EDUCATIONAL ACCOUNTABILITY REMADE

Comparability Without the Harms of Standardization: How Construct Comparability Promotes Authentic and Rigorous Evidence of Learning

center for innovation in education
leading with learning

FairTest
The National Center for Fair & Open Testing

COMPARABILITY WITHOUT THE HARMS OF STANDARDIZATION

How Construct Comparability Promotes Authentic and Rigorous Evidence of Learning

THE EAR TENTPOLE PAPER SERIES

01 From Compulsion to Mutual Assent: Pitching a New Tent for Educational Accountability

02 Elevating Community Trust As the First Principle of Accountability

03 Proving Education Inadequacy Without the Test Score Trap: How Inputs and Meaningful Outcomes Make the Case

04 Comparability Without the Harms of Standardization: How Construct Comparability Promotes Authentic and Rigorous Evidence of Learning

05 Authentic Assessment and Accountability as an Engine of Educator Renewal

INTRODUCTION

At its core, educational accountability should safeguard civil rights and bolster public trust by ensuring that every student is engaged in learning based on high expectations, and that measures of student achievement carry a common and trustworthy meaning regardless of zip code. Currently, the primary policy mechanism for these safeguards is through assessment comparability: common ways of defining and determining whether students have reached "proficiency" of common learning expectations.

Since the passage of No Child Left Behind (NCLB) in 2002, state systems have pursued assessment comparability through what is known as *data comparability*. The logic is that the measurement tool itself – the test items, time limits, security conditions, and scoring mechanisms – should be curriculum-neutral and identical in every classroom so that the resulting scores are effectively interchangeable (Lyons, et al, 2016). In other words, our current system is premised on the idea that scores from the same tests taken in different places produce scores that mean the same in either setting.

Orienting school accountability around data comparability was a significant undertaking. Building the current system required a massive, multi-billion dollar effort supported and encouraged by a private sector testing industry to develop and roll out state standards, build and validate state assessments, reorient instructional calendars to make room for testing, and erect

complex systems to enforce high-stakes consequences. In 2024 and 2025 alone, twenty years after these systems were first created, assessment administration still cost up to $40M per state, based on an analysis of 15 states (Ismael et al., 2026).

The response from the field has been equally consequential: schools emphasize tested subjects and grade levels over those that are not tested, teachers spend more time on tested material and test prep, students spend hours on computer-based test prep programs, parents with financial means fund a tutoring industry promising higher test scores, and students experience a range of responses from test anxiety to disengagement.

Many are now questioning the return on investment of this quarter-century venture. Without compromising the core purpose of accountability – ensuring high expectations, equity, and transparency – it is time to consider alternatives to data comparability. .

Today, there is better technology and new methods to achieve comparability a different way–not through *data comparability* but *construct comparability.* In this approach, the knowledge or skill being assessed (the "construct") is the basis of comparability, not the test design and assessment delivery mechanisms themselves.

In this paper, we propose how construct comparability can fuel an accountability system that need not require complete uniformity but can incorporate authentic and locally-relevant learning experiences. Here and through the Educational Accountability Remade paper series, we will show how this more robust evidence of student learning feeds a virtuous cycle of capacity-building, school improvement and increased student engagement in the joy of learning that creates the opportunity for a new renaissance in public education.

THE HARMS OF A SYSTEM BUILT AROUND DATA COMPARABILITY

A Faulty Theory of Change

When standardization and data comparability became pillars of the federal accountability system in 2002, researchers such as Richard Elmore (2004) noted that it rested on no clear theory of change. Beyond an implicit assumption that publishing comparable data would inform policy decisions, Elmore found no clear notion of how the system was expected to actually affect the behavior of the people in schools.

Studies of accountability rooted in standardized testing had shown improvements in scores in the wake of new testing (Hanushek & Raymond, 2005; Jacob, 2007), but critiques soon arose of a certain circular logic. "Learning" was measured solely according to the standardized test scores that the federal policy itself encouraged – akin to buying someone a step counter to

make them pay attention to their steps, and concluding that an increase in their daily steps means they are healthier. While a good sign, the score provides a narrow picture of health (Bae, 2008).

