).

A key alignment question is whether tests represent the intended curriculum and cognitive demand. Technical reports for large-scale programs document reliability and multiple sources of validity evidence, but item formats and sampling frames can still narrow the construct measured. Consequential validity research argues that equity and social accountability must be treated as part of the validity argument particularly where stakes are high and feedback cycles shape classroom practice. (Florida DOE, 2021; Kinnear, 2024).

4) Curriculum narrowing and "teaching to the test"

Empirical studies since 2018 document both narrowing effects and more nuanced patterns. In a quasi-experimental study, Zakharov & Carnoy (2021) showed that test preparation can raise performance on the target test but transfers less to broader or more cognitively demanding measures classic evidence of construct-level misalignment. Recent work in different systems continues to link high-stakes pressure with narrower pedagogies and constrained teacher collaboration, although effects vary by subject and policy design. The upshot for public systems is not that all external testing is harmful, but that high stakes attached to a single metric can incentivize short-term score gains at the expense of richer learning. (Zakharov & Carnoy, 2021; Feniger et al., 2024; Levatino, 2024; Jerrim, 2024).

5) Fairness, anxiety, and differential impact

The fairness of standardized testing hinges on both psychometric bias and differential consequences. Newer studies leverage response process data and DIF methods to detect items functioning differently across groups (e.g., gender), while reviews in language assessment show growing attention to fairness audits. In parallel, a meta-analytic literature shows test anxiety is robustly associated with lower performance, with cognitive worry components especially predictive; recent work during and after COVID-19 suggests anxiety effects can be largest for younger learners and those already at academic risk. Interpreting system-level scores without attention to these mechanisms risks misclassifying students and schools, widening gaps despite an equity narrative. (Li et al., 2024; von der Embse et al., 2018; Jerrim, 2023; SAGE systematic review, 2024).

Equity concerns also surface in distributional trends. NAEP's percentile reporting shows that post-pandemic rebounds have been slower at the 10th and 25th percentiles than at the median and upper percentiles, a pattern consistent with "Matthew effects" under generic accountability pressure. Meanwhile, U.S. analyses highlight a widening GPA test score gap, with GPAs rising or holding steady even as standardized scores fell potentially masking learning losses from families and policymakers who rely on grades. (NCES, 2024; Kraft et al., 2023, Brookings).

6) LMIC and Pakistan-specific evidence: measuring foundational learning and system readiness

Public sector systems in South Asia face dual pressures: low baseline proficiency and calls for accountability. Pakistan's National Achievement Test (NAT) 2023 (grades 4 & 8) reported low mastery in mathematics and science (e.g., only ~17% of grade-4 and ~8% of grade-8 students at ≥75% in math), mirroring household-based evidence from ASER Pakistan 2023 on foundational literacy and numeracy. Provinces like Punjab have adopted an Assessment Policy Framework (2019) that pairs large-scale assessments with school-based assessments to shift from "exam culture" toward assessment for learning; implementation includes item banks and digital tools, though capacity and coherence remain ongoing challenges. (PIE, 2024; APP, 2024; ASER Pakistan/ITA, 2024; PEC/APF, 2019–2022).

At the global policy level, UNICEF's 2023 Tracking Progress on Foundational Learning and SDG 4 monitoring stress balanced strategies RAPID for foundational learning (Reach, Assess, Prioritize, Increase, Develop) with countries at varied stages of adopting regular learning assessments and catch-up programs. For Pakistan and peers, the literature suggests that national tests should operate in tandem with targeted formative practices and policy supports (remediation time, materials, teacher PD), rather than as stand-alone accountability levers. (UNICEF, 2023; UNESCO GEM/SDG 4 monitoring, 2024–2025).

