



Journal of Educational Research & Human Growth

Volume 1, Issue 1 (2026)



Article Type: Critical Review

Rethinking Standardized Assessment: A Critical Review of Equity, Inclusion, and Student Development

Yengmen Miao^{1*} Jose Y. Haw²

1. Institute of Education, Tsinghua University, Beijing, China

2. Center for the Enhancement of Teaching and Learning, Faculty of Education, The University of Hong Kong, Pok Fu Lam, Hong Kong

Corresponding Author:

Jose Y. Haw

Article Info

Received:

25 May 2026

Revised:

24 July 2026

Accepted:

3 August 2026

DOI:

[10.67808/jerhgv1i1-003-260819093207](https://doi.org/10.67808/jerhgv1i1-003-260819093207)

Abstract

Background:

Assessment for accountability, selection, and performance monitoring continued to be a focus of standardization while the implications of that standardization went beyond measurement.

Methods:

This significant secondary analysis of evidence explored from the fields of scholarship, policy and institutional realms related to issues of equity, inclusion and student development. Materials were analyzed using a structured interpretive process that involved checking claims about control, reliability, validity, fairness, instructional effects, and developmental consequences.

Results:

Five recurring tensions emerged in the synthesis: comparable scores/validity in context; accountability/curriculum narrowing; formal equality/substantive inclusion; pressure on performance/well-being of students; efficiency of the system/Professional judgment. Where standardized tests allowed for comparison across schools, it also highlighted differential outcomes of attainment; however, standardized tests also reinforce inequalities of opportunity where varying levels of preparation, language proficiency, disability access and cultural knowledge were different. With high-stakes use increasing, there is an increase in teaching to the test, anxiety, labelling and strategic exclusion.

Conclusion:

The review recommended an assessor system of moderated teacher judgment, curriculum-based assessments, low-stakes standardized assessments, student portfolio, and equity audit. Reform was the change of purposes, consequences, governance and interpretation practices of tests.

Keywords

Standardized Assessment; Educational Equity; Inclusive Assessment; Student Development; Assessment Validity; Educational Accountability

1. Introduction

Standardized assessment has a strong influence on the education sector today due to potential benefits in comparison, efficiency and transparency. Common tests have been used by governments, school systems, universities, and employers to encapsulate achievement, highlight differences in achievement and to give away opportunities, and as an assessment of institutions (Ajjawi et al., 2022). These functions have contributed to stripping standardized scores of their subjective, mobile, and easy-to-administer looks. But, the apparent accuracy of a score can obscure the fact that classifications may be disputed by others, differences in the measurement of what was learned, by whom and under what circumstances, and some variations in the interpretations of resulting implications (Ansorger, 2021).

So that's no longer about how technically sound standardized tests are. The more fundamental issue relates to whether uniform measurement results in stealthily taken decisions that are consistent with equality in an unequal system, and educationally defensible (Apelehin et al., 2025). Pupils come to the assessment with varying experiences of experienced teachers, secure homes, food, internet, tutoring, language support, health insurance and culturally appropriate curriculum (Baniyadi et al., 2023). A common test could be implemented using the same procedures of a distribution yet measuring an opportunity distribution that was never equal. Formal equality can thus go hand in hand with equality which is essentially unfair.

The standardization also impacts the ways schools

teach and students' conceptions of learning. Teachers and leaders make pragmatic changes in practice when the consequences of results impact promotion, admission, funding or inspection and/or institutional reputation (Bulut et al., 2024). This response may be helpful in some cases of diffused curricula, but it may also have the negative effect of reducing the focus, creating privilege for knowledge easily scored through tests and examinations, and ultimately leaving less time available for creativity, discussion, collaboration, inquiry and social interactions during the implementation of the curriculum. The test moves from being a measurement device to a technology of control which influences behavior.

Consider three interconnected lenses for a rethinking of assessment, equity, inclusion and student development. Equity is about the fairness of the ways assessments and the outcomes of the assessment are distributed. Inclusion is about students with disabilities, multilingual learners, culturally diverse learners and those disadvantaged in learning to be able to demonstrate learning without irrelevant barriers (Cairney & Kippin, 2022). Student development focuses on the impact of assessment on motivation, identity, the building of confidence, curiosity, self-regulation, a sense of belonging and longer-term engagement. The same system has the potential of generating statistically similar scores and yet failing in each of the three areas. Uses and consequences of the use of standardization in assessment were critically appraised and analyzed secondarily. It was posed 4 questions.

- **RQ1:** What motives could these standardized tests have for legitimate educational purposes?
- **RQ2:** What was the impact of the test design and high-stakes testing on various groups of students?
- **RQ3:** How did standardized assessment impact teaching, motivation, identity and broader development?
- **RQ4:** What is an architecture for assessment that would do so without a single score becoming the predominant way things are judged educationally?

The paper made a basic claim that standardized measures should be used as partial evidence in a plural landscape of assessment, instead of as a statement of ability, merit, school quality or future prospects. Comparability proved to have its usefulness, but no

more than in the case of tightly drawn purposes, less than dire consequences, situational interpretations and truly weighty alternative evidence. It was thus necessary to carry out reform not only with better test items but also with other means. It involved rearranging institutional ties between measurement, curriculum, accountability, and professional judgement, as well as voice in the learning process and students.

2. Conceptual Foundations

Standardized assessment was defined as a measurement whose procedures of administration, scoring and interpretation are shared with the community. Standardized assessments were defined as a 'measurement administered, scored, and interpreted under the rules shared with the community (Equitable, 2025). By controlling for some of the variation due to how the test was administered and how items were scored, the results of every individual, school or across years could be compared. But reliability did not make for validity. A score might be consistently achieved, but it reflects a limited aspect of learning or may be based on a decision without reference to the evidence produced.

