CHAPTER 18

Measurement for Improvement

Sola Takahashi, Kara Jackson, Jon R. Norman, Marsha Ing, and Andrew E. Krumm

Improvement research rests on a foundational belief that transformative and lasting systemic change emerges as people within a given system learn through disciplined, ongoing inquiry as they enact their work. This chapter focuses on "measurement for improvement," or the deliberate and routine gathering, analysis, and interpretation of information with the distinct purpose of enhancing the learning of system actors as they test changes and improve processes that are at the heart of their work. Measurement for improvement happens within the context of a "learning loop" wherein what we will refer to as "improvers" articulate a theory of change, enact changes in the system, and gather data to inform their judgment about whether and in what ways the ideas worked and to guide associated changes to their theory and practice (Bryk et al., 2015; LeMahieu et al., 2017). Improvers include anyone participating in the continuous improvement effort, for example, teachers, instructional coaches, administrators, researchers, and others both located within and partnering with systems that are the focus of improvement.

Measurement for the purpose of *improvement* contrasts in important ways with dominant approaches to measurement in PK–12 education (Solberg et al., 1997). Especially in the United States, in light of the decades-long standards and accountability movement, a tremendous focus has been placed on annual lagging outcome measures for the purpose of accountability, and these data have been connected to rewards and consequences. This has resulted in a culture of data use that is often marked by judgment and a fear of data showing poor results. In contrast, measurement for improvement requires data that are connected to processes that are the object of improvement, collected and analyzed within the daily work lives of practitioners, and embedded in social routines and structures that enable collective sensemaking and actionable next steps. Generating and analyzing data for the purposes of improvement requires a culture of trust, which supports a willingness to acknowledge failures and make change.

Measurement for improvement also departs in critical ways from the role of measurement in the research, development, dissemination, and utilization (RDDU) model of educational research (Penuel et al., 2020). In an RDDU model, researchers and experts external to organizations use measurement to inform generalizable knowledge for the field, but not necessarily for the benefit of system members who are cast as the subjects of research. Improvement research calls for common ground among

researchers and system members, in which measurement activities in the context of ongoing collaboration leads to the leveraging of expertise and insight from both groups to improve teaching and learning in the here and now (Penuel et al., 2020).

Measurement for improvement is associated with the field of quality improvement in workplaces such as industry and hospitals (Provost & Murray, 2011; Solberg et al., 1997; Struebing, 1996). It was developed to support people across organizational silos, especially people on the front lines of work, to engage in broader systemic change efforts to test hypotheses about how to achieve an aim. While measurement for improvement has a decades-long history in other fields, its foray into the field of education is still at an early stage. We focus in this chapter on the measurement work that is integral to continuous improvement methods, which fall under the umbrella of improvement research.

Each of the five authors of this chapter has participated in continuous improvement efforts, spanning US K–12 education, postsecondary education, and the medical field, with an explicit focus on measurement.[1] We have each engaged to some degree with the Carnegie Foundation for the Advancement of Teaching, which has worked to advance continuous improvement methods in education. As such, each of our work has been influenced by key principles of networked improvement communities and the approach to improvement science as practiced by the Institute for Healthcare Improvement (to which the Carnegie Foundation looked as it started its work of continuous improvement in education). We are also practitioners of research methods traditionally utilized in educational research, and we see the distinctions between these established practices and the measurement needs specific to continuous improvement. We are also each deeply committed to redressing educational inequities in our work and research. While we have much to learn regarding how to leverage measurement for improvement to advance educational equity, we aim to identify how and in what ways equity concerns come to a head in this work, in an effort to support others to enact this work with equity at the center.

In the sections that follow, we first elaborate the broader context of measurement for improvement with a focus on the "learning loop" (Bryk et al., 2015). We then turn our attention to four key aspects of measurement for improvement, as represented in figure 18.1: (1) selection and design of measures for improvement; (2) establishment and maintenance of an analytic infrastructure; (3) building of social processes and routines for making sense of data; and (4) considerations of validity, or processes for assessing the quality of measures for the distinct purpose of improvement. Throughout our discussion of each of the aspects, we highlight issues of power and equity that arise, and we refer to specific examples of measurement for improvement when possible.[2] Finally, we reflect on challenges, as well as opportunities, in advancing measurement for improvement.

