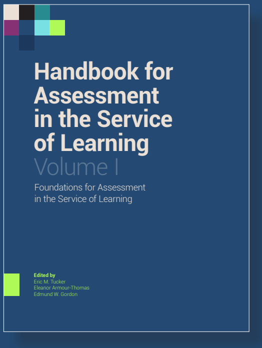
Practical Measurement for Improvement: Foundations, Design, Rigor

Paul G. LeMahieu and Paul Cobb

Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker



Handbook for Assessment in the Service of Learning
Volume I
Foundations for Assessment in the Service of Learning

Edited by
Eric M. Tucker
Eleanor Armour-Thomas
Edmund W. Gordon

ISBN: 978-1-945764-33-2

Suggested Citation:
LeMahieu, P., & Cobb, P. (2025). Practical measurement for improvement: Foundations, design, rigor. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Practical Measurement for Improvement: Foundations, Design, Rigor

Paul G. LeMahieu and Paul Cobb

Educational systems today face persistent challenges that demand not only innovation but disciplined learning about what works, for whom, and under what conditions. Improvement science has emerged in response to this challenge, offering a structured, iterative, and evidence-based approach to addressing complex problems of practice. At the heart of this approach lies "practical measurement," a form of assessment that is embedded within the flow of professional practice and is designed to support real-time learning and continuous improvement.

This essay foregrounds three critical aspects of practical measurement in education: (1) the theoretical foundations of improvement science and its implications for measurement; (2) the design and implementation of practical measures; and (3) the technical quality and validity concerns that must be addressed to ensure responsible and equitable use. In contrast to discussions of abstract psychometrics and summative evaluations, this synthesis emphasizes how measurement can function as a tool that frontline educators, school leaders, and improvement teams can use to drive meaningful change.

By examining the purposes, essential attributes, uses, and technical criteria for practical measurement, this essay aims to articulate a coherent framework that supports its rigorous and responsible application in educational contexts. Importantly, this vision of measurement aligns with and reinforces other bodies of work that seek to reposition assessment as a tool for learning and advancing instruction, particularly the work on *Assessment in the Service of Learning* (AISL) championed by Edmund W. Gordon and colleagues (e.g., The Gordon Commission, 2013; Gordon & Rajagopalan, 2016; Gordon, 2020). AISL similarly foregrounds the role of assessment in directly improving teaching and learning, emphasizing formative, diagnostic, and student-centered approaches over primarily evaluative summative judgments. Practical measurement and AISL share a commitment to equity, actionable feedback, and a shift in power toward learners and front-line educators. Practical measurement, with its emphasis on real-time data use, co-design with practitioners, and improving teaching and learning represents a concrete instantiation of AISL principles in action.

Moreover, the knowledge developed through research on and experience with practical measurement, particularly how these measures can be integrated into supports for teachers and how validity-in-use can be attained, offers useful insights that can advance the goals and practice of AISL. The similarities in purposes for and contexts of these two assessment traditions lead to complementary similarities in both how we assess and how we can use the resulting data to guide improvement. Further strengthening the bridge between these traditions will enrich both and move the field toward more just and effective uses of assessment.

Theoretical Foundations of Practical Measurement for Improvement

Improvement science in education is grounded in six interlocking principles that reframe both the purpose and the practice of inquiry for improvement:

- Make the work problem-centered and user-focused,
- View variability in performance as the problem to solve,
- See the system that produces current outcomes,
- Use measurement to inform judgment and improvement,
- Apply disciplined inquiry, and
- Accelerate learning through networked collaboration
  (Bryk et al., 2015; LeMahieu et al., 2017).

Measurement, while appearing only in the fourth principle, is foundational to all six. Without evidence generated from practice, the identification of problems, diagnosis of causes, evaluation of interventions, and broader system learning would be impossible. Practical measures are the instruments that supply this evidence, not in abstract or delayed form, but in direct, timely, and actionable ways.