By the early 2000s, policies attempted to compel improvement by attaching stakes to the standardized assessments, like school closures, governance takeovers, and other consequences for students and teachers who perform poorly. The term "high-stakes" tests contrasted these assessments with "low-stakes" tests, like the National Assessment of Education Progress (NAEP) and the Programme for International Student Assessment (PISA), which collected data through standardized instruments but were not assigned to students or severe consequences. Studies comparing performance between high- and low-stakes tests found muddied evidence of the efficacy of attaching high-stakes consequences; some studies found that high-stakes tests were indeed accompanied by benefits in learning math by younger students who were not significantly behind while others found divergent results between high-stakes and low-stakes tests, with scores improving on the former despite worsening on the latter (Jacob, 2007; Rothstein et al., 2008; Jennings & Lauen, 2016). An international comparison found that countries that had adopted systems based on data comparability experienced worsening results compared to similar nations that did not, a possibly perverse effect of public testing on practice (Adamson et al., 2016).

Studies have also found other particularly pernicious effects further undermining federal accountability's tenuous theory of change. Campbell's Law, stating that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor," has borne out in many ways. Beyond the most egregious examples of cheating (Atlanta, 2015; New York, 2011), much has been written about the narrowing of curriculum, teaching to the test, teaching to the "bubble kids," and disengagement of students from the classroom (see, for example, Payne-Tsoupros, 2010; Jennings & Lauen, 2016; Heilig, 2011; Rothstein, 2008).

On balance, despite some initial hope, the clear need to address broad and persistent inequities in the education system remain thirty years later. Systemic reporting of data-comparable assessments and high-stakes consequences for low performance have produced neither widespread improvements in achievement nor the changes in behavior needed to eliminate the glaring inequities of race, socioeconomics, language, and disability that the test scores so dramatically highlight.

A Mismatch with the Science of Learning

This data-comparability approach stands in stark contrast to what has been more recently learned in cognitive psychology and neuroscience about learning. These fields tell us that students, especially adolescent learners, learn better through emotionally engaged social process-

es and contextualized inquiry than through abstract, decontextualized, stand-alone tasks; yet, such tasks are key features of the data-comparability approach (Gottlieb, Yang & Immordino-Yang, 2024, Immordino-Yang et al.., 2024; Farrington et al., 2012). Moreover, the generic tasks of standardized tests are not as context-neutral as they seem, because students construct meaning through unique lived experiences, and thus can interact with any particular assessment task in idiosyncratic ways. Student motivation to perform on measurement tasks that feel detached from authentic and relevant learning can often result in a less than true assessment of learning based on half-hearted engagement (National Academies, 2018).

Another flaw of point-in-time standardized testing is that it systematically devalues another cognitive process essential to student growth and understanding: the iterative process of revision. Revision enables educators to convey that they value the learning process, enabling students to deepen learning and thinking. Standardized assessments, based on point-in-time administration of learning against tasks scored a single time, underscores the myth that learning is all or nothing.

Individualism at the Expense of Relationships and Social Capital

Relationships are central to both classroom learning and lifelong success. Research shows that strong peer-to-peer relationships and trust between students and adults are prerequisite conditions for improving learning in schools (Bryk and Schneider, 2002; Wortham & Jackson, 2012). Relationships also impact students' future economic success through the expansion of their social capital (Manno, 2021; Fullan et al., 2015). The experience of collaboratively developing understanding with others, especially across lines of difference, is no less than central to a functioning democratic society (Nieto, 2018; Levinson, 2011; Ambrosio, 2013).

Yet, by measuring – and thereby prioritizing – individual achievement on externally constructed tasks, standardized test-based accountability undermines rather than bolsters relationships. Literature on how accountability is experienced emphasizes that this one-way version of accountability—a score on a test motivates a change—is out of step with actual human experience rooted in multidirectional relationships—one's experience with multiple people and multiple assessments yields incremental changes in behavior. (Fahey & Koster, 2010; O'Kelly & Dubnick, 2010). Placing priority on assessments that are externally designed and scored bypasses teachers' more intimate knowledge of students' strengths and capabilities in favor of the limited information provided by the data-comparable tests (Farris-Berg & Dirkswager, 2013; Kostogriz & Doecke, 2011; Ryan et al., 2017; Ramirez et al., 2018). While this pretense of objectivity may seem like a boon for equity, its shallowness relative to the deep contextual knowledge held in the teacher-student relationship devalues teacher judgment, turns a child's asset-based hope into deficit-based shame, and leaves students, teachers, and parents to question the relevance of state tests to the outcomes that matter most to them.