Context of Measurement for Improvement

As discussed earlier, a continuous improvement endeavor is marked by a "learning loop" composed of three mutually informing and evolving facets of the work: a

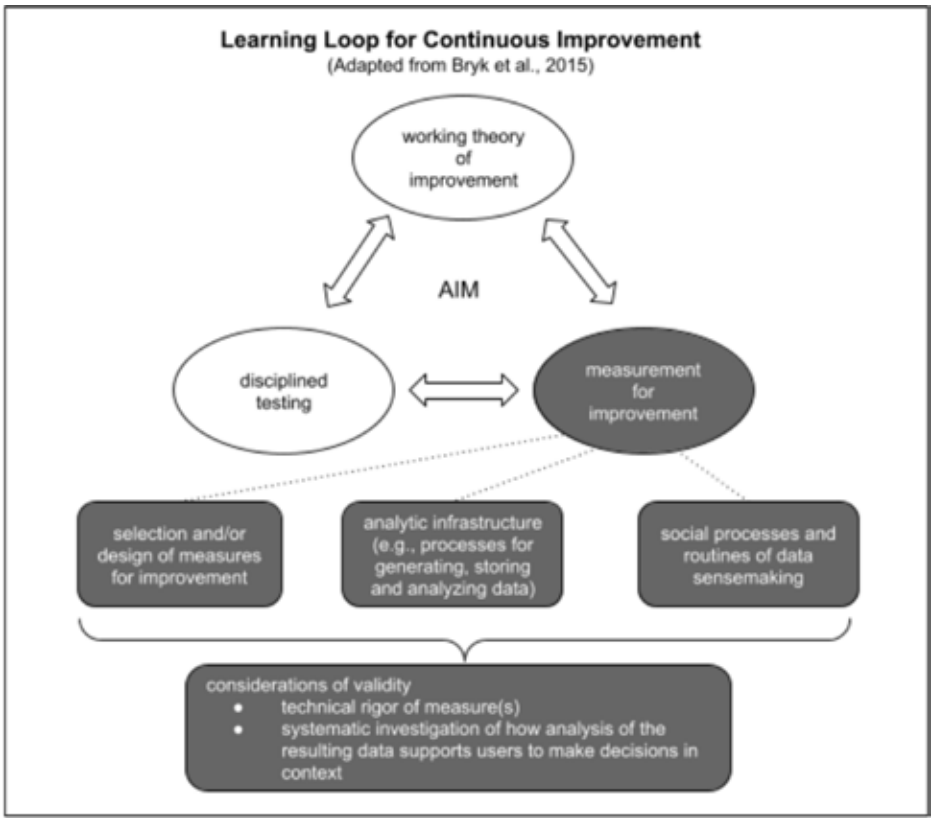

Figure 18.1. Learning Loop for Continuous Improvement
Adapted from Bryk et al. (2015).

working theory of improvement, disciplined testing, and measurement for improvement (figure 18.1). All three facets are centered on a specific aim, which articulates the aspiration for the work and includes specifics about what will be achieved, for and by whom, and when. To illustrate the learning loop, and specifically the role of measurement for improvement in the loop, we will draw on the example of the Building a Teaching Effectiveness Network (BTEN). BTEN was a networked improvement community based at the Carnegie Foundation for the Advancement of Teaching, which focused on improving the development and retention of early career teachers (Bryk et al., 2015; Hannan et al., 2015). The network included schools in the Austin Independent School District and Baltimore City Schools.

The first node of the learning loop is the *working theory of improvement*. This is an articulation of an improvement team's hypotheses about how the aim can be achieved. Theories of improvement specify the particular aspects of the system that are to be changed, and do so with practicality and parsimony: only those aspects of the system that are key levers in a causal logic chain within the system that leads to the outcome, and are within an individual's or team's locus of control, are focused on in improvement efforts (Bennett & Provost, 2015; Bryk et al., 2015).

The BTEN theory of improvement was initially developed through conversations among a broad array of experts on early career teachers, including academic scholars and educators who lead new teacher induction efforts. Improvers identified several processes of the PK–12 district systems that might serve as levers for improving teacher retention and development (e.g., hiring processes, professional development, feedback processes). From these aspects, partnering district teams identified the feedback process as a starting place. This prompted another phase of developing the improvement theory, whereby improvers specified aspects of the system where changes could improve the feedback that early career teachers received. These included improving the process by which teachers were observed and received feedback, the processes through which early career teachers are provided support as their next steps of development are identified, and the coordination among feedback providers.

It is crucial to generate and further specify a theory of improvement that closely reflects the realities of those who are experiencing the problem that is the target of improvement. Doing so requires eliciting the voices of the users of the system, like students, teachers, families, and community members, and valuing their unique expertise (Bryk et al., 2015; Hough et al., 2017). In BTEN, early career teachers were actively involved in surfacing the coordination of feedback among many providers as a key issue that was otherwise overlooked.