Practical measurement stands in contrast to traditional assessment systems that typically privilege purposes such as accountability or academic research (Solberg, L., Mosser, and MacDonald. 1997). These systems usually employ standardized assessments that are administered infrequently, measure only distal outcomes, and return results too late to inform ongoing improvements to practice. Moreover, these systems tend to be disconnected from the specific concerns and goals of practitioners. In improvement science, however, the goal is not only to track what is happening, but to understand "why," "how," and "for whom" change is occurring. Practical measurement is specifically designed for this purpose.

The centrality of practitioners in this process marks a significant epistemological shift. Rather than treating frontline educators as implementers of measurement-driven prescriptions, improvement science positions them as co-inquirers—designing, testing, and refining practices while contributing to shared knowledge. Measurement becomes a generative process, not just an evaluative one.

Design and Use of Practical Measures

Practical measures are not simply shorter or quicker versions of traditional assessments; they are conceptually distinct. They are designed to be:

- *Aligned* with a working theory of practice improvement that indicates what needs to be measured,
- *Relevant and meaningful* to those closest to the work and responsible for students' intellectual, social, and moral development,
- *Actionable*, informing specific decisions about changes that are likely to be improvements,
- *Minimally burdensome*, fitting within educators' existing workflows, and
- *Timely*, both in frequency of administration and in providing needed feedback (Takahashi et al., 2022).

These criteria enable practical measures to support disciplined inquiry cycles such as Plan-Do-Study-Act (PDSA), providing real-time feedback on whether a change is leads to improvement.

Clarifying Purpose: Improvement vs. Accountability vs. Research

A central premise of practical measurement is that the "purpose" of an assessment should dictate its appropriate design, implementation, and use. Measurement for improvement differs in crucial ways from both accountability-driven testing and from measures developed for use in traditional academic research.

Accountability measures are often summative, standardized, and high stakes. They are typically externally mandated and designed to evaluate whether schools, teachers, or students have met predetermined benchmarks. Accountability testing happens infrequently and is usually extremely time-consuming as it has to cover broad performance domains. The lengthy timeframes for processing and returning results make accountability assessments lagging indicators that offer little guidance for real-time adaptation of practice. They reinforce what the system currently values most.

Research measures are usually designed for internal validity, generalizability, and theory testing. They may assess constructs of interest to researchers that are not directly actionable or even relevant to practitioners' concerns. They are comprehensive and measure all aspects of constructs that might be relevant to the theory under examination.

Practical measures for improvement, in contrast, are designed to be integrated into educators' daily work. They capture leading indicators and thus enable practitioners to determine whether a change they have made is an improvement, and what adjustments they might need to make. Their primary goal is to support practitioners' ongoing learning from and in practice, not to determine whether benchmarks have been attained or to investigate relationships between theoretical constructs.

Design Considerations

Designing effective practical measures involves balancing technical rigor with usability. This includes:

- Anchoring items to specific factors relevant to the theory of improvement and the changes being tested–be they the processes or tools being changed or their intended and unintended outcomes,
- Ensuring that measures are sensitive to changes in practice,
- Structuring data collection and interpretation so that they are integrated into existing routines, (existing routines, and)
- Designing reporting formats to facilitate interpretation and action.

Iterative refinement of measures is essential. Initial versions of a measure may not capture important differences in practice, may be misunderstood by users, or may produce data too late to be helpful to practitioners. Developers must remain attentive to feedback, patterns of use, and the interpretability of results.

Myths and Misconceptions

As interest in improvement science has grown, so too have misunderstandings about the role and nature of practical measurement. Four common myths are especially important to address:

Myth 1: Practical Measures Are "Quick and Dirty"

The descriptor "practical" can misleadingly suggest casual or imprecise design. In fact, the opposite is true. Effective practical measurement requires high levels of rigor, creativity, and iterative testing. These measures must be simultaneously predictive of meaningful outcomes and usable within the real constraints of classroom, school, and district work processes. Far from being quick and dirty, they are often the result of assiduous development processes involving repeated design-test-revise cycles.