Teachers, who had a minimal role in designing state level assessments, receive data that has been analyzed without them and many respond, unsurprisingly, with skepticism. Their more intimate knowledge of students' strengths and individual contexts are devalued in favor of what the data-comparable tests show (Farris-Berg & Dirkswager, 2013; Kostogriz & Doecke, 2011; Ryan et al., 2017; Ramirez et al., 2018). Faced with the pressure to raise individual test scores, instruction itself becomes increasingly individualized, deficit-focused and intervention-based rather than focused on meaningful learning and engagement with peers.

What is described above is not just bad luck or poor design. Rather, it reflects a contradiction at the heart of the data-comparability approach. To make headway on these issues requires addressing the central elements of that system.

THE BENEFITS OF A SYSTEM BUILT AROUND CONSTRUCT COMPARABILITY

Recent research by psychometrician Robert Mislevy suggests an alternative approach to comparing learning across systems (Mislevy, et al, 2025). Mislevy draws a distinction between *data comparability*, described above, and *construct comparability*, wherein performance on two assessments that are contextually different may be compared if the construct they measure (such as a standard or competency) is the same and performance is evaluated against common definitions of mastery for that construct. Thus, rather than insisting on standard, externally designed tasks, an accountability system based on construct comparability allows multiple, varied assessments of the same knowledge and skills.

Construct comparability aligns with design principles put forward by Bennett (2023) who has argued that learning and performance cannot be separated from the social, cultural, and linguistic settings in which students live and build knowledge. Mislevy cites an example in which students of the Blackfoot tribe of Alberta, Canada are studying thermodynamics through tipi construction. He makes the case that this particular evaluation of student learning of heat transfer in its sociocultural context is comparable to and more fair than other assessments of the same construct.

For this approach to support rigor and high expectations for all learners, it must include two key features drawn from Mislevy's work.

1. A clear, common definition of the underlying construct being assessed, which may include articulation of what performance looks like at different levels of sophistication ("proficient" or otherwise). This articulation is often provided through rubrics that describe what student work products or performance look like at levels that demonstrate proficiency, with contrasting descriptions to identify what it looks like when not yet at that level.
2. A validity argument for how different local tasks require the same level of intellectual work relative to the construct. This argument rests, in part, on processes through which the peo-

ple scoring the assessments build common understanding and calibrate their ratings with each other.

For example, consider the construct of evidence-based historical argumentation, including a student's ability to analyze primary and secondary sources, identify competing interpretations, and construct a written argument supported by evidence. A student in a school serving a predominantly immigrant community might demonstrate this construct through a research paper examining the labor organizing history of their own neighborhood. A student in a rural school might demonstrate it through an investigation of the economic history of their region's agricultural industry. The tasks look different, the sources are different, and the community contexts are different, but both students are being asked to do the same intellectual work. In both cases, their work is held to the same set of criteria for proficient demonstration of evidence-based historical argumentation.

This approach allows learners to demonstrate complex capabilities in locally meaningful projects. Such a system is more likely to boost engagement, deepen learning, strengthen relationships, and build social capital than the current system emphasizing individual performance on standardized tests.



The benefits extend to educators as well: by participating in defining constructs and creating performance rubrics, and by collaboratively scoring local tasks through moderation studies, educators deepen both their understanding of the standards and competencies they are teaching and their assessment literacy. This can be seen in the case study below and also has been found in projects such as New Hampshire's Performance Assessment of Competency Education (PACE) (Thacker & Becker, 2017). In fact, this work *is* the very work of good teaching. Unlike the

current system, where data comparability relies on standardized assessments that sit outside the daily work of the classroom, investments in a system built around construct comparability are direct investments in teaching and learning.

The case study that follows is one example of a robust performance-assessment system that illustrates the possibilities inherent in an approach to accountability built on construct comparability. This is not the only system of performance assessment currently operating in the United States. However, the New York Performance Standards Consortium's offers an important example from which to learn, given its scale (12,000+ students), enduring footprint (over 25 years), and well-developed approaches to establishing validity grounded in shared values around teaching and learning held across its network of schools

CASE STUDY: THE NEW YORK PERFORMANCE STANDARDS CONSORTIUM

The New York Performance Standards Consortium (the Consortium), a network that now includes 38 public schools across New York State, provides insights on what this system can look like in practice. The Consortium has built community while validating results against metrics of graduation, college enrollment, and college persistence. Through a set of curricularly-embedded, in-depth, performance-based assessments, Consortium schools assess students' readiness for high school graduation and post-secondary life, using an approach designed to ensure those assessments are constructed and implemented through valid and reliable methods.

