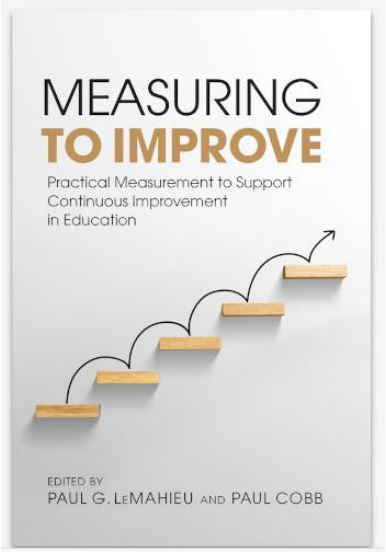
Education Review
Reseñas Educativas Resenhas Educativas

January 7, 2026 ISSN 1094-5296

LeMahieu, P. G., & Cobb, P. (Eds.) (2025). *Measuring to improve: Practical measurement to support continuous improvement in education*. Harvard Education Press.

280 pp. **ISBN: 9781682539675**

Reviewed by Audrey Amrein-Beardsley
Arizona State University
United States



Measuring to Improve edited by Paul G. LeMahieu of the Carnegie Foundation for the Advancement of Teaching and Paul Cobb, Professor Emeritus at Vanderbilt University, provides a thorough, interconnected, and accessible framework for how those working within, and leading, educational organizations can use measurements and measurement-based systems in practical ways. Following their lead, herein, I define practical measurement as measurements and measurement-based systems designed for real-world, real-time, day-to-day use, to support improvements in student learning over time, as assessed, evaluated, and understood through multiple measures of student achievement.

Instead of treating educational measurement as an exercise in compliance or an enterprise primarily meant for accountability purposes, the book editors and chapter authors (many of whom are also affiliated with the Carnegie Foundation for the Advancement) position educational measurement tools as learning apparatuses that support educators and educational leaders in both diagnosing problems and monitoring progress towards resolving said problems. This is to be achieved by designing educational measurements for use within and across schools and districts, as embedded in realistic organizational and instructional contexts, rather than being abstracted from everyday practice. Indeed, educational improvement (and reform) requires different kinds of data than those generated by traditional large-scale assessments (such as the standardized tests required under No Child Left Behind (NCLB, U.S. Department of Education, 2002). Rather, educators and educational leaders must lean on data that are timely, specific, descriptive, and much more closely connected to the instructional and organizational practices educators and educational leaders are working to change.

Amrein-Beardsley, A. (2026, January 7). Review of *Measuring to improve: Practical measurement to support continuous improvement in education*, by P. G. LeMahieu & P. Cobb. *Education Review, 33*.
https://doi.org/10.14507/er.v33.4401

In pursuit of these objectives, the late W. James Popham (2008), whose lasting influence on modern perspectives regarding formative assessment and measurement systems endures, warned that the utility of an assessment diminishes as its distance from the classroom environment increases. This assertion highlights that the instructional efficacy of any assessment or assessment-based system is contingent not solely on its technical complexity, but also on its integration with classroom practice and its ability to guide prompt pedagogical choices.

Consequently, this book furnishes a robust conceptual framework alongside actionable advice for educators and educational administrators endeavoring to integrate measurement into their continuous improvement efforts. The editors and contributing authors contend that educational systems achieve optimal enhancement when those directly involved in the work possess dependable, practical information, prioritizing inquiry, learning, and collaborative problem-solving over the utilization of summative data for external evaluation, accountability, or critical decision-making. These principles collectively define the book's primary contribution and the standards by which I evaluate its efficacy.

Framing

Informed by this perspective, first, I examine the book's four principal strengths, considering the contributions of various authors and chapters. I emphasize the book's coherent conceptual framework, its foundation in measurement practices that prioritize improvement and formative assessment, and its potential value for individuals involved in educational enhancement and reform initiatives. Subsequently, I address four areas of the book that I found less persuasive or inadequately developed, offering, with due respect and from my professional viewpoint, suggestions for enhancing the book's overall impact through greater clarity, balance, or more critical analysis.

It is also essential to acknowledge that this review will be structured around the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014, henceforth referred to as the *Standards*). In my own research, I consistently use the *Standards*, especially when evaluating different types of educational assessment and assessment systems, regardless of their purpose or context. I use this same approach here. Additionally, I connect some of my observations to Michael T. Kane's (2013) interpretive/use argument (IUA) model of validity. Kane's model views validity as an argument based on evidence and theory, linking the suggested meanings and uses of scores to their intended and unintended results.