The second node of the learning loop is *disciplined testing*. At the more granular level of the working theory are specific changes and actions that will be taken in the system to achieve the aim. The changes are enacted through a testing framework. Whether this framework is a plan-do-study-act cycle or another form of inquiry cycle, a cycle typically involves a set of phases in which improvers (a) conjecture how the planned change is expected to result in an improvement; (b) enact the change and collect data to gauge the effects of the change; and (c) analyze the data to determine if the hypothesis is upheld, negated, or needs revision. Series of inquiry cycles—repeating the test in new sites and testing ideas at a larger scale—lead to a greater "degree of belief" (Langley et al., 2009) in the changes and actions, until they are ready for widespread adoption (or elimination). In the BTEN work, changes were made to the three facets of an effective feedback system: the feedback processes themselves, coordination efforts, and support for early career teachers to improve their practice. The changes included testing different ways to schedule regular observations and meetings between teachers and their observers, developing a conversation protocol that principals could use when giving feedback to teachers, and establishing routines for different feedback providers of the same teachers to coordinate their messaging and support, among others. Numerous inquiry cycles were carried out for each change idea in the relevant contexts.

The third node of the learning loop is *measurement*, which provides evidence for the working theory and for change efforts. Here, measurement does not refer to a single measure but, rather, a set of measurement efforts that support the learning of improvers. A family of measures (also referred to as a "system of measures") consists of outcome measures that capture information about the overall performance of a system, process measures that are proximally connected to specific aspects of the system (e.g., structures, norms, processes) that are high-leverage areas for the aim of the

improvement work, and balancing measures that enable improvers to keep an eye on other aspects of the system that may be inadvertently impacted by the improvement efforts in one part of the system (Bryk et al., 2015; Langley et al., 2009; Provost & Murray, 2011). While outcome measures tend to be already generated by education systems (e.g., graduation rates, students' grades in courses, test scores), systems often lack process measures (Bennett, 2018; Provost & Murray, 2011).

At the level of outcomes, the BTEN system of measures included the retention and instructional ratings of early career teachers. At the more granular level of key processes, the BTEN measures included the frequency of feedback provided to early career teachers, teachers' perceptions of the value of the feedback they received, and the coherence of feedback across feedback providers (Bryk et al., 2015). Early career teachers at the participating schools completed a survey approximately once every six weeks that solicited input on their perceptions and experiences. Feedback providers (e.g., administrators, department heads) completed an online form after each observation, which provided information about the frequency of classroom observations and the content of feedback conversations. Thus, as improvers made changes—such as the feedback conversation protocols or the observation and feedback scheduling routines—improvers examined the resulting data to see if the teachers' perceptions or feedback frequency matched hoped-for outcomes.

In what follows, we focus especially on measures that are proximal to the day-to-day work of improvement efforts. Ideally, improvers repeatedly administer measures for improvement to take a "temperature" read on whether the actions they and others have taken are leading to the desired changes. The resulting data informs conjectures about needed revisions to the improvement process and/or confirms that things are unfolding in a desired direction, and therefore users should "stay the course." Others have used the terms *practical measures* (Bryk et al., 2015; Yeager et al., 2013) and *pragmatic measures* (Kosovich et al., 2019) to describe these kinds of frequent, proximal measures that inform improvement. We focus on proximal measures in this chapter because they are essential for continuous improvement, they are not commonplace in education settings and are often missing at the start of an improvement effort, and they are commonly misunderstood as informal measurement.

Selecting and Designing Measures for Improvement

A crucial task in continuous improvement entails identifying the kind of information needed to quickly assess changes that people are making in practice. Especially in education, this is a non-trivial task. Different from other professions like health care, the kind of information readily available in educational systems are often summative measures of organizational outcomes that are too infrequent, too lagging, and too distal from system processes to serve as directly useful indicators of processes. As a result, improvers often need to figure out how to harvest information that already exists in the environment about specific processes, or design new measures that are tailored to

their particular improvement aims and context. In this section, as shown in table 18.1, we discuss a set of five essential features of measures for the purposes of improvement.

The first essential feature is that the measure is *closely tied to the working theory of improvement*. This relationship was discussed earlier in the chapter as the connection between the nodes of the learning loop—namely, between the theory of improvement and the family of measures to inform improvement efforts.