Although the Consortium maintains smaller teacher to student ratios to accomplish this performance assessment design and evaluation work by classroom educators, member schools are funded using the formulas as other district schools; Consortium schools accomplish this economy of scale by prioritizing direct classroom instruction and advisory staffing roles over large departmental frameworks, heavy administrative overhead, and extensive non-instructional staffing.

The following paragraphs review this model in greater depth and illustrate how construct comparability can be used to support accountability systems that promote school cultures aligned to deeper learning.

Results and Effects on Equity

A comparison of Consortium graduates' success in college against that of their peers found that they achieved higher first-semester GPAs and persisted at higher rates in college (Fine & Pyiomak, 2020). This is evidence of "predictive validity" – the system is a valid predictor of the outcomes it claims to predict, a claim that is unmatched by the state's standardized assessment system. Further, graduation rates for English Language Learners and students with disabilities are significantly higher in Consortium schools than in the rest of New York City, evidence that the system is functioning equitably in practice.

Construct Comparability in the Consortium Model

To graduate from a Consortium school, students must complete four performance-based assessment tasks (PBATs) in writing and present these in an oral defense: an analytic literary essay, social studies analytical research paper, original science experiment or engineering design, and math problem-solving narrative. To complete these tasks, each student is expected to demonstrate skills of extended inquiry, evidence-based argumentation, oral defense and discussion, reflection and revision. Teacher and student agency and ownership

are built into task design, as is the learning arising from revision. Teachers design tasks that grow out of the curriculum of their courses. Students choose their own research questions and present work they have authored. Student work products are evaluated by the teacher as well as external evaluators, including outside experts and community members, using shared rubrics.

The engagement of students, teachers, and community members together in the Consortium assessment process is a fundamental departure from common practice in state assessment and yields deeper connection to community and to local knowledge, nurturing student-teacher relationships and student-community connections instead of disrupting or ignoring them. Yet, this departure is at odds with data comparability as it yields larger and more varied forms of assessments and student work products than standardized assessments do. The system instead demonstrates key components of a construct comparability approach to accountability.

Central to the process of ensuring the validity of the system across contexts are biannual moderation studies through which Consortium educators share information on scoring of assessments and samples of student work. Representatives from the 38 schools build inter-rater reliability through calibration exercises in which they review tasks for alignment to the shared rubrics and to Webb's Depth of Knowledge (2002) and review student work produced for those tasks. Through this process, inter-rater reliability is developed across schools while individual schools also regularly receive specific feedback on quality of their tasks and the alignment of their school-based scoring to the moderated score. School-based moderation studies complement this practice at the school level. In addition, rescoring of every written PBAT and oral defenses by colleagues and external evaluators provides ongoing accountability and feedback to teachers.



While such comparisons across multiple locally designed tasks were possible 28 years ago at the founding of the Consortium, the sophistication with which they can be performed is far greater today. This is due to the data storage, sharing, and comparison through online technology not previously available. As policy and educational leaders consider scaling similar systems based on construct comparability, these efficiencies will play a larger role.

Construct Comparability in the Context of the Science of Learning

The Consortium's model is aligned to the growing understanding from neuroscience and cognitive psychology of how adolescents learn, where students are assessed without undermining, and ideally while cultivating, students' sense of value and the opportunity for the exploration and deliberation of ethical and social issues (Immordino-Yang, 2016; Polakow-Suransky, Immordino-Yang, & Knecht, 2018).

This approach represents a fundamental departure from typical state-based assessment systems and one that would be untenable in these other systems. It treats learning as non-linear, socioculturally situated, and best evidenced through extended performance rather than recall under time pressure. Students graduate having demonstrated deep disciplinary knowledge and the intellectual habits necessary for success in college, careers and civic life, a necessity whose achievement is backed by the research on college success cited earlier. Student engagement in topic selection, deliberate and creative inquiry around a topic, and revision and reflection are all supported by the neuroscientific research.

Comparability in the Context of Relationships and Community

Relationships, social capital, and localized community and cultural knowledge are also deeply ingrained in the Consortium model. When students engage in an Oral Defense of each PBAT, it is to a panel of community members: the community engagement and social capital central to learning are thus central to the process of accountability. Common rubrics, collaboratively developed and used by teachers and external evaluators, ensure local judgment connects to a shared standard of quality. Teachers engage with students to select topics and shape questions they will explore in their PBATs, an element intended to build teachers' ownership of the assessment process and students' ownership of learning. The system is explicitly designed to build on students' funds of knowledge and lived experiences. Curriculum and tasks are designed by teachers in response to their students and in the context of the course. Students choose topics that connect to their own questions and communities.