Validity was considered to be a question of "how" to use and interpret findings. Of note, it was the propriety of a specific inference from a particular score, under specific circumstances, that was being questioned, and not the validity of the test in the abstract. Therefore, a mathematics score may need to be translated to reflect performance on one or more numerical tasks sampled from the measurement, but cannot necessarily be used to automatically assume performance on other measures such as creativity, persistence, quality of teaching or the effectiveness of the school (Fien et al., 2021). High-stakes systems tended to have loose interpretations, which were moved beyond the level of evidence.

Equity and Equality are distinguished. Equality would mean following consistent procedures, equality would require attention to unequal conditions, and/or consequences; equity would not. An equitable assessment system had students rate their experience as students trying to learn the content that was assessed, have students rate whether there were any factors that limited their ability to learn that

content, and have students rate whether there were any factors that hurt their performance more so than other groups of students. It also explored who would benefit from the use of sorting by scores, and who would run the risk of being misclassified.

Inclusion was seen as the inclusion with dignity and interpretative justice. Accommodations (be it providing extra time, assistive technology, accessible formats, breaks, or alternative modes of responding) would decrease the barriers but, in themselves, were not a sign of inclusion (Garcia et al., 2025). Exclusionary tests might occur when the language requirements have nothing to do with the underlying concept being assessed, when the language examples require a certain type of background knowledge that a certain social or cultural group may not possess, or when rules for administration are based on the idea that all learners are in some way deviant.

A wide approach to the conception of student development was taken. Education was about academic knowledge, but also acquired agency, identity, relationships, ethical judgement, emotion control, communication, imagination and agency to act collaboratively. These outcomes were sampled only as part of standardized tests. They also indirectly and unconsciously exerted their developmental impact by giving feedback, comparison, ranking information, preparation information and placement decisions.

The assessment was made based on the following assessment ecology approach (Garver, 2020). Assessment was not considered as one event, but a set of related purposes, actors, instruments, interpretations, incentives and consequences. A technically sound test might have negative consequences when coupled with other consequences. A small, standardized measure, however, might be useful in addition to teacher evidence, discussion of diagnostics and low-stakes monitoring. So the ecology, the whole idea, is what was important, and not the instrument in itself.

3. Methodology

This study was considered to be a critical review and qualitative secondary analysis. This perspective was chosen because the research question needed interpretation in various fields of study, from

measurement theory, educational policy, inclusion, psychology, sociology, and the classroom context. The aim was not to arrive at some sort of pooled effect size, but to analyse some of the claims, inconsistencies, mechanisms, and institutional results that tend to recur.

Scholarly reviews, empirical studies, policy analyses, technical reports, institutional evaluations and documented student and/or teacher perspectives about standardised assessment constituted a purposive evidence corpus. Materials were deemed relevant based on their consideration of at least one of four areas: measurement validity, educational equity, inclusive access, and/or student development. Primary, secondary and transitional education evidence was kept that shed light on mechanisms at the system level. In order to eliminate purely technical studies without implications for their applications or for their interpretation, the author excluded from the above-mentioned collection those papers.

Structured extraction was used to analyse the data obtained from the participants' responses in the course of the study. For the assessment, each source was assessed for the following: Purpose, population, stakes, construct measured, reported benefits, reported harms, affected groups, contextual moderators, and proposed reforms. Some of the original codes involved 'Comparability', 'Opportunity to learn', 'Cultural loading', 'Language demand', 'Disability access', 'Curriculum narrowing', 'Testing to the test', 'Anxiety', 'Motivation', 'Labelling', 'Score inflation', 'Professional judgements', 'Accountability Incentives'.

A comparison of codes was made between different

types of evidence and themes. Any findings that contradicted each other were retained. For instance, the use of standardized testing was coded both as a mechanism for the revealing of achievement gaps as well as a mechanism that can re-create an unequal opportunity. Potentially beneficial curriculum review and potentially harmful narrowing (based on intensity, alignment and stakes) were coded as examples of test preparation. This way of doing things avoided the self-serving prescription "one size fits all" for standardized assessment, as well as the assumption that standardized assessment universally harmed.

There were five themes that emerged as the valid results: constrained validity, distributive inequity, incomplete inclusion, developmental consequences and accountability distortion. These themes were then woven into a well-designed assessment ecology that included standardized measures, but delineated their limited role and the complementary functions of teacher judgments, performance tasks, portfolio, formative assessments, and student reflections.

Reflexivity was considered by separating evidence pertaining to the properties of tests from evidence pertaining to the use of tests. Different purposes have been distinguished from the consequences that were observed. Where claims relied on advocacy, commercial interest or small samples, these were addressed in a considered manner. No subjects were recruited, and data were not discernible. Thus, the need for ethical approval was not necessary. The review does, however, make a moral critique of the effects of classification, exclusion and high-stakes decision-making.

Table 1. Critical-review design, evidence boundaries, and interpretive safeguards

Element	Specification	Analytical contribution
Review design	Critical review and qualitative secondary analysis	Supported interpretation across measurement, policy, inclusion, psychology, and classroom practice.
Evidence corpus	Scholarly reviews, empirical studies, technical reports, policy analyses, institutional evaluations, and documented perspectives	Included evidence addressing validity, equity, inclusion, or student development.