From this perspective, my initial criticism of this publication centers on the editors' and authors' inconsistent (or lack of) engagement with the *Standards*. A more direct and explicit incorporation of the *Standards* throughout this book would have been beneficial, especially considering the status of the *Standards* as the most nationally and internationally recognized and used framework for defining, assessing, and protecting the intended and unintended applications and outcomes of the practical strategies proposed by the book's editors and authors. I should add, though, that this observation is accompanied by a crucial technical qualification: the *Standards*, in their current form, are not particularly accessible

documents, especially for practitioners or those focused on improvement; their language and structure can present significant accessibility hurdles. However, the book's relevance to the broader field of measurement could have been improved by more direct involvement from book editors and authors, even via translations focused on increased accessibility. This would have made the book more useful, accessible, and *Standards*-aligned for readers looking for both theoretical and practical guidance.

Notwithstanding this critique, I acknowledge the book's considerable merit while also positioning it within the enduring discourse surrounding educational measurement. This discourse primarily concerns the design, interpretation, and application of measurement tools to more effectively foster student learning and facilitate organizational advancement. My objective is not to criticize the book for its omissions, but to engage with it as a significant and pertinent contribution that encourages ongoing discussion. Consequently, I commence by identifying what I consider the book's most notable strengths.

Strengths

One of the book's most important strengths is its sustained and explicit rejection of the idea that measurement designed for improvement can, or should, serve high-stakes accountability purposes. Across chapters, the authors repeatedly emphasize that measures attached to external judgment, sanctions, or even some types of evaluation will inevitably distort practice, narrow instruction, and undermine validity. I define validity herein, following the *Standards*, as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests," including careful attention to the consequences of those uses (AERA et al., 2014, p. 11). This stance is articulated most clearly in Chapter 1 (LeMahieu et al.), it is empirically reinforced in Chapters 3 (Jackson et al.) and 5 (Allensworth), and it is reiterated in Chapter 8 (Cobb & LeMahieu).

I find this position particularly compelling, likely because it aligns most closely with my own work documenting the inordinate number of harms that too often come from test-based accountability, especially when systems become overly reliant on summative uses (misuses and abuses) of the accountability indicators derived from tests (Amrein & Berliner, 2002; Amrein-Beardsley, 2014). This stance also reflects the *Standards'* explicit insistence that all validity claims must attend to the consequences of score interpretation and use (AERA et al., 2014), as well as Kane's interpretive/use argument framework (IUA), which conceptualizes validity as an evidentiary and theoretical argument linking proposed score interpretations and uses to their intended and unintended consequences (Kane, 2013). Taken together, the editors' and chapter authors' insistence that practical measures be insulated from high-stakes uses reflects what is, in my opinion, a highly principled understanding of validity as fundamentally tied to purpose, context, and, most critically, use.

A second major strength is the book editors' and authors' consistent grounding of educational measurement in the realities of instructional and organizational work. Rather than treating assessment as a technical overlay, they show how practical measures are embedded (and should be more explicitly recognized and

embraced as embedded) in routines, relationships, and professional judgment. Indeed, measurements and measurement-based systems are most harmful when they marginalize or displace human judgment, rather than support it, particularly in contexts where professional expertise is essential for responsibly interpreting, in this case, test-based evidence (Amrein-Beardsley, 2014). This theme cuts across Chapter 2's Carpe College Access Network (Takahashi & Norman), Chapter 6's exit-ticket system for ninth-grade literacy (Larbi-Cherif et al.), and Chapter 4's Using Sources Tool within the National Writing Project (Friedrich & Bear).

What I also find particularly convincing here is how the authors of these chapters demonstrate that educational measurement becomes meaningful, but only when it is accessible, interpretable, and actionable by those closest to real educational work, in real time and practice. Their collective emphasis on timeliness, minimal burden on teachers and students, and shared sense-making resonates strongly with formative assessment principles, again, as closely aligned with Popham (2008), as well as the *Standards*, which frame validity (or validity of inference) as a function of how evidence is interpreted and used for inference-making purposes (see also Kane, 2013). In this respect, the book stands in stark contrast to the often unclear and slow data streams typical of large-scale testing, which often prioritize distal reporting over understanding for improvement. This difference highlights the book's main contribution: shifting assessment toward learning, rather than just following rules. It also challenges established patterns that have limited the field's ability to use assessment to support learning, rather than just for accountability, often by itself.

In this context, the book's exploration of validity is, in my view, unusually thorough, especially for a work focused on using educational assessments to improve teaching. Chapter 7 (Smith) makes a particularly important contribution by distinguishing between validity-for-use and validity-in-use. This distinction closely reflects both the *Standards* (AERA et al., 2014) and Kane's (2013) IUA-oriented focus on validity arguments based on interpretation and use.