A companion paper addressing "Community Trust" goes deeper into the ways in which families, business and community members can and have participated in the accountability processes associated with such a system. By advising on system design, observing student

defenses, providing student learning opportunities outside the classroom, and posing new authentic problems for student teams to address, they make the learning more meaningful and engaging for students and more relevant to caregiver and community needs.

A NEW BASIS FOR ACCOUNTABILITY REQUIRES NEW SYSTEM ELEMENTS

Not unlike the 25-year effort that reoriented American education around data comparability, a new model for accountability premised on construct comparability will require development. Yet unlike before, investments in this new system directly benefit teaching and learning by building educator capacity, not by financing external testing companies.

Shared Infrastructure

States and districts can begin by building shared infrastructure and piloting local accountability systems based on construct comparability. States and intermediary organizations can support this effort by:

- fostering environments for school communities to collaboratively develop assessments adapted to local contexts,
- establishing regional calibration hubs,
- curating open-source performance task banks, and
- fostering peer-review networks across participating districts.

Peer-review networks turn the routine work of scoring assessments into high-yield professional learning, building inter-rater scoring reliability while simultaneously sharpening daily classroom instruction. By treating system capacity as an incremental, developmental process rather than an all-or-nothing mandate, state leaders can expand participation organically as regional expertise deepens, proving at scale that technical rigor and local agency can reinforce one another.



Scaling such a system statewide does not mean replicating all assessments or asserting a comprehensive and mandatory set of construct-defining rubrics. Rather, it requires establishing shared purpose, then creating the necessary support and oversight to enable networks and communities to build assessment systems aligned to their own values and shared purposes. States and districts can enable this by setting clear standards for what high quality locally-developed construct-comparable systems entail, building networks, providing funding, and sharing expertise along with system oversight. As states consider how to bring their accountability systems in line with current research, care must be taken to consider how to carry through the key components, including the importance of community context, student and educator voice, educator learning, curriculum design, and classroom practice. Given the operational complexities of developing these new systems, we will be providing a practical roadmap for system development in future writings.

Thoughts on the role of generative AI

The New York Performance Standards Consortium has proven that a system of accountability based on construct comparability can be built without support of artificial intelligence. Consortium leaders do not currently use AI for any system-level infrastructure purposes,citing a strong preference for human-centered systems development and oversight.

At the same time, one can imagine that other local systems might experiment with using AI to support personalized learning environments, competency mapping, automated curriculum development, adaptive tutoring, formative assessment, and learner analytics (Ahmedtelba et al., 2026; Hong, 2025; Mukashova et al., 2025; Venter et al., 2024). For this paper, we take a neutral position on the use of Generative AI, with the added caution that any use of these powerful tools requires a set of principles based on careful research and planning to avoid bias and unintended consequences. One example of such unintended consequences is the necessity to change the

process and design of student assessments based on the growing risk of not knowing whether work is produced by students or by AI. A recent study introduced the idea of a "scalability-verification paradox" that the more AI is used in scaling systems, the more concern arises regarding AI-produced evidence of learning (Ahmedtelba, et al., 2026).

CONCLUSION: COMPARABILITY AS A CATALYST FOR SYSTEMIC RENEWAL

The current design of school accountability exposes a dilemma: comparable information is necessary for accountability, but the existing paradigm of data comparability calls for collecting that information in a way that often runs counter to how children learn. Our proposed remedy is to invest in a new paradigm of construct comparability, which values evidence of learning constructed within local communities. This approach allows for a richer, more community-connected form of education that can be scaled across diverse settings while maintaining rigorous accountability and revealing inequities in the system, provided appropriate safeguards are in place.

Several networks, including the New York Performance Assessment Consortium, The New Hampshire PACE districts, the Local Laboratories of Learning (L3) districts in Kentucky, and the Massachusetts Consortium for Educational Innovative Assessment (MCIEA), are already showing how moderation processes can bring together educators within schools and across districts to improve scoring reliability and address validity claims, while simultaneously serving to identify gaps and weaknesses in the system. These processes are not only necessary for ensuring equity but also directly benefit the system by increasing educators' assessment literacy and improving the assessments themselves.

Local educators hear consistently from business and industry partners who do not place much stock in current accountability measures. Instead, they are eager for indicators that provide a clearer picture of graduate competencies and accomplishments that translate seamlessly into their workplaces. The success of networks like the New York Performance Standards Consortium serves as both a proof-of-concept and a harbinger of what is possible. By embracing performance-based assessments that are locally-contextualized, lean into school-based relationships, and assess a broader set of skills and abilities,– and by constructing a system of comparability around them – we can move beyond the limitations of the past and build an educational system that truly respects both the science of learning and the complexity of the communities we serve.