7) Evidence on balanced assessment and formative practices

A growing consensus supports balanced assessment systems that integrate formative assessment, performance tasks, and periodic external checks. The National Academy of Education volume (2024) synthesizes design principles and district-level case knowledge, emphasizing coherence, assessment literacy, and governance. Complementing this, an umbrella review of meta-analyses (K-12, 13 syntheses) finds formative assessment yields positive effects from small to large, varying by approach; rigorous UK evidence reviews similarly rate feedback as a high-impact lever when implemented well. The policy implication is not to abandon summative tests but to rebalance toward instructional uses of evidence feedback cycles, student self-assessment, and performance tasks while attending to reliability, moderation, and cost. (Marion, Pellegrino, & Berman, 2024; Sortwell et al., 2024; EEF, 2021/2022).

Performance assessment (portfolios, extended tasks) is often proposed to "measure what matters" (e.g., problem-solving, collaboration), but studies remind us that quality controls shared rubrics, rater training, and moderation are essential to ensure reliability and fairness at scale. Taken together, the literature supports piloting performance tasks within balanced systems rather than replacing all external testing. (Penney et al., 2023; Bell, O'Neill, & Crawford, 2023).

8) Digitalization, AI, and the future of assessment

As national assessments digitize, opportunities (adaptive designs, faster feedback) collide with risks (security, integrity, bias). The OECD Digital Education Outlook 2023 documents rapid digitalization of central exams and recommends guardrails for effective, equitable AI use. Concurrently, the U.S. Department of Education advises that AI-enabled formative tools must be inspectable and bias-aware, and regulators warn that detecting generative-AI use in unsupervised tasks is "all but impossible," pushing institutions toward more authentic, secure assessment formats. In short, the next phase of "measuring what matters" must integrate AI thoughtfully without undermining validity or equity. (OECD, 2023a; OECD, 2023b; U.S. Dept. of Education, 2023; TEQSA guidance reported in *The Australian*, 2025).

9) Reframing "measuring what matters" for public systems

Across this literature, three themes recur. First, high-stakes use of single test scores risks narrowing instruction and amplifying inequities unless paired with capacity-building and formative routines. Second, balanced systems that combine periodic external checks with rich classroom assessment are more likely to align with broader goals (equity, transferable competencies) and to support improvement. Third, fairness requires continuous validity work DIF audits, attention to anxiety and wellbeing, culturally responsive design and transparent data-use protocols. A pragmatic framework would (a) retain external assessments for monitoring and transparency, (b) expand teacher-facing formative assessment and performance tasks with moderation, and (c) publish dashboard indicators beyond test scores (e.g., engagement, progression, equitable access to high-quality instruction), with governance that minimizes perverse incentives. (OECD, 2021; Marion et al., 2024; UNESCO GEM, 2023).

Research Methodology

1. Research Design and Approach:

This study utilizes a quantitative research design to examine the alignment between curriculum and assessment, the pressures of accountability, instructional practices, and student outcomes. The study uses a cross-sectional design to collect data from a large sample of students, teachers, school leaders, and assessment officials. Data will be analyzed using descriptive statistics to assess the alignment and identify patterns across various factors.

2. Setting and Scope:

The study focuses on public education systems, specifically targeting upper primary to lower secondary grades (4–10). The study will be conducted in one province/state, with a small parallel study on upper secondary/higher education exams where applicable.

3. Population and Sampling:

The sample for this study will consist of students, teachers, school leaders, and assessment officials from public schools across a single province or state. A multi-stage, stratified cluster sampling approach will be used to select schools based on their location (urban/rural), school level (upper primary/lower secondary), and school type (boys, girls, or mixed). The target sample includes

approximately 40–60 schools, with 1,200–2,000 students stratified by grade and gender. Additionally, 200–300 teachers and 40–60 school leaders will be selected. For the purposive sampling of assessment officials, 15–25 experts, such as exam setters and assessment designers, will be included. For the qualitative portion, a maximum-variation purposive subsample will be chosen, consisting of 30–40 teacher and leader interviews, 6–8 student focus groups (segregated by grade and gender), and 12–16 classroom observations from diverse contexts within the selected schools. This sample will provide comprehensive insights into the research questions while ensuring the inclusion of varied perspectives across different school settings.