Analytical unit	Claims about purposes, constructs, stakes, mechanisms, consequences, affected groups, and proposed reforms	Separated properties of tests from consequences of their use.
Coding domains	Comparability, opportunity to learn, language, disability, cultural loading, curriculum narrowing, anxiety, motivation, accountability, and alternatives	Preserved positive, negative, mixed, and conditional evidence.
Synthesis method	Comparative coding, thematic grouping, contradiction analysis, and integrative interpretation	Generated five critical themes and a balanced assessment ecology.
Safeguards	Reflexive distinction between intended purposes and observed effects; scrutiny of advocacy and commercial claims	Reduced overgeneralization and maintained interpretive transparency.

4. Findings

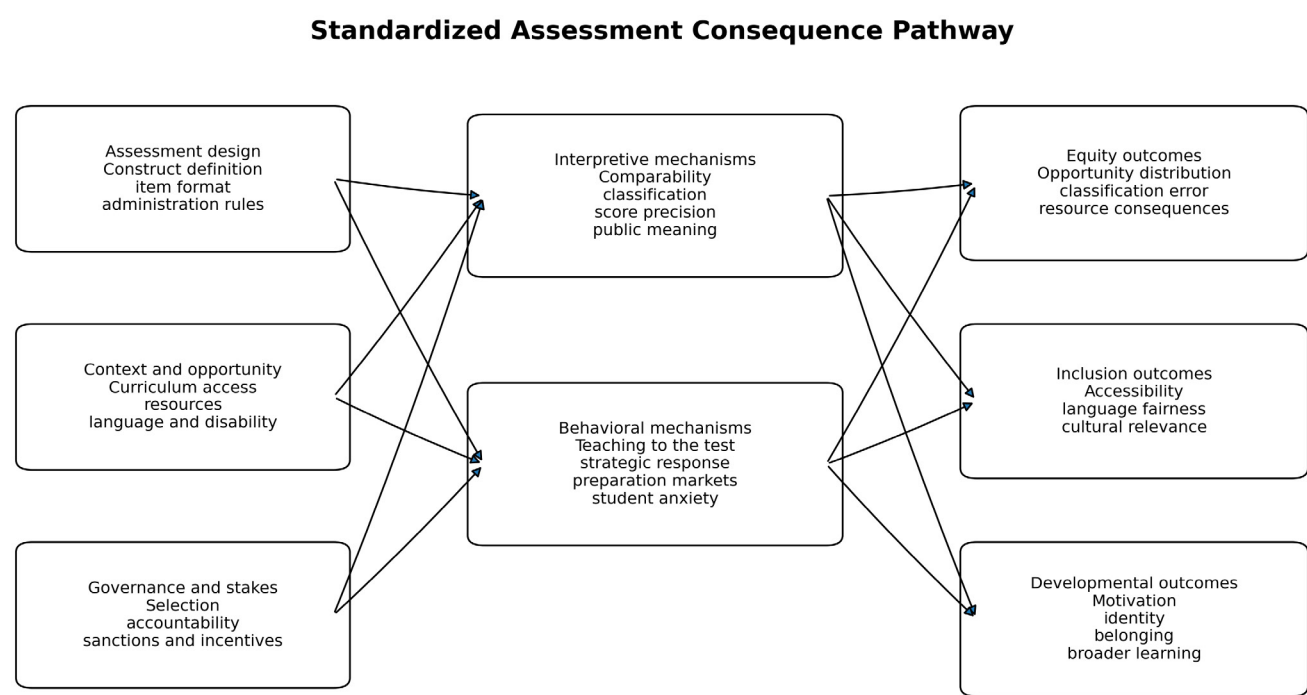
4.1 Constrained validity and the problem of overinterpretation

Selected knowledge was summarised efficiently in standardized tests, which were conducted under controlled conditions. Out of these, their best was comparability. Systems honed in on overall performance trends and tracked progress and identified population groups in need of attention through common tasks and scoring (Hargreaves, 2020). This information might prove to be a wake-up call if they generally felt complacent, looking at the differences that had been discovered.

However, there was a loss in their validity when scores were used as overall indicators. Only basic knowledge, content and limited time and response formats were tested. They tended to favour swift performance of individuals, problem-solving without context and knowledge of test conventions. It was hard to convey complex skills and abilities like collaboration, oral rationale, enquiry, creativity, practical skills and ethics with traditional items.

Overinterpretation was when insufficient data were used to rate teachers, schools or intelligence, potential or institutional quality. These inferences omitted to consider sampling error, curriculum variation, student mobility, language, disability or opportunity to learn. Numbers, because they seemed “objective,” lead the decision makers to ignore the “assumptions” needed to make the leap from performance with the test items to making decisions about merit (Hauer et al., 2023).

Also, false certainty was the score’s precision. Even if a small difference occurred within a plausible range of measurement, this difference might result in a placement or access difference. The notion of cut scores pushed continuous categorizations of performance to profiling/aggregating into categories (e.g., proficient, below standard, and eligible), thereby creating the impression that there were fixed classes of different kinds of students near the cut scores who were in the transition zone. It was an administratively expedient classification that was educationally precarious.



The same instrument produced different consequences when purpose, stakes, opportunity, accessibility, and interpretation changed.

Figure 1. Pathways through which standardized assessment design and governance shaped equity, inclusion, and development

4.2 Distributive equity and unequal opportunities to perform

The review revealed that accumulated opportunity was sometimes assessed as part of a standardized assessment process in addition to current attainment. The students who came from wealthier families tended to be more likely to have access to tutoring, a quiet space to study, enrichment activities, amongst other things, whilst adults in the households also tended to be more knowledgeable about the education system. Test preparation markets were able to translate monetary investments into “test score” advantages, especially when test-taking was necessary for admission to top programs or institutions.