Myth 2: Researchers Design, Practitioners Use

The traditional division of labor between knowledge producers and users undermines the development of useful and meaningful tools to support ongoing learning and development. Improvement science insists on collaborative co-design: practitioners, researchers, content experts, and improvement specialists working together to determine what needs to be measured, design instruments, and test usability. This not only enhances technical quality but ensures that measures are interpreted and acted on in ways that respect the complexity of practice.

Myth 3: A Single Measure Can Serve Both Accountability and Improvement

Measures designed for accountability often distort the learning processes they aim to monitor. The high stakes of many assessments can encourage superficial compliance, teaching to the test, or gaming. Conversely, measures for improvement require space for safe exploration, including failure and iterative refinement. Attempting to use the same measure for both purposes compromises each. It is therefore essential that systems create dual infrastructures: one for accountability and one for improvement (LeMahieu and Wallace 1986; LeMahieu and Reilly, 2004).

Myth 4: Any Data is Better Than No Data

While data are essential, poor-quality or misaligned data can do more harm than good. Misleading indicators can obscure problems, foster false confidence, or direct attention away from real causes. Practical measurement emphasizes the need for the "right" data, aligned with the improvement aims, interpretable by practitioners, and capable of guiding productive action.

Validity Considerations

Although practical measures are situated in real-world settings, they must still meet standards of technical quality. Two complementary concepts: "*validity-for-use*" and "*validity-in-use*" are essential for understanding their trustworthiness (See, for example, Messick, 1989; Shepherd, 1997; Moss, Girard, and Haniford, 2006; Bond 2013; Smith, 2025).

Validity-for-Use

This refers to whether a measure accurately captures what it is intended to assess. It can be conceived and examined in a number of ways:

- Face Validity: Does the measure appear appropriate to users?
- Content Validity: Does it cover the relevant domain?
- Construct Validity: Does it assess elements of the theory of improvement that are the focus of specific improvement efforts?
- Predictive Validity: Does it correlate with future outcomes?
- Concurrent Validity: Does it align with other established measures?

Importantly, because practical measures are integrated into current work routines, their design must be especially sensitive to trade-offs between brevity and construct coverage. Strategies such as mapping onto existing research findings, cognitive interviews with users, and triangulation with other data sources can help build a robust validity argument. Validity-for-use is typically conceived of as a characteristic of the measure itself.

Validity-in-Use

Validity-in-use goes beyond psychometric properties of the instrument itself to consider *how* practitioners might interpret the resulting data and *what* actions those data prompt. A technically valid instrument can be misused if it does not produce data that can inform and support constructive action or if users lack the support, shared understanding, or conducive context needed to use it productively. This compels a co-creative process that attends to systems of use as well as to the measure itself. This precludes, for example:

- Measures interpreted in deficit-based ways that may exacerbate inequities;
- Data that are used for evaluation rather than improvement can provoke anxiety and surface-level compliance; and
- In the absence of routines for collaborative sensemaking, even good data may fail to lead to improvement.

Validity-in-use requires that its actual use be aligned with its intended purpose, and that its use is supported by adequate infrastructure, professional learning, and leadership. Validity-in-use focuses not so much on the characteristics of the measure itself but on how the measure is used and in the contexts in which it is used.

Implications and Challenges

In our experience, the development and use of practical measurements for improvement raise several issues and challenges that must be addressed if the measures are to be used appropriately and most beneficially to inform improvements of practice. These include (LeMahieu and Cobb, 2025):

Equity and Inclusion

Practical measurement holds particular promise for advancing equity. The enduring questions of improvement science: "what works, for whom, and under what conditions" compel attention to variability in performance and thus to extant inequities, thereby highlighting where changes are (or are not) benefiting historically underserved populations. However, this potential will be realized only if measures are co-developed with attention to diverse contexts and interpreted in ways that avoid deficit framing. Measurement must support—not obscure—efforts to reduce disparities in opportunity and outcomes.