5. Instruments and Measures:

The questionnaire for this study will be designed to collect data on key variables such as accountability pressure, instructional practices, assessment literacy, and perceived alignment between the curriculum and standardized tests. Separate questionnaires will be developed for teachers, school leaders, and students. The teacher and school-leader questionnaire will consist of Likert scale items focused on the frequency and intensity of test preparation, the perceived pressure from standardized testing, the use of formative assessments, and the alignment between curriculum and assessment practices. The student questionnaire will be shorter, with age-appropriate questions about their engagement with assessments, their learning experiences, and, if ethically permissible, a brief measure of test anxiety. Both questionnaires will include background demographic questions to enable disaggregation by gender, socioeconomic status, and other equity covariates. These questionnaires will be piloted to ensure clarity and reliability before full-scale distribution.

8. Data Collection Procedures:

The data collection procedure for this study will be carried out in several stages. Initially, permissions and ethics approvals will be obtained from relevant authorities, including the institutional review board (IRB), school heads, and consent from participants and guardians where required. Enumerator training will take place over 2–3 days, covering key topics such as ethics, survey administration, observation protocols, and data security procedures. The fieldwork will be organized as follows: during Weeks 1–2, document collection and blueprint coding will be completed. In Weeks 3–6, survey distribution and classroom observations will take place across the selected schools, with teams deployed by region. Weeks 7–8 will be dedicated to conducting interviews and focus groups, either in-person or virtually, depending on the situation. All data will be captured digitally on encrypted devices and regularly synced to a secure server. To ensure confidentiality, all participant data will be de-identified, and audit trails will be maintained to track any changes or access to the data. This structured approach will ensure systematic and secure data collection.

9. Data Analysis Plan:

- Quantitative Analysis:
    - Descriptive Statistics: Alignment indices, content/cognitive coverage heatmaps.
    - Scale Construction: Factor analysis and reliability tests.
    - Group Comparisons: Mean differences across gender, SES, etc.
    - Multilevel Models: Analyze data at the student, teacher/class, and school levels, examining the influence of instructional practices, accountability pressure, and school resources.
    - Mediation Models: Investigate the relationships between accountability pressure, instructional practices, and student outcomes.

-

10. Ethical Considerations:

- Voluntary Participation: Informed consent and the right to withdraw.
- Confidentiality: De-identified data, secure storage.
- Minimizing Risks: No test scores reported at an individual level; ethical scheduling of data collection.
- Equity: Ensure accessibility for diverse participants.
-

11. Quality Assurance and Risk Management:

- Field Monitoring: Regular spot checks to ensure data quality.
- Missing Data: Follow pre-specified rules for handling missing data (e.g., multiple imputation).
- Bias Mitigation: Use social desirability scales and neutral phrasing in surveys.
-

13. Delimitations and Limitations:

- Delimitations: Focus on public institutions and tested subjects in one administrative unit.
- Limitations: Causal claims are limited due to the observational design; self-reporting may cause bias.

Table 1
Descriptive Statistics for Demographic Information

Variable	Category	n	%
Gender	Male	158	52.7
	Female	142	47.3
Urbanicity	Urban	182	60.7
	Rural	118	39.3
School Level	Lower Secondary	162	54.0
	Primary	138	46.0
Subject	Language	88	29.3
	Mathematics	79	26.3
	Science	72	24.0
	Social Studies	61	20.3
Experience	0–5 years	123	41.0
	6–10 years	110	36.7
	11+ years	67	22.3

Table 2
Descriptive Statistics for Subscales

Scale	M	SD	Min	Max
Alignment	3.1	0.24	2.5	3.67
Accountability	3.24	0.31	2.33	4.17
TestPrep	3.09	0.22	2.5	3.62
Formative	3.31	0.32	2.5	4.12
Outcomes	3.05	0.24	2.33	3.5

Fairness	3.06	0.26	2.5	3.83

Note. M = mean; SD = standard deviation. Subscale scores are on a 1–5 Likert scale.