Another aspect that influenced performance was school resources. Opportunity to learn was affected by experienced teachers, technology, libraries, coherent curricula and small class size and specialist support. Without the ability to account for unequal “inputs” or contributions, there was danger in holding schools accountable for their “outputs” (Herrera et al., 2024).

Though the use of standardized testing could help ensure equity if used diagnostically. Disaggregated results allowed persistent gaps to be seen and excluded average performance to mask exclusion. This was to change based on the equity value. Results that lead to further resources, curriculum support and investigation could be used to enhance accountability for the underserved groups through testing. The same information could also exacerbate inequalities if failure to engage in this process in the context of results would lead to sanctions, closure or reputational damage.

Strategic behaviour was the result of high-stakes consequences. Schools might focus resources on pupils with scores close to the thresholds, pupils who have not yet reached the thresholds might be excluded from the groups used for testing, schools might discourage pupils who need support from taking tests, or schools might ‘narrow’ teaching for groups which might be considered at risk. Not all of these reactions were necessarily from unethical people, but were typical responses to incentive structures.

4.3 Inclusion, accessibility, and construct-irrelevant barriers

Students with disabilities experienced difficulties where test formats required skills that wasn't the target construct. For example, a science test may be transformed into a reading speed test, a mathematics point may be based on visual presentation, and a written answer may negatively affect a person due to motor issues. However, some barriers were lowered through accommodations, while inconsistent eligibility and the complexity of administering the accommodations restricted access.

The review recognized that there is a conflict of standardization and adaptation. The stronger the uniformity, the easier it was to maintain a level of comparability; with individual access, the more flexible the access (Huber & Helm, 2020). It was not necessary that the conditions of each student be exactly the same for the conditions to be considered fair. It called for some modification in administration to break through the irrelevant obstacles, but it did not bring about any alterations in the construct. This distinction called for design and professional judgment.

Issues were also similar for multilingual learners. Testing in a dominant language may result in incorrect inferences about knowledge of a subject matter vs the language proficiency. Sometimes, minimally changed text enhanced comprehension, while at other times inadequate simplifications increased or decreased the meaning of the disciplinary language and/or changed the level of difficulty. There were also risks involved with translation, as the concepts, contexts, and expectations of the rhetorical listeners did not align in a seamless manner.

Interpretation was shaped by familiarity with the culture/s. There might be benefits to item contexts, names, experiences and reasoning expectations that benefit students who live up to the expectations of the test. Some differential functioning will be identified through statistical checks, but this does not solve technical problems with cultural issues. This meant including a variety of design teams, conducting cognitive interviews, having accessible trials and subsequent review following implementation.

4.4 Curriculum narrowing and instructional

distortion

The testing environment of high stakes impacted the priorities of the curriculum. Arts, physical education, citizenship, practical learning, and interdisciplinary inquiry were the subjects and/or skills that were at risk of being reduced, while subjects/skills included in the accountability metrics received more time. Teaching shifted to variations in the form of questions and regular topics.

Educationally responsible alignment was present in some cases. The common assessments and standards might help promote focus on basal literacy and numeracy. Troubles started to crop up when mimicry took the place of alignment. Test-type questions were repeated, which resulted in an increased familiarity but led to the displacement of explanation/exploration and transfer. Students acquired skills in noticing answer patterns in the absence of long-term understanding.

Teachers who were under pressure would use scripted programs that test to narrow targets, and often, frequent benchmark tests and data meetings would focus on predicted test scores. There was more assessment information, much of which was used for accountability purposes, not learning (Jagers et al., 2025). Time taken to administer and analyze interim tests might be at the expense of time for classroom support.

Additionally, professional judgment was impacted. There was potential for teachers to increasingly privilege external scores over classroom evidence, although the evidence in some cases was signs of deeper learning. On the other hand, if testing is distrusted, resistance to comparative information which could be of benefit may surface. Nor submission to scores nor rejection of measurement was an appropriate solution for a balanced system.

4.5 Motivation, identity, and student development

Assessment was standardized and messages of ability and/or value communicated. Students frequently saw scores as making judgments on the nature of the person's intelligence as opposed to a performance in certain conditions. Students may become marginalized in low classification if they are excluded from the intervention or access to advanced work or determine grouping.

Concern about tests affected and decreased focus, working memory, and perseverance. For some, a level of challenge may have encouraged them to prepare better, where high stakes may have increased their fear of failing or the repercussions of failure for either themselves, their teachers or their school. Anxiety wasn't equally spread across the board. For students who have had negative experiences in the past, those who are perfectionist, or those with disabilities or language concerns, or who think that others may not be expecting them to excel, there may be some more interference.

Motivations shifted to a performance orientation when success was considered in terms of ranking. The students began to place less value on grades and scores as compared to curiosity, revision, cooperation, or intellectual risk (Khasawneh & Khasawneh, 2023). For the untested activities, they did not seem to be as valuable. Sometimes assignments may be used for external comparison, which provides motivation for students; on the other hand, it is possible that some students may experience avoidance because they believe that they would get low marks.

Relations to the consequences of development. The pressure of competitiveness may negatively affect peer cooperation, and the pressure at school may affect the teachers' interaction with their students. The teacher could focus on students who would have an impact on the performance indicators, which may unwittingly diminish the attention given to others.

The review was not in favour of a removal of evaluation. Clear expectations, feedback and opportunity for students to demonstrate progress benefited students. The developmental theme was about the nature and way in which the assessment was to be done. Feedback that was specific, revisable and related to strategies was a positive support to growth. One benchmark score offered little advice as it gave an overall picture of performance and not what to do to improve.