REFERENCES

Adamson, F., Astrand, B., & Darling-Hammond, L. (2016). Global education reform: How privatization and public investment influence education outcomes. Routledge Press.

Ahmedtelba, E. , Mitereva, S. and Ponce, M. (2026). The Future of Competency-Based Education and Generative-AI: Solving the Scalability-Verification Paradox through Process-Based Assessment. Creative Education, 17, 1851-1872. doi: 10.4236/ce.2026.1710105.

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

Bae, S. (2018). Redesigning systems of school accountability: A multiple measures approach to accountability and support. *Education Policy Analysis Archives,* 26(8). https://doi.org/10.14507/epaa.26.2920

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. Educational Assessment, 28(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Elmore, R. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In N. Fuhrman & R. Elmore (Eds.) *Redesigning accountability systems for education*. Teachers College Press.

Ercikan, K. et al. Cross Cultural Validity and Comparability in Assessments of Complex Constructs. (2023)

Fahey, G. and F. Köster (2019), "Means, ends and meaning in accountability for strategic education governance", *OECD Education Working Papers*, No. 204, OECD Publishing, Paris, https://doi.org/10.1787/1d516b5c-en.

Farrington, C. A.; Roderick, M.; Allensworth, E.; Nagaoka, J.; Keyes, T. S.; Johnson, D. W.; Beechum, N. O., (2012). Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review. Consortium on Chicago School Research.

Farris-Berg, K., & Dirkswager, E. (2013). Trusting teachers with school success: What happens when teachers call the shots. Rowan and Littlefield Education. https://doi.org/10.5040/9798216416951

Fine, M., & Pryiomka, K. (2020). Assessing college readiness through authentic student work: How the City University of New York and the New York Performance Standards Consortium are collaborating toward equity. Learning Policy Institute, Palo Alto, CA.

Foster, N. and M. Piacentini (eds.) (2023), Innovating Assessments to Measure and Support Complex Skills, OECD Publishing, Paris, https://doi.org/10.1787/e5f3e341-en. Chapter 11.

Fullan, M., Rincon-Gallardo, S., & Hargreaves, A. (2015). Professional capital as accountability. *Education Policy Analysis Archives,* 23(15). https://doi.org/10.14507/epaa.v23.1998

Gill, B.P., Lerner, J.S., & Meosky, P. (2016). Reimagining accountability in K–12 education. *Behavioral Science & Policy 2*(1), 57-70. https://dx.doi.org/10.1353/bsp.2016.0007.

Gottlieb, R.J.M., Yang, X.F. & Immordino-Yang, M.H. (2024). Diverse adolescents' transcendent thinking predicts young adult psychosocial outcomes via brain network development. Sci Rep 14, 6254. https://doi.org/10.1038/s41598-024-56800-0

Hall, A. T., Frank, D. D., & Buckley, M. R. (2017). An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *Journal of Organizational Behavior*, 38, 204–224.

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

Heilig, J. V. (2011). As good as advertised? Tracking urban student progress through high school in an environment of accountability, *American Secondary Education*, 39(3), 17-41.

Hong, L. (2025). Development and Validation of a Competency-Based Ladder Pathway for AI Literacy Enhancement among Higher Vocational Students. *Scientific Reports, 15,* Article No. 29898. [Google Scholar] [CrossRef] [PubMed]

Immordino-Yang, M. H. (2016). *Emotions, learning, and the brain: Exploring the educational implications of affective neuroscience*. W. W. Norton & Company.

Immordino-Yang, M.H. (2025). The power of the adolescent brain and mind Kappan

Immordino-Yang, M. H., Kundrak, C., Knecht, D., & Matthew, J. (2025/12/01). Civic reasoning depends on transcendent thinking: Implications of adolescent brain development for SEL. *Social and Emotional Learning: Research, Practice, and Policy.* 10.1016/j.sel.2024.100067

Ismael, K., Jasiura, K., Lee, Y., Vranek, J. (2026). Rethinking the Test Pile: A National Study of K-8 Academic Assessments. Education First.