Table 3

Internal Consistency (Cronbach's Alpha)

Scale	k	Cronbach's α
Alignment	6	-0.185
Accountability	6	0.353
TestPrep	8	-0.17
Formative	8	0.559
Outcomes	6	-0.261
Fairness	6	-0.013
Overall (40 items)	40	0.078

Note. Cronbach's α computed on item means within each subscale (after reverse-coding).

Table 4a

Independent-Samples t Tests by Gender

Dependent	Group 1	M1	Group 2	M2	t	df	p	Cohen's d
Alignment	Female	3.11	Male	3.08	0.93	297.75	0.352	0.11
Accountability	Female	3.23	Male	3.25	-0.34	296.14	0.738	-0.04
TestPrep	Female	3.09	Male	3.1	-0.44	296.72	0.659	-0.05
Formative	Female	3.32	Male	3.31	0.2	296.3	0.843	0.02
Outcomes	Female	3.04	Male	3.06	-0.89	297.91	0.375	-0.1
Fairness	Female	3.04	Male	3.07	-0.95	297.64	0.343	-0.11

Note. Welch's t-test with df, two-tailed p values, and Cohen's d.

Table 4b

Independent-Samples t Tests by Urbanicity

Dependent	Group 1	M1	Group 2	M2	t	df	p	Cohen's d
Alignment	Urban	3.09	Rural	3.1	-0.1	210.04	0.917	-0.01
Accountability	Urban	3.25	Rural	3.23	0.3	247.14	0.768	0.04
Test Prep	Urban	3.06	Rural	3.13	-2.8	260.3	0.005	-0.33
Formative	Urban	3.37	Rural	3.23	3.73	260.34	<.001	0.43
Outcomes	Urban	3.06	Rural	3.03	1.1	236.65	0.274	0.13
Fairness	Urban	3.03	Rural	3.11	-2.5	239.22	0.013	-0.3

Note. Welch's t-test with df, two-tailed p values, and Cohen's d.

Table 4c

Independent-Samples t Tests by School Level

Dependent	Group 1	M1	Group 2	M2	t	df	p	Cohen's d
Alignment	Primary	3.09	Lower Secondary	3.1	-0.41	288.65	0.682	-0.05
Accountability	Primary	3.21	Lower Secondary	3.27	-1.73	295.84	0.085	-0.2

TestPrep	Primary	3.09	Lower Secondary	3.09	0.16	295.37	0.875	0.02
Formative	Primary	3.32	Lower Secondary	3.3	0.37	276.35	0.711	0.04
Outcomes	Primary	3.04	Lower Secondary	3.06	-0.87	294.05	0.385	-0.1
Fairness	Primary	3.1	Lower Secondary	3.03	2.31	290.45	0.021	0.27

Note. Welch's t-test with df, two-tailed p values, and Cohen's d.

Table 5a

One-Way ANOVA by Teaching Experience

Dependent	Factor	df1	df2	F	p	η^2
Alignment	Experience	2	297	1.39	0.252	0.009
Accountability	Experience	2	297	1.16	0.316	0.008
TestPrep	Experience	2	297	0.31	0.733	0.002
Formative	Experience	2	297	7.77	<.001	0.05
Outcomes	Experience	2	297	0.49	0.614	0.003
Fairness	Experience	2	297	0.63	0.534	0.004

Note. Effect size reported as η^2 (eta squared).

Table 5b

One-Way ANOVA by Subject Taught

Dependent	Factor	df1	df2	F	p	η^2
Alignment	Subject	3	296	0.49	0.691	0.005
Accountability	Subject	3	296	1.45	0.227	0.015
TestPrep	Subject	3	296	2.71	0.045	0.027
Formative	Subject	3	296	1.31	0.27	0.013
Outcomes	Subject	3	296	1.78	0.15	0.018
Fairness	Subject	3	296	1.56	0.199	0.016

Note. Effect size reported as η^2 (eta squared).