4.6 Accountability, governance, and institutional behaviour

Standardized assessment was most problematic when it came to bear responsibilities it could not carry. Scores were chosen by policymakers as they were easily visible and scalable and could be communicated

to the political sphere. But issues like the quality of a school or the effectiveness of its teachers would not be amenable to a single measure or scale.

Feedback loops were created with accountability systems. Targets shaped teachers, selection of students, allocation of resources, teacher and student report system and teacher leader priorities. When scores became "high-stakes indicators, the whole thing was 'tainted' as it served as a 'neutral' observation, then as participants got used to scores, they became more like a 'trained' observation. An increase in results may be due to learning, familiarity, narrowing of the curriculum, selective participation and/or change in cohort composition.

The quality of governance was related to transparency. Stakeholders wanted to be informed about what was being tested, how the scores were scaffolded, how uncertainty was shown, the options for accommodations, and how the results informed decisions. Proprietary systems also tended to evade scrutiny, reducing public accountability.

Standardized samples may need to be used more heavily in low-stakes monitoring, and richer evidence and review could be used for making individual decisions of a high-stakes nature. This process was thus not viewed as a way of introducing optional exceptions, but was a requirement, subject to appeals, contextual information and human judgment.

4.7 Alternatives and the case for a balanced assessment ecology

None of the alternatives completely addressed all the limitations. Teacher assessment had been used to reflect ongoing achievement and understanding of the context, though there was potential for personal expectations, workload and the lack of uniformity of standards to be a factor. Portfolios were growing in complexity, but were potentially difficult to moderate and compare. The performance tasks were expensive in terms of design and scoring and measured the application and reasoning. The self-assessment of students was assisting in the realization of agency, but it was not a substitute for external assessment.

The synthesis, thus, favoured a combination (rather than substitution). Limited diagnostic signalling could be provided with standardization of tests on a system-level basis. There is potential to use

classroom assessment to inform classroom practice. Achievement over time could be moderated by teacher judgment. Application, communication and collaboration can be captured with performance tasks. Portfolios might evidence growth, revisions and a range of strengths (Meylani, 2024). Student reflection is the opportunity to develop metacognition and ownership. Multiple measures only presented any value if there was integration and balance between the measures. New additions to the system in the form of portfolios would have no effect on the system of decisive SLPs. Progression, reporting and evaluation required some

formal weight to be given to the alternates in this. Moderation processes needed to be put in place to allow for consistency, whilst not forcing elements of evidence to a common template. Equity auditing was also a part of the balanced assessment. Systems should consider issues of participation, accommodation, classification error, subgroups, curriculum impact, and preparation input. To interpret and apply student and teacher views. They had achieved the best results in the field of assessment reform, where there had been an improvement in information quality and in the educational process.

Table 2. Critical synthesis of benefits, risks, and reform requirements

Dimension	Potential contribution	Critical risk	Required safeguard
Validity	Reliable comparison of selected content	Scores are overinterpreted as intelligence, potential, teacher quality, or school quality.	Restrict claims to the measured construct and intended decision.
Equity	Disaggregated results exposed persistent attainment gaps	Unequal preparation, tutoring, resources, and opportunity shaped performance	Audit opportunity to learn and resource consequences.
Inclusion	Common expectations and accommodations supported participation	Language, disability, format, and cultural familiarity created irrelevant barriers	Design accessible tasks and permit construct-preserving flexibility.
Curriculum	Standards encouraged attention to foundational knowledge	High stakes, narrowed subjects, pedagogy, and intellectual risk	Monitor curriculum completeness and reduce punitive consequences.
Student development	Clear expectations sometimes encouraged preparation	Anxiety, labelling, ranking, and score-based identity weakened motivation	Use specific feedback, revision, and multiple demonstrations.
Governance	Scalable data-supported monitoring and public reporting	Thresholds, rankings, and sanctions generated strategic behavior	Use proportional stakes, transparency, appeals, and contextual review.
Professional judgment	External evidence challenged inconsistent local expectations	Scores displaced richer classroom knowledge or provoked total rejection	Moderate professional evidence and develop assessment literacy.

5. Discussion

The review revealed that standardization of assessment is neither in itself fair nor unfair. Its impact varied according to the intentions, objectives, and risks associated with its use, how long it was used, and

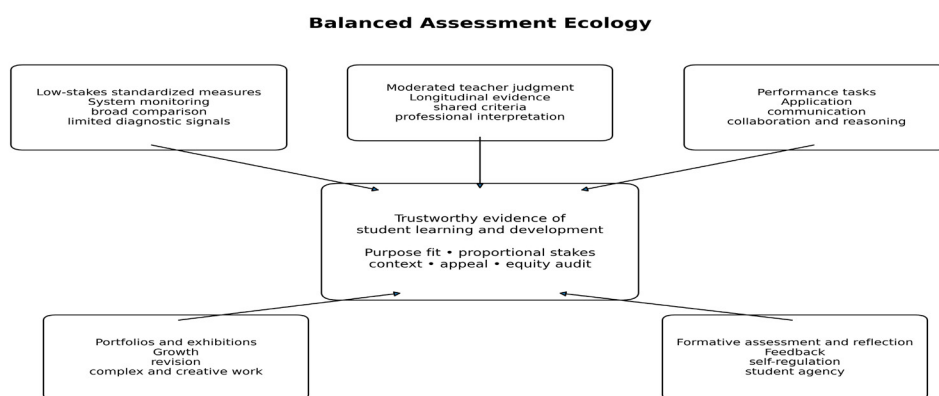
the context. Comparability was the one dimension of quality that was provided by standardization. Common measures may show patterns and be used in a narrow evidentiary role to assist in monitoring systems. As dominant judgments they created, they