This approach to validity is not just a theoretical idea, though. It is also shown in earlier chapters, especially Chapter 3, which discusses the Practical Measures for Reasoning about Rigor (PMRR) project (Jackson et al.). In this chapter, the authors demonstrate how even well-designed measures can become invalid if they're used incorrectly or interpreted in a way that focuses on weaknesses. As a result, after reading this chapter, I found myself reconsidering, and perhaps even partially changing, my earlier criticism that the book editors and authors did not engage the *Standards* enough.

The logic underpinning modern validity theory was, as demonstrated in this chapter and throughout the book, consistently present, even when not explicitly identified. Consequently, this perspective represents a significant contribution of this book. The editors and numerous chapter authors acknowledge that even well-constructed assessments can be detrimental if they are not aligned with suitable objectives, contexts, and user perspectives. This understanding, while a long-standing principle within contemporary validity theory, is not always emphasized in policy-oriented measurement reforms, which often (though not invariably) treat scores as self-explanatory or automatically actionable (Kane, 2013). Accordingly,

this book's consistent, albeit occasionally implicit, consideration of contemporary validity theory represents, in this author's view, its third most significant asset, especially for those interested in the principled application of measurement in real-world scenarios.

Finally, it is crucial to acknowledge that equity is not merely a rhetorical show within this text; instead, it functions as a fundamental design principle that directly influences the selection of what is measured, the methodologies employed, and the interpretation and application of the findings. As a case in point, the authors of Chapters 2 (Takahashi & Norman), 3 (Jackson et al.), 4 (Friedrich & Bear), and 6 (Larbi-Cherif et al.) prioritize historically marginalized students by focusing on their instructional opportunities, lived experiences, and access to meaningful learning, rather than relying solely on distant outcome measures. Chapter authors illustrate how educational interventions and the measurement-driven policies and systems that can support them can be used to shift focus away from perceived, socially constructed student deficiencies. Rather, they emphasize system-level conditions, instructional practices, and organizational routines that adults can change through their power, authority, and agency. They, therefore, effectively promote an equity-focused view of measurement, emphasizing responsibility, improvement, and collective learning, rather than surveillance, categorization, or blame.

Additionally, I found the authors' decisions in Chapters 3 and 6, which explicitly prioritize students' perceptions and experiences as valid evidence, to be compelling. This perspective further critiques prevailing accountability models that prioritize results over actual experiences within classrooms. Conversely, student-centered assessment positions students as primary and reliable sources of information, especially concerning the circumstances they perceive as conducive to their learning. This approach is also consistent with the *Standards*, as the authors highlight the necessity for validity evidence evaluations to consider fairness, subgroup disparities, and the intended and unintended impacts of score application across different groups (AERA et al., 2014). Incorporating student perspectives into measurement systems broadens the definition of evidence, strengthening the validity of the inferences drawn by ensuring that interpretations are based on how different student groups experience schooling.

I certainly commend the editors and chapter authors for their work, and I acknowledge their accomplishments beyond the scope of this review. I now turn to my four main critiques.

Critique

First, despite the editors' and authors' repeated cautions about accountability misuse, they tend to underplay how extremely difficult it is, in practice, to protect improvement-oriented measures, particularly formative assessments, also after they begin to demonstrate their value. The authors of Chapters 1 (LeMahieu et al.) and 8 (Cobb & LeMahieu) clearly call out the persistent tensions between improvement and accountability; however, they stop short of fully grappling with the structural realities of educational systems, in which district leaders, states, and external funders routinely seek (or are forced) to repurpose formative and summative

assessment data for monitoring, evaluation, comparison, and high-stakes decision-making. Notably, the *Standards* caution against score use, like in these instances noted, that exceeds or departs from the evidentiary basis originally established for interpretation (see also Kane, 2013).

Based on my own work examining value-added models (VAMs) and other primarily federally induced test-based accountability systems and policies, I remain skeptical that principled assessment design alone is sufficient to prevent such repurposing once political and policy incentives are in play (Amrein-Beardsley, 2008; Amrein-Beardsley & Collins, 2012; Amrein-Beardsley & Close, 2019; Amrein-Beardsley & Geiger, 2020). More explicit attention to governance structures, policy safeguards, and constraints such as collective bargaining agreements would have strengthened the book's guidance, accordingly, especially for sustaining improvement-oriented measurement in high-stakes environments. Absent such protections, even well-designed assessment measures are routinely absorbed into accountability regimes that privilege comparability, surveillance, and sanctioning over learning, professional judgment, and instructional improvement. This is a pattern that, from a validity perspective, reflects failures of use- and consequence-related warrants, as articulated in both the *Standards* and Kane's IUA, especially when those most empowered to repurpose assessment data are least constrained to do so.