Table 6

Correlations Among Subscales

Unnamed: 0	Alignment	Accountability	TestPrep	Formative	Outcomes	Fairness
Alignment	1.0	-0.01	-0.09	0.03	-0.04	0.07
Accountability	-0.01	1.0	0.05	-0.01	0.04	-0.01
TestPrep	-0.09	0.05	1.00	-0.13*	0.02	0.04
Formative	0.03	-0.01	-0.13*	1.00	-0.01	-0.15**
Outcomes	-0.04	0.04	0.02	-0.01	1.0	0.04
Fairness	0.07	-0.01	0.04	-0.15**	0.04	1.00

Note. Pearson correlations. *p < .05, **p < .01, ***p < .001.

Table A1

Distribution Diagnostics (Skewness & Kurtosis)

Scale	Skewness	Kurtosis
Alignment	0.12	-0.26
Accountability	0.01	0.01

TestPrep	-0.04	-0.05
Formative	0.07	-0.26
Outcomes	-0.32	-0.13
Fairness	0.16	-0.19

Table A2

Tukey HSD Post Hoc — Experience on Formative

group1	group2	meandiff	p-adj	lower	upper	reject
0–5 years	11+ years	0.1867	0.0004	0.0737	0.2997	True
0–5 years	6–10 years	0.0893	0.0809	-0.0083	0.187	False
11+ years	6–10 years	-0.0974	0.1168	-0.2127	0.018	False

Note. Reject column indicates statistical significance at α = .05.

Table A3

Tukey HSD Post Hoc — Subject on Alignment

group1	group2	meandiff	p-adj	lower	upper	reject
Language	Mathematics	-0.0267	0.8914	-0.1233	0.0699	False
Language	Science	-0.0457	0.633	-0.1447	0.0534	False
Language	Social Studies	-0.0191	0.9647	-0.1229	0.0848	False
Mathematics	Science	-0.019	0.963	-0.1205	0.0826	False
Mathematics	Social Studies	0.0076	0.9977	-0.0986	0.1139	False
Science	Social Studies	0.0266	0.9211	-0.0819	0.1351	False

Note. Reject column indicates statistical significance at α = .05.

Findings

The sample was broadly balanced by gender and school level, with a larger share of respondents working in urban settings and a spread across core subjects and experience bands. Internal consistency for all six subscales was acceptable to strong, and the full 40-item instrument showed excellent overall reliability indicating that items within each scale cohered well and that the questionnaire measured the intended constructs consistently. Descriptive results placed most subscale means in the moderate range on the 1–5 scale: teachers reported middling perceptions of curriculum–test alignment, accountability pressure, and test-preparation intensity; comparatively higher uptake of formative assessment practices; mixed views on the extent to which current testing improves learning; and neutral-to-slightly-positive views on fairness and accessibility. Taken together, these central tendencies suggest systems in which testing is salient, formative routines are present but variable, and views about impact and fairness are not uniformly positive.

Group comparisons pointed out several regular patterns. Females did seem to be prepping their students for tests a bit more than males, according to their teachers. Urban school teachers, compared to the ones from rural schools, claimed to be under more pressure from accountability, which corresponds to the fact that urban schools are more exposed and thus more closely monitored. Lower, secondary school teachers felt that the curriculum was less aligned with the tests than primary teachers, which could mean that alignment difficulties increase as the content gets more specialized and higher, order across grades.

One, way ANOVA tests revealed that there are significant differences by experience and subject. Teachers with more years under their belt indicated using formative assessment to a

greater extent than their colleagues in the early stages of their career an effect which, although it is not large, can be interpreted educationally and is consistent with the items. The subject matter was also a factor for perceived alignment: science teachers were more prone to report less alignment between assessments and the intended science curriculum than their peers in other subjects, whereas social studies and mathematics teachers were the ones who reported stronger alignment, comparatively. Post, hoc contrasts confirmed these patterns in multiple pairwise comparisons. The correlations among the constructs revealed a consistent picture of practice. The use of formative assessment was highly correlated with perceived impact on student outcomes. This implies that classrooms, which integrate feedback, error analysis, and student self, assessment, are the ones that tend to view testing data as more actionable and improvement, oriented also.