reduced and simplified the complexity of education, and amplified inequality. Numerous test-related ill effects stemmed from the consequences of tests, and not just from the design of tests. A moderately valid test lost its validity when tied to punitive sanctions, hard limits or only-in programmes. Those who are using more will need more technical improvement; that wouldn't help them (Murillo & Hidalgo, 2020). From equal treatment to equal opportunity for performance of learning was a requirement for equity. This transition didn't imply no standards (Tobón et al., 2020). That involved taking a look at opportunity to learn, individual language, disability issues, cultural context and consequences of error. It also required the understanding that changes in scoring might not indicate skill learning differences, but might just be a relation to different environments (Walker et al., 2023). For inclusion, the design for the variability of the learner was a wider way of thinking from the start. Some use of adaptations was still required, but the use of the one-to-many modes and varying the timing made exceptional options less necessary (Paramole & Adeoye, 2024). But there was a requirement for restraint in the interpretation of "inclusion. Language and/or format effects should not blind a rater to a subject knowledge impact. The broadened criterion for assessment quality was the result of student development. Only reliability and prediction should not be the criteria to determine if a system is a good one or not. It is also to be reviewed based on the effect it had on motivation, identity, belongingness, curiosity, and the desire to learn. It would be impossible to view tests which guaranteed

a correct ranking together with a diminished interest in education as an educational success. The findings favored a balanced approach for assessing an ecology that is centred on the criteria of "fitness for purpose. There were a number of purposes, all of which involved different types of evidence: system monitoring, classroom diagnosis, certification, selection, and student reflection. Purpose separation was accomplished, and the overinterpretation was eliminated, enabling the selection of the forms used in the assessment based on their strengths (Smith & Holloway, 2020).

Professional judgment has been a key. While human judgment may be biased, there were no values or discretion eliminated by removing human judgment, but they were simply moved to test design, algorithms, thresholds, and policy rules. Disciplined judgment with moderation, evidence, transparency and appeal was the appropriate response. Likewise, teachers should not have to make decisions but are not held accountable, nor be pushed to administer external scores (Stacey et al., 2020).

A second recommendation was for the idea of assessment literacy as a system capacity to be explored. Some understanding of uncertainty, purpose, validity and appropriate inference by students, their families, teachers, leaders and policymakers was needed (Taylor, 2022). A score's significance beyond the bounds of their literacy. As a result, such public communication needs to focus on ranges, patterns and what elements of context should be taken into account, rather than on league-table certainties.



Balance required formal decision weight, moderation, accessibility, transparency, and continuous review of subgroup consequences.

Figure 2. Balanced assessment ecology integrating standardized, professional, performance, portfolio, and formative evidence.

6. Policy And Practice Framework

A reform concept was devised based on six principles. First, the instrument selection serves the purpose. All decisions in relation to assessments should be included, and the evidence needed. Individual progression should not be set on the basis of tests set up for system monitoring.

Secondly, there shall be proportionality to the validity. The more at stake, the more needed the evidence, the more the possibility for reviewing, and the more safeguards there were, the more safeguards there were for the situation. Access/exclusion/retention/institutional sanction should not be based on one score only.

Third, curriculum-completeness should be the focus of assessment. Systems need to look at whether tested content pushed aside/overlooked larger educational objectives. The quality of a school should be assessed, as well as against its academic attainment, by the participation and well-being of the student, the feeling of “inclusion,” creative work and developmental progress.

Fourth, be careful not to simply add to inclusion; try to design it overall. Accessible formats and language review, multiple response options, and AT should be designed in rather than modified. Accessible formats, language review, multiple response options, AT, and flexible administration should be considered or built into development, rather than modification. This should look as simple, consistent and stigma-free as possible when processes are being decided on the accommodations.

Moderation of professional evidence. It was necessary to coordinate teacher judgments, portfolios, and performance tasks, to have illustrations available for use, to review work together and periodically calibrate. Moderation enhanced comparability and maintained the local richness.

Sixth, ongoing equity auditing needs to take place. The following should be taken into account by officials: score distributions, misclassifications,

non-participation, accommodation usage, access to test preparation, narrowing of curriculum, and downstream placement. Different outcomes of assessment are not a technical consequence, but instead should trigger a change in policy if harm is done to one or more groups as a result of that assessment.

The framework recommended that schools develop an evidence profile with standardized elements of little value, an assessment process that was based on curriculum, an extended performance task, moderated teacher judgment, student reflection and “contextual notes” for sharing with students. The profile should display ‘change over time’, not just ‘endpoint’. Follow-up feedback should be given to students, in an informal way, outlining strengths, further steps and support.

Where it wasn’t necessary to have an individual student’s score, there should be sampling for accountability. The use of samples would decrease the burden of testing, strategic distortion, and maintain information about the population. Institutional evaluation needs to be a combination of outcomes and resources, inclusion and process evidence. When using review teams, the focus should help to identify the causes for concerns and not make judgments about quality based on a number.

Multiple pathways should be used for selection, where assessment is concerned. Common measures might be supplemented by prior attainment, contextual information, portfolios, interviews, recommendations and/or demonstrated performance. Appeal and reconsideration to be part of decisions, especially in regions of thresholds.