A second essential feature of measures for improvement is that the measures (and resulting data) are *meaningful to a broad range of improvers*, including those who are not trained in the technical aspects of measurement. Measurement for the purposes of improvement requires input from many types of experts, for example, experts on content (e.g., language comprehension), on practice (e.g., knowledge about teaching), on data and analytics (e.g., how to analyze and visualize data in meaningful ways), and on the impact of the system on those it intends to serve (e.g., insight about the experience of those receiving education services, such as students and families). It is crucial that the range of role groups who will be administering the measure and analyzing the resulting data are involved in the creation of the measure. For example, those working on developing the BTEN measures included researchers, system improvement coaches, and district leaders. Members of this team sought insights from early career teachers during the measurement development process, including conducting cognitive interviews with some of these teachers to understand how they interpreted survey items. The early and frequent involvement of key system members through the development phase helped ensure that the measures were ultimately written in a language that made sense and were meaningful to those using the data.

In addition, it is imperative that teams developing measures include perspectives of people who have often been marginalized or ignored within a system, within and across the role groups. Deciding on measures is ultimately a decision about what matters in the system—who gets to weigh in on this decision is thus tremendously consequential. Attending to multiple forms of expertise, and diverse perspectives, helps lift often excluded voices in the process of creating a measure. By including the representation of historically marginalized groups, improvement teams will be in a better position to identify measures that matter for the broader community and to build partnerships that can achieve the equity aims of the work.

The third essential feature is that the *resulting data are actionable*. By actionable, we mean that analysis of the resulting data will provide information to improvers that they can use to better understand their improvement efforts and can be examined to signal opportunities for further improvement. For example, BTEN system members

Table 18.1. Essential Features of Measures for Improvement

1. The measure is closely tied to the working theory of improvement.
2. What is being measured is meaningful to its users.
3. Resulting data are actionable.
4. Key data activities are minimally burdensome to users, including administration of the measure, data collection, and analysis of the results in successive inquiry cycles.
5. Data collection and analysis processes are timely.

used the data in various ways to inform their next steps. In one case, there was increasing evidence that teachers found the feedback they received across providers lacked coherence. In response, a school team worked to structure better communication across feedback providers working with the same early career teachers. The improvement of this communication process was evident in the next round of data where nearly all of the survey completers reported coherent feedback.

Data on their own do not suggest actions that people should take. However, for the purposes of improvement, it is important that any measure is "close to practice," meaning that people can use the data to identify patterns and variation in the practices that improvers are trying to affect (although they may need collegial support in order to do so, which we discuss later in the section on social processes of data use). As Moss (2016) has argued, especially for the purposes of improvement, it is important that people take what Murnane and colleagues (2009) refer to as a "conceptual" approach to data, as opposed to an "instrumental" approach. Instrumental approaches "entail using the results to make decisions [based] directly on . . . data without considering why . . . scores are low" while conceptual approaches focus on "identifying patterns . . . followed by systematic exploration of possible explanations, [which] requires collection and examination of other data" (p. 271).

The fourth key feature is that data collection and analysis processes are *minimally burdensome* to users, including that they are minimally disruptive of natural work processes. When possible, as mentioned earlier, it is helpful when improvers can make use of data that is already being collected in the local environment. Doing so eliminates the need for accompanying data collection tasks. As an example, researchers and Chicago Public Schools central office leaders used extant data, including data on students' attendance and grades, to develop "ninth-grade early warning indicators" for dropping out of high school (Allensworth, 2013). The data were already collected, but new ways of reporting the data allowed various role groups (e.g., teachers, school leaders, parents, students) to use the early warning indicators to inform the implementation of strategies to improve achievement and graduation rates.

However, in cases when new data need to be generated, it is important to ensure that the measure is simple and quick to administer. For example, measures that have a large number of items or take too long to administer are unlikely to be useful for the purposes of improvement because they won't be sustainable over the course of an improvement effort. To the extent that gathering data can be embedded in improvers' daily workflow, this can reduce the sense of data gathering as additional work.

Relatedly, the fifth key feature is that data collection and analysis processes are *timely*. Data are more likely to fulfill a feedback function when improvers can access the data shortly after they enact change efforts. Data are also optimally useful at a certain cadence and regularity—the exact cadence and regularity depend on the measure and the context of use. The extent to which measurement activities are minimally burdensome, as well as improvers' access to timely data, are dependent upon the existence of an analytic infrastructure (discussed in the following section) that supports the collection and analysis of measures for improvement in quick and efficient ways.