Capacity Building

Using practical measures productively requires new capabilities across role groups. Teachers need support in interpreting data as they enact cycles of improvement. Coaches, facilitators, and leaders must cultivate collaborative data use practices. Researchers must develop new competencies in designing measures that are psychometrically sound *and* practically usable.

Sustainability and Scaling

Practical measurement is not a one-size-fits-all endeavor. What works in one setting may not translate easily to another. Nonetheless, by building libraries of field-tested measures, open-access repositories, and adaptable tools, resources can be created that support local adaptation while retaining shared learning. Networks for improvement can accelerate this process by testing and refining tools across a range of contexts, not with the fidelity of standardization but with integrity in adaptation as the focus.

Participation in Development

The essential attributes of practical measurement for improvement compel new thinking about how best to ensure that they fully realize their promise. Traditional thinking about assessment would place primary responsibility for (and therefore agency in) developing measures with the technical experts of the psychometric community. In doing so, a number of difficulties can arise. This is, in part, because traditional procedures for determining and assuring quality in assessments can constrain the form and focus of assessment. This too often provides evidence that teachers do not find useful. Practical measures must adopt forms and formats of assessment that provide evidence and analytics that practitioners find relevant, meaningful, and actionable. In our experience, this is most effectively accomplished when practitioners are centrally involved in development efforts, with those providing technical expertise included as members of the development team.

Conclusion

Practical measurement for improvement represents a fundamental rethinking of the role of data in educational improvement. It is not merely a technical tool, but a social and organizational practice—an engine for professional learning, informed judgment, and improvement. Properly conceived and skillfully implemented, practical measures can help educators see problems more clearly, test ideas more effectively, and work toward equity with greater efficacy and efficiency.

However, realizing this vision requires ongoing attention to design quality, contextual appropriateness, and the social dynamics of data interpretation and use. It demands that educators, researchers, and leaders reimagine their roles, not as isolated actors, but as partners in inquiry. By embedding rigor into relevance, and structure into responsiveness, practical measurement helps fulfill the core promise of improvement science: that we can, in fact, *"get better at getting better."*

References

Bond, L. (2013). Toward a measurement science capable of informing and improving teaching and learning. In *Gordon Commission on the Future of Assessment in Education, To assess, to teach, to learn: A vision for the future of assessment (Technical Report).* Educational Testing Service.

Bryk, A. S., Gómez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better.* Harvard Education Press.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment (Technical Report).* Educational Testing Service.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W., & Rajagopalan, K. (2016). New approaches to assessment that move in the right direction. In E. W. Gordon (Ed.), *The testing and learning revolution: The future of assessment in education* (pp. 107–146). Palgrave Macmillan.

LeMahieu, P. G., Bryk, A. S., Grunow, A., & Gómez, L. M. (2017). Working to improve: Seven approaches to improvement science in education. *Quality Assurance in Education, 25*(1). Emerald Publishing.

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

LeMahieu, P. G., & Reilly, E. C. (2004). Systems of coherence and resonance: Assessment for education and assessment of education. *Yearbook of the National Society for the Study of Education, 103*(2), 189–202.

LeMahieu, P. G., & Wallace, R. C., Jr. (1986). Up against the wall: Psychometrics meets praxis. *Educational Measurement: Issues and Practice, 5*(1), 12–16.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*(1), 109–162.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8.

Smith, T. (2025). Validity and technical quality of practical measures. In P. G. LeMahieu & P. Cobb (Eds.), *Measuring to improve: Practical measures for improving teaching and learning.* Harvard Education Press.

Solberg, L. I., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: Improvement, accountability, and research. *The Joint Commission Journal on Quality Improvement, 23*(3), 135–147.

Takahashi, S., Peurach, D. J., Russell, J. L., Cohen-Vogel, L., & Penuel, W. (2022). Measurement for improvement. In D. J. Peurach, J. L. Russell, L. Cohen-Vogel, & W. Penuel (Eds.), *Handbook on improvement research in education.* Rowman & Littlefield.