Jacob, B.A. (2007). Test-based accountability and student achievement: An investigation of different performance on NAEP and state assessments. *National Bureau of Economic Research, Working Paper 12817.*

Jennings, J. L. & Lauen, D. L. (2016). Accountability, inequality, and achievement: The effects of the No Child Left Behind Act on multiple measures of student learning. *RSF: Russell Sage Foundation Journal of the Social Sciences,* 2(5), 220-241. https://doi.org/10.7758/rsf.2016.2.5.11

Jimenez, L., Boser, U. The Way Forward for State Standardized Tests. Center for American Progress. (2021)

Joyce, T.J., Khosroshahi, M.A., Truelsch, S. ,Gentsch, K. & Du, K. Standardized Test Scores and Academic Performance at a Public University System. National Bureau of Economic Research (March 2026).

Kostogriz, A., & Doecke, B. (2011). Standards-based accountability: Reification, responsibility and the ethical subject. *Teaching Education*, 22(44), 397-412. https://doi.org/10.1080/10476210.2011.587870

Levinson, M. 2011. Democracy, accountability, and education. Theory and Research in Education 9, no. 2: 125-144. https://doi.org/10.1177/1477878511409622

Lyons, S., Christopherson, S., Brown, J.C., Ihlenfeldt, S., & Pace, L. (2025). Evaluating Technical Quality in the Context of Assessment Innovation: Policy Implications from a Case Study. Lyons Assessment Consulting, Wisconsin Center for Education Products and Services, KnowledgeWorks.

Lyons, S., Marion, S.F., Pace, L., & Williams, M. (2016). Addressing Accountability Issues including Comparability in the Design and Implementation of an Innovative Assessment and Accountability System. www.knowledgeworks.org and www.nciea.org.

Lyons, S., Oliveri, M. E., & Poe, M. (2025). A framework for enacting equity aims in assessment use. In C. M. Evans & C. S. Taylor (Eds.), *Culturally responsive assessment in classrooms and large-scale contexts.* Routledge.

Lyons, S. (2026) The Illusion of Comparability: Rethinking What Test Scores Can Mean. (Unpublished Paper)

MacNeil, K. A., Butler, D. L., & Schneller, L. M. (2021). From accountability to shared responsibility: A case study of a multi-layered education change initiative. *Journal of Educational Change.* https://doi.org/10.1007/s10833-021-09440-4

Manno, B. V. (2021). New Pathways for Education Reform. *Stanford Social Innovation Review*, *19*(3), 40–47. https://doi.org/10.48558/VVJ3-FE39

Meier, D., & Knoester, M. (2017). Beyond Testing: Seven Assessments of Students and Schools Better than Standardized Testing. Teachers College Press.

Mislevy R. J,, Oliveri M. E., Slomp,, D., Crop Eared Wolf, A., Elliot N. (2025). An Evidentiary-Reasoning Lens for Socioculturally Responsive Assessment. Taylor and Francis Group.

Mukashova, A., Tussupov, J., Sandugash, S., & Mukhavnova, A. (2025). AI-driven framework for automated competency formalization: From professional standards to adaptive learning outcomes. *Frontiers in Computer Science* (7). https://doi.org/10.3389/fcomp.2025.1710358

National Academies of Sciences, Engineering, and Medicine. 2018. How People Learn II: Learners, Contexts, and Cultures. Washington, DC: The National Academies Press. https://www.nationalacademies.org/read/24783/chapter/1

Newton, X. A., Lohmeier, J. H., Thompson, S. R., Oh, B., Dmelio, J. R., & Trainor, A. (2022). Rethinking alternative education accountability indicators: A multilevel contextual perspective. *Performance Improvement Quarterly,* 34(4), 455-489. https://doi.org/10.1002/piq.21358

Nieto, S. (2018). Language, culture, and teaching: Critical perspectives. Routledge Press.

NH PACE Technical Manual. 2016, 2017. Susan Lyons, Carla Evans, Scott Marion, and Jeri Thompson National Center for the Improvement of Educational Assessment (July 28, 2017)

NCSA 2023 Plenary Lunch Session, (Susan Lyons) Tuesday, June 27, 2023

O'Kelly, C., & Dubnick, M. J. (2019). Dissecting the semantics of accountability and its misuse. In H.L. Paanakker, A. Masters, & L. Huberts (Eds.), *Quality of governance: Values and violations* (pp. 45-79). Palgrave Macmillan.

Payen-Tsoupros, C. (2010). No child left behind: Disincentives to focus instruction on students above the passing threshold. *Journal of Law and Education,* 39(4), 471-501.

Polakow-Suransky, Immordino-Yang, M.H., Knecht, D. EdWeek. (2018) Why Science Says School Is Boring.