Second, a cross-cutting concern of mine involves the book's implicit assumptions about capacity, or assumptions about the extent to which school and district leaders possess not only the human resources with technical expertise to generate and analyze data, but also possess the organizational stability, leadership continuity, and institutional supports necessary to sustain improvement-oriented measurements and measurement-based systems over time. The *Standards* emphasize the importance of these assumptions because the validity of score interpretations and applications hinges on the specific contexts in which these measurements and measurement-based systems are employed, considering available resources and support systems. Furthermore, in accordance with Kane's IUA framework, capacity and contextual conditions are critical, albeit frequently implicit, assumptions that influence assertions regarding score utilization, resulting outcomes, and decision-making processes.

Authors of Chapters 2 (Takahashi & Norman), 3 (Jackson et al.), 5 (Allensworth), and 6 (Larbi-Cherif et al.) all examine contexts distinguished by significant analytical proficiency, comparatively stable leadership, and strong research-to-practice collaborations. Although these case studies are undoubtedly valuable and persuasive, they, without sufficient qualifications, also risk portraying idealized representations of the practical requirements for effective measurement implementation.

In many schools and districts, especially those serving historically marginalized groups, educational leaders often lack the necessary staff, time, data systems, and analytical support seen in the examples advanced. Without a clear discussion of the required adjustments, oversimplifications, as well as resource limitations and trade-offs, practical measurement could become a reform that mainly benefits systems with existing resources. This would reduce its usefulness.

Third, while the editors and chapter authors consider validity, reliability is often addressed indirectly rather than directly. Authors of Chapters 3 (Jackson et al.), 4 (Friedrich & Bear), and 6 (Larbi-Cherif et al.) discuss rubrics, surveys, and exit tickets, all of which are appropriately designed to be contextually and interpretively sensitive. Although such sensitivity is frequently advantageous for formative assessment and improvement initiatives, the authors could have provided more guidance to assist readers in systematically evaluating acceptable levels of consistency, stability, and measurement error across various applications and interpretations. The *Standards* unequivocally state that reliability is intrinsically linked to validity claims; consequently, reliability must be assessed, evaluated, and reported, especially given the increasing prevalence of surveys and other subjective measures, such as rubrics, within and across state- and district-level systems that rely upon multiple measures. Furthermore, reliability must be assessed in relation to the assessments' intended interpretations and applications (see also Kane, 2013).

Indeed, more specific guidance on how to assess reliability in the low-stakes, practice-based measurement situations discussed in this book would likely better help practitioners balance interpretive flexibility with appropriate caution. This would also reduce the risk of overestimating the accuracy of measures that, while useful for learning, might be especially unreliable when used outside their intended (and validated) contexts.

Finally, while equity is a central theme in this book, some chapter authors, particularly Chapters 4 (Friedrich & Bear) and 6 (Larbi-Cherif et al.), make claims about equity impacts that are more suggested than proven. More specific evidence would have been helpful, showing how the practices described changed teaching methods, organizational structures, or learning outcomes for specific student groups over time.

Evidence could have been drawn from various sources, including, but not limited to, documentation of improved access to advanced courses for historically marginalized students, a decrease in the use of exclusionary disciplinary measures for comparable students, the redistribution of instructional resources or support services to benefit these students, enhanced instructional coherence for multilingual learners or students with disabilities, or differential improvements in learning outcomes, also specifically for historically marginalized student populations. Absent such evidence, there exists the possibility that equity remains an aspirational objective rather than a verifiable outcome of the practical measurement systems that have been both conceptualized and promoted in this book.

From a measurement standpoint, this apprehension is consistent with the *Standards*, which prioritize fairness, consideration of the consequences of test application, and the necessity for validity evidence that directly addresses subgroup effects. Likewise, this concern aligns with Kane's IUA, which mandates explicit justifications for assertions regarding score applications and their effects, encompassing equity-related consequences. However, this critique does not aim to undermine the book's focus on equity. Instead, it emphasizes the necessity of ongoing scrutiny regarding the practical realization of equity objectives within improvement-focused measurements and measurement-driven systems.