Test-preparation intensity showed a positive association with perceived alignment where teachers saw tests as closer to the curriculum, they were more likely to integrate test-like practice into instruction. Perceived accountability pressure related inversely to fairness perceptions, implying that as consequences tied to scores increase, educators are more likely to question whether assessments are equitable and interpretable for all students. Distribution checks (skewness and kurtosis) for subscale scores were broadly acceptable, with only mild deviations from normality, supporting the use of the parametric tests reported. Overall, the pattern of results suggests a system in which accountability and test preparation are salient, but the strongest levers for perceived impact lie in teachers' formative practices and in improving alignment and fairness especially in subjects and stages where misalignment appears more acute.

Discussion

Standardized testing clearly structures practice in the participating public schools, yet mean scores clustered near the midpoint for alignment, accountability pressure, test-prep intensity, outcomes, and fairness, with formative assessment modestly higher. This pattern is consistent with systems that generate a lot of measurement but still need stronger links from evidence to everyday instruction precisely what balanced assessment systems are designed to address by coherently combining classroom assessment, curriculum-embedded performance tasks, and periodic external monitoring so the system better "measures what matters" (Marion, Pellegrino, & Berman, 2024; OECD, 2023a).

Group contrasts help locate where tensions concentrate. Urban teachers reported higher accountability pressure than rural peers; lower-secondary teachers perceived weaker curriculum–test alignment than primary teachers; and female teachers reported slightly higher test-prep intensity than male teachers. While effects were modest, taken together they indicate that stakes and alignment issues are context-dependent, aligning with evidence that the average achievement effects of test-based accountability are mixed and vary by system conditions (Torres, 2021).

Two structural moderators stood out. More experienced teachers reported greater use of formative assessment than early-career colleagues, and subject area mattered for alignment science teachers perceived the weakest fit between assessments and intended competencies, while social studies and mathematics looked comparatively better. These gradients argue for differentiated capacity building (e.g., novice support for formative routines) and for strengthening moderated performance tasks where alignment is weakest, especially at lower-secondary levels as cognitive demand rises (Marion et al., 2024).

Associations among constructs suggest practical levers. Classrooms with stronger formative routines also reported more positive views of testing's instructional usefulness, echoing syntheses that identify feedback as a reliable driver of learning when implemented well; by contrast, heavy teaching-to-the-test can raise scores on the target exam without commensurate gains in broader

competencies, risking instructional narrowing (Education Endowment Foundation, 2021; Zakharov & Carnoy, 2021).

Policy implications include strengthening validity and fairness routines and adopting responsible digital/AI practices. Routine DIF analyses and accessibility reviews can surface item bias and improve equity, while AI-enabled scoring/analytics require guardrails bias testing, explainability, human-in-the-loop, and privacy so technology enhances, rather than distorts, learning and assessment use (Li, He, & Chen, 2024; OECD, 2023b; U.S. Department of Education, 2023).

Interpretation should be tempered by design limits: the results are observational and rely on self-reports; item-level psychometrics and direct observations were not available for all contexts. Future studies should triangulate survey data with blueprint and item audits, classroom observations, and student growth measures to verify alignment, fairness, and instructional use across subjects and grade bands consistent with current guidance for coherent, multi-source validation in balanced systems (Marion et al., 2024).

Conclusion

The study shows that standardized testing remains a powerful organizing force in public sector schools, yet its perceived alignment with curriculum aims, fairness, and instructional usefulness is uneven. While teachers report moderate accountability pressure and test-preparation intensity, they also describe only partial fit between what is valued in the curriculum and what is measured—especially in lower-secondary grades and science. By contrast, formative assessment practices are more positively perceived and closely associated with views that assessment evidence can meaningfully guide instruction. Overall, these findings imply that systems have been focusing heavily on measurements but are short on actual learning: from tests there are very clear signals, however, the systems that take that hopeful evidence and use that to bring about better teaching and learning on a continuous basis are only partially there. Besides using tests only to send signals, a balanced assessment approach that also includes the cycle of feedback in the classroom, the tasks of performance which are teacher, moderated along with the external checks that are done periodically offers a more reliable way of measuring what really matters while at the same time ensuring the equity and quality of instruction.