Implementation required investment. It took time and expertise to assess performance, to moderate performances, to determine accessibility and to provide feedback. Rich assessment couldn’t be obtained by simply requesting teachers to include more assessment into current assignments. The focus should be shifted away from over-testing, towards commercial preparation and the multi-production of data, towards collaboration and the learning of professionals.

Table 3. Implementation roadmap for equitable and developmentally responsible assessment reform

Reform stage	Required action	Monitoring evidence
1. Clarify purpose	State the decision, construct, population, and intended interpretation before selecting an instrument	Purpose-use register; validity review; prohibited uses
2. Reduce score dominance	Prevent one result from determining access, retention, selection, or institutional sanction	Multiple-evidence rule; threshold review; appeal rate
3. Build inclusion	Embed accessible formats, language review, assistive technology, and flexible administration	Participation gaps; accommodation uptake; accessibility testing
4. Protect curriculum breadth	Monitor displacement of arts, inquiry, citizenship, practical learning, and collaboration	Instructional-time audit; curriculum breadth indicators
5. Moderate professional evidence	Use shared criteria, exemplars, collaborative review, and calibration for teacher judgment	Inter-rater consistency; moderation completion; feedback quality
6. Audit equity effects	Examine subgroup classification, preparation access, error consequences, and downstream placement	False-positive and false-negative patterns; resource response
7. Report development	Include growth, revision, student reflection, well-being, and learning strategies	Narrative profiles; portfolio quality; student agency indicators
8. Review governance	Disclose scoring, uncertainty, data use, commercial interests, and decision rules	Transparency audit; independent oversight; stakeholder trust

7. Limitations and Future Research

This review was a synthesis of multiple pieces of evidence without a formal meta-analysis. The ages of the materials included ranged from elementary to high school; they came from various jurisdictions and included different types of tests, with varying stakes and methodological quality. Mechanisms and recurring patterns were identified for the analysis, but the prevalence and magnitude of effect for each of these were not estimated.

The secondary design relied on reference evidence, and there was evidence, published references and publicly documented evidence. Students who did not participate in testing, were not engaged in learning, and/or did not have access to research may have been underrepresented. Favourable interpretations also might be placed in a commercial or government report.

Standardized assessment included any of the following: Achievement tests, admission testing,

accountability tests, benchmarks, and screening instruments. They had quite divergent objectives. Drawing inferences should accordingly be made based on context and not as general negations.

Future research should study the relative effectiveness of balanced assessment systems through a longer time frame and across a number of areas, including achievement, equity, well-being, teacher workload and public trust. Research should look at errors in classification according to overlapping identities, and not each of the four of them separately, disabling, disadvantaged, language and gender. Another research area should be the impact on transparency, bias and professional judgment in the use of artificial intelligence, automated scoring, and adaptive testing. There is a need for an increased focus on pupil voice. The schedules of feedback, reporting, appeals and assessment should involve young people in their design. The cost research was also necessary as richer systems could require some extra time and effort for the professional to manage, but would save money spent on redundant testing, remediation and disengagement in systems costlier than these.

8. Conclusion

The idea of standardized assessment was good because it provided a kind of comparability, but it was educationally unhealthy to view the numbers as absolute, context-laden judgments. That design of tests, along with the stakes, interpretation, accountability incentives and opportunity to learn, was demonstrated to influence equity, inclusion and student development in the review.

The fair system didn't end measurement. It limited each of the measures to the ends that it was able to serve. When making judgments, standardized evidence should inform, but not dictate, judgment. Multiple sources, context review, accessible participation, and meaning of appeal were required for high-stakes decisions.

The proposed framework of a balanced assessment ecology involved a mixture of low-stakes standardised assessments, moderated teacher judgements, in situ performance-based assessment, portfolios, summative feedback and student reflections. With this design, shared expectations were maintained, and there were different approaches to expressing learning.

This necessitated a change of paradigm in thinking about assessment from that of identifying pupil rank to that of identifying development. The key issue was how to create credible evidence whilst not restricting the breadth of education, or portraying more positively than they are in fact, students learning. Systems were more likely to facilitate accountability and humane education when they were in balance that were both comparable and contextual, attended to professional judgment, and were inclusive and developmentally focused.

Declarations

Author Contributions

Yengmen Miao conceived the study, developed the conceptual framework, identified the research problem, designed the review approach, conducted the literature search and evidence selection, and prepared the initial manuscript draft. Jose Y. Haw developed the methodological framework, carried out the qualitative secondary analysis and thematic synthesis, critically interpreted the findings, refined

the policy and practice implications, substantially revised the manuscript for intellectual content, and supervised the overall scholarly development of the study. Both authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgement:

The author acknowledges educators, learners, and researchers whose documented experiences informed this critical secondary review.

Funding statement:

This research received no external funding and relied solely on publicly available secondary evidence sources.

Competing interests:

The author declares no competing interests influencing the review design, analysis, interpretation, or reported conclusions.

Ethics approval:

Not applicable because this study analyzed publicly available, non-identifiable secondary evidence.

Data availability:

No primary dataset was generated; the analytical framework and synthesis are fully reported herein.