These criticisms, when considered collectively, are not intended to negate the book's central objectives; rather, they serve to expand on them. My criticisms acknowledge the intricate nature of the editors' and chapter authors' undertaking, alongside the structural, political, and organizational limitations inherent in the practical application of improvement-focused measurements and measurement-based systems. Crucially, none of these reservations diminishes the book's fundamental contribution: a principled, practice-oriented perspective on educational measurement and measurement-based systems that emphasizes learning, equity, and professional judgment, rather than mere adherence to regulations and control. The editors and chapter authors, instead, offer suggestions for future research, design, and policy work that could build on and strengthen the foundation laid throughout this book.

In fact, *Measuring to Improve* provides a clear and ethically sound approach to rethinking educational measurement. Its main strength is its consistent refusal to separate technical quality from its purpose, how it is understood, how it is used, and its effects, a position that directly supports the *Standards*.

Validity, at its core, concerns the suitability of score interpretations and applications within defined settings. This perspective aligns with Kane's IUA framework, which emphasizes the necessity of explicit, evidence-supported justifications that connect measurement assertions to their intended applications and outcomes. Consequently, this book offers a constructive counterpoint to historically established and consistently dominant accountability frameworks, which have often prioritized technical improvements and compliance objectives, particularly among federal and state educational policymakers, while simultaneously disregarding contextual factors, professional expertise, and the downstream impacts (i.e., both positive and negative, intended and unintended consequences), that arise from the interpretation and subsequent utilization, misuse, or abuse of assessment data.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/open-access-files.html

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). https://doi.org/10.14507/epaa.v10n18.2002

Amrein-Beardsley, A. (2008). *Methodological concerns about the Education Value-Added Assessment System (EVAAS). Educational Researcher, 37*(2), 65–75. https://doi.org/10.3102/0013189X0831642

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Routledge.

Amrein-Beardsley, A., & Close, K. (2019). Teacher-level value-added models (VAMs) on trial: Empirical and pragmatic issues of concern across five court cases. *Educational Policy*. https://doi.org/10.1177/0895904819843593

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (EVAAS): A review of the intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). https://epaa.asu.edu/index.php/epaa/article/view/1096

Amrein-Beardsley, A., & Geiger, T. (2020). Methodological concerns about the Education Value-Added Assessment System (EVAAS): Validity, reliability, and bias. *SAGE Open, 10*(2). https://doi.org/10.1177/2158244020922224

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

Popham, W. J. (2008). *Transformative assessment*. ASCD.

U.S. Department of Education. (2002). *No Child Left Behind Act of 2001* (Pub. L. No. 107–110, 115 Stat. 1425). https://www.govinfo.gov/content/pkg/PLAW-107publ110/pdf/PLAW-107publ110.pdf

About the Reviewer

Audrey Amrein-Beardsley is a professor in the Educational Policy and Evaluation Division of the Mary Lou Fulton College for Teaching and Learning Innovation. Her research interests include educational policies, measurement and measurement-based systems, research methods, and educational policies and systems based upon high-stakes tests and value-added models (VAMs). She is author of over 100 peer-reviewed and editorially reviewed journal articles and two academic books on these topics. She has been recognized as being one of the "Top Edu-Scholars in the Nation," honored by *Education Week* for being a university-based academic contributing to the public discourse and current thought surrounding America's public schools. Her research has also been highlighted in popular press outlets, including National Public Radio (NPR), *The New York Times*, *USA Today*, *The Washington Post*, *Education Week*, and HBO's *Last Week Tonight with John Oliver*. She has also served as an expert witness on behalf of many educators across states and districts, as per some of the lawsuits about the consequential (mis)uses of test scores.

About the Editors

Paul G. LeMahieu is senior advisor to the president at the Carnegie Foundation and graduate faculty in the College of Education, University of Hawai'i – Mānoa. LeMahieu served as superintendent of education for the State of Hawai'i, serving 190,000 students; prior to that he was undersecretary for educational policy and research

for the State of Delaware. He has been president of the National Association of Test Directors and vice president of the American Educational Research Association. Paul's current professional interests focus on the adaptation of improvement science methodologies for application in networks in education. Paul has a doctorate (Ph.D.) from the University of Pittsburgh, a master's (M.Ed.) from Harvard University, and a bachelor's (A.B.) from Yale College.

Paul Cobb is professor emeritus at Vanderbilt University. His work focuses on improving the quality of mathematics teaching and student learning on a large scale. He is currently involved in a project that is developing practical measure of key aspects of high-quality mathematics and investigating their use as levers for as well as measures of instructional improvement. He received Hans Freudenthal Medal for cumulative research program over the prior ten years from the International Commission on Mathematics Instruction (ICMI) in 2005, and the Silver Scribner Award from American Educational Research Association in 2010 for research over the past 10 years that contributes to our understanding of learning and instruction.

Education Review

Reseñas Educativas **Resenhas Educativas**