Queensland Curriculum & Assessment Authority, QCAA. (2014). School-based Assessment: The Queensland System.

Ramirez, G., Hooper, S., Kersting, N., Feguson, R., & Yeager, D. (2018). Teacher math anxiety relates to adolescent students' math achievement. *AERA Open,* 4(1). https://doi.org/10.1177/2332858418756052

Rothstein, R. Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right.* Teachers College Press.

Rothman, M. (2024). The experience of accountability for students who reclaim success: The case study of an alternative high school. [Doctoral Dissertation, Fielding Graduate University.] ProQuest Dissertations & Theses.

Ryan, S., von der Embse, N., Pendergast, L., Saeki, E., Segool, N., Schwing, S. (2017). Leaving the teaching profession: The role of teacher stress and educational accountability policies on turnover intent. *Teaching and Teacher Education,* 66, 1-11. https://doi.org/10.1016/j.tate.2017.03.016

Shepard, L. AFT. (2021) Ambitious Teaching and Equitable Assessment: A Vision for Prioritizing Learning, Not Testing. https://www.aft.org/ae/fall2021/shepard

Thacker, A. & Becker, D. (2017). Formative Evaluation of New Hampshire's Performance Assessment of Competency Education (PACE) Summary Report. HumRRO. (Center for Innovation in Education).

Troppe, P., Osowski, M., Wolfson, M., Ristow, L., Lomax, E., Thacker, A., & Schultz, S. (2023). Evaluating the Federal Innovative Assessment Demonstration Authority: Early Implementation and Progress of State Efforts to Develop New Statewide Academic Assessments (NCEE 2023-004). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee This report also is available on the IES website at http://ies.ed.gov/ncee.

Velotti, L., & Justice, J. B. (2016). Operationalizing Giddens's recursive model of accountability. *Public Performance and Management Review,* 40(2), 310-335. https://doi.org/10.1080/15309576.2016.1207549

Venter, J. W., Coetzee, S. A., & Schmulian, A. (2024). Exploring the Use of Artificial Intelligence (AI) in the Delivery of Effective Feedback. *Assessment & Evaluation in Higher Education, 50,* 516-536. [Google Scholar] [CrossRef]

Webb, N. Depth-of-Knowledge Levels for Four Content Areas. (2002).

Wortham, S., & Jackson, K. (2012). Relational education: Applying Gergen's work to educational research and practice. *Psychological Studies,* 57(2), 164-171. https://doi.org/10.1007/s12646-011-0120-z

This paper was authored by Paul Leather, Partner at the Center for Innovation in Education, Alexandra (AJ) Rathmann-Noonan, Executive Director of NYPSC, and Michael Rothman, Senior Director, P20 Pathways Development at City University of New York, with editing by Jenny Poon and contributions from the Project EAR team.

SPONSORING ORGANIZATIONS

The Center for Innovation in Education supports education leaders to create effective, durable, community-owned solutions to complex problems.

The National Center for Fair & Open Testing (FairTest) works to end the misuses and flaws of standardized testing and to ensure that evaluation of students, teachers and schools is fair, open, valid and educationally beneficial.

SUGGESTED CITATION

Leather, P., Rathmann-Noonan, A., & Rothman, M. (November 2026). *Comparability without the harms of standardization.* Education Accountability Remade.

PROJECT EAR TEAM

MICA BAUM-TUCCILLO
Director of Research, New York Performance Standards Consortium; Senior Research Fellow, Public Science Project, City University of New York

CHRISTINE DONFRANCESCO
Senior Policy Analyst, Assessment and Accountability, National Education Association

HARRY FEDER
Executive Director, National Center for Fair & Open Testing (FairTest)

MICHELLE FINE
Distinguished Professor, Psychology, Urban Education, Liberal Studies, Women's and Gender Studies, American Studies, Public Scholarship, LGBTQ Studies, City University of New York

PAUL LEATHER
Partner, Education Policy and State and National Relationships, Center for Innovation in Education

TONY MONFILETTO
Fmr Director, Instituto del Puente at Future Focused Education

PAM MORAN
Executive Director, Virginia School Consortium for Learning

KAREN PERRY
Executive Director, University of Kentucky Center for Next Generation Leadership

JENNIFER POON
Partner, Center for Innovation in Education

ALEXANDRA (AJ) RATHMANN-NOONAN
Executive Director, NYPSC

MICHAEL ROTHMAN
Senior Director, P20 Pathways Development, City University of New York

DOANNIE TRAN
Managing Partner, Center for Innovation in Education