References

- Ajjawi, R., Tai, J., Boud, D., & de St Jorre, T. J. (2022). *Assessment for inclusion in higher education: Promoting equity and social justice in assessment*. Taylor & Francis.
- Ansorger, J. (2021). An analysis of education reforms and assessment in the core subjects using an adapted Maslow's hierarchy: Pre and post COVID-19. *Education Sciences*, 11(8), 376. <https://doi.org/https://doi.org/10.3390/educsci11080376>
- Apelehin, A. A., Imohiosen, C. E., Ajuluchukwu, P., Abutu, D. E., Udeh, C., & Iguma, D. (2025). *Assessment and evaluation for social improvement in education: Strategies for equity and fairness*. *International Journal of Social Science Exceptional*

- Research. 2025b, 4(1), 119-125.
- Baniasadi, A., Salehi, K., Khodaie, E., Bagheri Noaparast, K., & Izanloo, B. (2023). Fairness in classroom assessment: A systematic review. *The Asia-Pacific Education Researcher*, 32(1), 91-109. <https://doi.org/https://doi.org/10.1007/s40299-021-00636-z>
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., & Walsh, C. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. arXiv preprint arXiv:2406.18900. <https://doi.org/https://doi.org/10.48550/arXiv.2406.18900>
- Cairney, P., & Kippin, S. (2022). The future of education equity policy in a COVID-19 world: A qualitative systematic review of lessons from education policymaking. *Open Research Europe*, 1, 78. <https://doi.org/https://doi.org/10.12688/openreseurope.13834.2>
- Equitable, I. (2025). Transforming Educational Assessment. Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy, 99. <https://doi.org/https://doi.org/10.1002/rrq.391>
- Fien, H., Chard, D. J., & Baker, S. K. (2021). Can the evidence revolution and multi-tiered systems of support improve education equity and reading achievement? *Reading Research Quarterly*, 56, S105-S118.
- Garcia, M. B., Rosak-Szyrocka, J., Yilmaz, R., Metwally, A. H. S., Acut, D. P., Ofosu-Ampong, K., Erdoğan, F., Fung, C. Y., & Bozkurt, A. (2025). Rethinking educational assessment in the age of generative AI: actionable strategies to mitigate academic dishonesty. In *Pitfalls of AI integration in education: Skill obsolescence, misuse, and bias* (pp. 1-24). IGI Global Scientific Publishing. <https://doi.org/DOI:10.4018/979-8-3373-0122-8.ch001>
- Garver, R. (2020). Evaluative relationships: teacher accountability and professional culture. *Journal of Education Policy*, 35(5), 623-647. <https://doi.org/DOI:10.1080/02680939.2019.1566972>
- Hargreaves, A. (2020). Large-scale assessments and their effects: The case of mid-stakes tests in Ontario. *Journal of educational change*, 21(3), 393-420. <https://doi.org/https://doi.org/10.1007/s10833-020-09380-5>
- Hauer, K. E., Park, Y. S., Bullock, J. L., & Tekian, A. (2023). "My assessments are biased!" Measurement and sociocultural approaches to achieve fairness in assessment in medical education. *Academic Medicine*, 98(Supplement_2), S16-S27. <https://doi.org/https://doi.org/10.1097/ACM.0000000000005245>
- Herrera, S., Herrera, S. G., Murry, K. G., & Cabral, R. M. (2024). Assessment of culturally and linguistically diverse students. *BoD—Books on Demand*. <https://doi.org/https://doi.org/10.1007/s11092-020-09322-y>
- Huber, S. G., & Helm, C. (2020). COVID-19 and schooling: evaluation, assessment and accountability in times of crises—reacting quickly to explore key issues for policy, practice and research with the school barometer: Huber SG, Helm C. *Educational assessment, evaluation and accountability*, 32(2), 237-270.
- Jagers, R. J., Rivas-Drake, D., & Williams, B. (2025). Transformative social and emotional learning (SEL): Toward SEL in service of educational equity and excellence. In *Social and emotional learning* (pp. 39-62). Routledge.
- Khasawneh, M. A. S., & Khasawneh, Y. J. A. (2023). Achieving assessment equity and fairness: Identifying and eliminating bias in assessment tools and practices. <https://doi.org/https://doi.org/10.20944/preprints202306.0730.v1>
- Meylani, R. (2024). A comparative analysis of traditional and modern approaches to assessment and evaluation in education. *Bati anadolu eğitim*

bilimleri dergisi, 15(1), 520-555. <https://doi.org/https://doi.org/10.51460/baebd.1386737>

Murillo, F. J., & Hidalgo, N. (2020). Fair student assessment: A phenomenographic study on teachers' conceptions. *Studies in Educational Evaluation*, 65, 100860. <https://doi.org/https://doi.org/10.1016/j.stueduc.2020.100860>

Paramole, O. C., & Adeoye, M. A. (2024). Reassessing standardized tests: Evaluating their effectiveness in school performance measurement. *Curricula: Journal of Curriculum Development*, 3(2), 217-234. <https://doi.org/https://doi.org/10.17509/curricula.v3i2.74535>

Smith, W. C., & Holloway, J. (2020). School testing culture and teacher satisfaction. *Educational assessment, evaluation and accountability*, 32(4), 461-479. <https://doi.org/https://doi.org/10.1007/s11092-020-09342-8>

Stacey, M., Talbot, D., Buchanan, J., & Mayer, D. (2020). The development of an Australian teacher performance assessment: Lessons from the international literature. *Asia-Pacific Journal of Teacher Education*, 48(5), 508-519.

Taylor, C. S. (2022). Culturally and socially responsible assessment: Theory, research, and practice. Teachers College Press. <https://doi.org/doi.10.14507/er.v31.3737>

Tobón, S., Juárez-Hernández, L. G., Herrera-Meza, S. R., & Núñez, C. (2020). Assessing school principal leadership practices. Validity and reliability of a rubric. *Educación XX1*, 23(2), 187-210. <https://doi.org/DOI:10.5944/educXX1.23894>

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally responsive assessment: Provisional principles. *ETS Research Report Series*, 2023(1), 1-24. <https://doi.org/https://doi.org/10.1002/ets2.12374>