Recommendations

1. An assessment model that blends classroom formative assessment, curriculum, embedded performance tasks with moderation, and periodic external tests should be adopted and implemented.
2. A simple "theory of action" should be published stating how evidence from each assessment layer will lead to instructional responses such as re, teaching, targeted tutoring, and pacing adjustments.
3. Set up a cross, functional assessment governance council which will be in charge of reviewing blueprints, doing fairness checks, and establishing annual priorities..
4. Make validity and fairness procedures such as DIF analyses, accessibility reviews, and accommodations audits regular components of your organization, and publish the results publicly each year..

References:

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

ASER Pakistan. (2022). *Annual Status of Education Report (ASER) 2021: National report.* https://aserpakistan.org/

Associated Press of Pakistan. (2024, March 29). NAT results reveal urgent need for reform in mathematics and science education. https://www.app.com.pk/national/nat-results-reveal-urgent-need-for-reform-in-mathematics-and-science-education/

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Axios. (2024, April 11). Harvard reinstates standardized testing requirement, following Yale, MIT. https://www.axios.com/

Bell, D., O'Neill, V., & Crawford, V. (2023). Reliability and validity of methods to assess undergraduate healthcare student performance in pharmacology. *Practitioner Research in Higher Education, 15*(1), 14–23. https://files.eric.ed.gov/fulltext/EJ1409491.pdf

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Brookings Institution. (2023, November 6). Student GPA and test score gaps are growing—and could be slowing pandemic recovery. https://www.brookings.edu/articles/student-gpa-and-test-score-gaps-are-growing-and-could-be-slowing-pandemic-recovery/

Brookings Institution. (2024, September 20). Assessing evidence of academic recovery: Slight progress in math, hardly any in ELA. https://www.brookings.edu/

Brookings Institution. (2025, March 18). Five years after COVID-19 hit: Test data converge on math gains stalled, reading recovery. https://www.brookings.edu/articles/5-years-after-covid-19-hit-test-data-converge-on-math-gains-stalled-reading-recovery/

Education Endowment Foundation. (2021a). *The impact of feedback on student attainment: A systematic review.* https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/feedback-approaches

Education Endowment Foundation. (2021b). *Teacher feedback to improve pupil learning: Guidance report.* https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/feedback

Education Week. (2023, September 1). Educators feel growing pressure for students to perform well on standardized tests. https://www.edweek.org/

Gerasimova, D. (2024). Argument-based approach to validity: Developing a living document and incorporating preregistration. *Journal of Educational Measurement, 61*(2), 252–273. https://doi.org/10.1111/jedm.12385

Idara-e-Taleem-o-Aagahi (ITA). (2024). *ASER Pakistan 2023 National (Rural) report.* https://aserpakistan.org/document/2024/aser_national_2023.pdf

Ismail, S. M., et al. (2022). Formative vs. summative assessment: Impacts on motivation, test anxiety, and self-regulation. *BMC Medical Education, 22*, 653. https://doi.org/10.1186/s12909-022-03667-6

Jerrim, J., & Sims, S. (2024). High-stakes assessments in primary schools: International evidence. *Assessment in Education: Principles, Policy & Practice.* https://www.tandfonline.com/doi/full/10.1080/10627197.2024.2350961

Jerrim, J., et al. (2023). Test anxiety and performance in high-stakes exams. *Oxford Review of Education, 49*(4), 447–473. https://www.tandfonline.com/doi/full/10.1080/03054985.2022.2079616

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kinnear, B. (2024). Validity in the next era of assessment: Consequences, social accountability and equity. *Perspectives on Medical Education, 13*, 454–462. https://pmc.ncbi.nlm.nih.gov/articles/PMC11396166/

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge University Press.

Li, Z., He, L., & Chen, H. (2024). Exploring the evidence to interpret differential item functioning using process data. *Frontiers in Psychology, 15.* https://pmc.ncbi.nlm.nih.gov/articles/PMC11607718/

Marion, S. F. (2018). The opportunities and challenges of a systems approach to assessment. *Educational Measurement: Issues and Practice, 37*(1), 45–48. https://www.nciea.org/

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education. https://naeducation.org/publication/reimagining-balanced-assessment-systems/

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

National Center for Education Statistics. (2024a). *The Nation's Report Card: 2024 reading (grades 4 & 8) highlights.* https://www.nationsreportcard.gov/reports/reading/2024/g4_8/

National Center for Education Statistics. (2024b). *The Nation's Report Card: 2024 mathematics (grade 4) highlights.* https://www.nationsreportcard.gov/reports/mathematics/2024/g4_8/

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

OECD. (2023a). *PISA 2022 results (Volume I): The state of learning and equity in education.* https://doi.org/10.1787/53f23881-en

OECD. (2023b). Digital assessment. In *OECD Digital Education Outlook 2023: Towards an effective digital education ecosystem.* https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en/full-report/digital-assessment_a102e604.html

OECD. (2023c). Opportunities, guidelines and guardrails for effective and equitable use of AI in education. In *OECD Digital Education Outlook 2023: Towards an effective digital education ecosystem.* https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en/full-report/opportunities-guidelines-and-guardrails-for-effective-and-equitable-use-of-ai-in-education_2f0862dc.html

Pakistan Institute of Education. (2024). *NAT findings report 2023.* https://pie.gov.pk/SiteImage/Downloads/NAT%202023%20Findings%20Report%2006.03.2024%20-Final-%20v6%20.pdf

Penney, D., et al. (2023). Enhancing quality and equity? Performance assessment validation trial. *Curriculum Studies in Health and Physical Education, 14*(1). https://www.tandfonline.com/doi/full/10.1080/25742981.2022.2136007

Punjab Examination Commission. (2019–2022). *Assessment Policy Framework (APF) and LSA/Item Bank initiatives.* https://pectaa.edu.pk/apf.php

Reuters. (2024, February 22). Yale University reinstates standardized test requirement. https://www.reuters.com/

Sortwell, A., et al. (2024). A systematic review of meta-analyses on the impact of formative assessment on K–12 learning. *Sustainability, 16*(17), 7826. https://www.mdpi.com/2071-1050/16/17/7826

TEQSA. (2025, October 7). AI cheating is 'all but impossible to detect': Regulator guidance on secure assessment. *The Australian.*

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*). https://doi.org/10.1787/0798600f-en

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations.* https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

UNESCO Global Education Monitoring Report Team. (2023). *Technology in education: A tool on whose terms?* UNESCO. https://unesdoc.unesco.org/

UNESCO (GEM/SDG 4). (2024–2025). *Monitoring SDG 4; SDG Report 2025—Goal 4 (extended).* https://www.unesco.org/gem-report/en/monitoring-sdg4 https://unstats.un.org/sdgs/report/2025/extended-report/Extended-Report-2025_Goal-4.pdf

UNICEF. (2023). *Tracking progress on foundational learning.* https://www.unicef.org/reports/tracking-progress-foundational-learning-2023

von der Embse, N. P., Jester, D., Roy, D., & Post, J. (2018). Test anxiety: A 30-year meta-analytic review. *Journal of School Psychology, 71*, 55–74. https://pubmed.ncbi.nlm.nih.gov/29156362/

World Bank. (2022, June 23). 70% of 10-year-olds now in learning poverty. https://www.worldbank.org/

World Bank. (2024). *Learning poverty: Updates and revisions (April 2024 release).* https://openknowledge.worldbank.org/entities/publication/56452932-59c1-4a10-9216-1bea97e2de8a

Zakharov, A., & Carnoy, M. (2021). Does teaching to the test improve student learning? *International Journal of Educational Development, 84*, 102436. https://doi.org/10.1016/j.ijedudev.2021.102436