Establishing and Maintaining an Analytic Infrastructure: Processes for Administering Measures and Generating, Storing, and Analyzing Data

As shown in figure 18.1, another critical aspect of productively using measures for improvement concerns the establishment and maintenance of an analytic infrastructure for administering the measurement instruments, generating data, storing data, and analyzing data in a timely manner. Understood broadly, an infrastructure refers to "pervasive enabling resources in network form" (Bowker et al., 2010, p. 98), and it consists of the interaction of people, tools, and processes. Ultimately, the goal of developing an analytic infrastructure to support improvement efforts is to build capacity in a system to enable the right data to reach the right users in a timely and efficient way, across the system and over time (Star & Ruhleder, 1996).

Improvement efforts in education often require the generation of new data, storage, and analytic processes to facilitate the learning of improvers as they make changes in their system. For example, in BTEN, the data that informed the continuous improvement efforts primarily consisted of a teacher survey administered once every six to eight weeks and a classroom observation tool used by instructional leaders who gave feedback to early career teachers. These new instruments and their associated data required new processes of collecting and storing the data, analyzing the data, and reporting back the data in visualizations and summary statistics. To ensure that data collection and analysis was minimally burdensome and timely, many participating schools moved from varying modes of collecting teacher observation and feedback data to a singular online system. While it took time to set up the online system and to train feedback providers on how to use it, the new tool did not require additional data collection above and beyond what was collected previously, such as the date and observation notes. The improvement teams were then able to easily access information about feedback frequency. Ultimately, entering the data into the online platform enabled the ability to track key processes while embedding the data collection into the regular workflow of system members. In addition, the improvement team members responsible for analyzing the data created a process so that site leaders could receive the teacher survey data within five days after the survey window closed. Once these systems were established, they served improvers with timely data with minimal effort and resources. More generally, while establishing new processes necessarily takes additional time and work, once they are running, they should ultimately reduce the burden that would have been placed on individuals. A robust infrastructure is "self-renewing" (P. LeMahieu, personal communication, February 9, 2021) and adds capacity in the long term.

Establishing an analytic infrastructure for the purposes of improvement requires more than having trained data analysts with the proper tools, although that is essential. In education, organizational leaders, administrators, analysts, and teachers are some of the key participants who build an analytic infrastructure. Those within the

organization might partner with people in other organizations who play the role of an "analytic partner," who bring technical research skills, social skills, and knowledge about the problem at hand, to serve as thought partners and collaborators (Sherer et al., 2019). All of these improvers, within and outside the educational organization, set a vision and priorities, allocate resources to enable the work, make decisions about what data are being captured and how, utilize technologies to collect data and facilitate the transformation of data into information, make sense of the information to shape action, and provide feedback on the effectiveness of the infrastructure. The development of the BTEN online observation tool required the collaboration of principals who would be using it, the online tool developer who was an outside consultant, the district leaders who oversaw the support for early career teachers, and improvement specialists who brought insight into the types and cadence of data that inform improvement efforts. As principals and teachers used the online observation tool, they provided feedback on how it worked (and how it didn't) to enable refinement over time.

An analytic infrastructure reflects the values and priorities of a community, and key decision points along the processes of the data infrastructure matter for advancing an equity agenda. For example, whose perspectives and practices are being measured, and who will therefore be represented in the data that is generated? How will data be visualized, and how will those decisions shape both the kinds of questions users ask of the data, and the resulting actions users take? One key issue concerns deciding how to attend to variation within a system and over time, including when it makes sense to disaggregate data and along what dimensions. Will users be able to easily investigate issues of variation, including what is typical variation, and what is undesirable variation (Berwick, 1991)? Centering equity also means attending to who is involved in making these decisions, and what feedback loops are established to enable a variety of voices to shape the infrastructure (Ahn et al., 2019).

Building Social Processes and Routines for Data Sensemaking

It is tempting to believe that the administration of the "right measures" and analysis of the resulting data will, rather straightforwardly, lead improvers to revise their existing theories toward a more robust and accurate theory of change. However, research has shown that when education professionals engage in data interpretation activities, it is rarely in a straightforward manner where information feeds directly into concrete decisions (Coburn et al., 2009; Spillane & Miele, 2007). Thus, another key aspect of measurement for improvement concerns attending carefully to the social processes and routines of data sensemaking (see figure 18.1). Social and organizational conditions and power relations matter for any data sense-making routine in an educational setting (Coburn & Turner, 2011), but, as we discuss below, they matter in a distinct way in the context of a continuous improvement context.

One critical condition regards norms of interaction that support deep and sometimes uncomfortable inquiry into practice during data sense-making routines—transparency and psychological safety. A continuous improvement approach depends on acknowledgment of practices that are not going well, and a willingness to innovate and take risks in practice. In many educational systems, there are long-standing norms of interaction that are not conducive to a continuous improvement approach. For example, it is common for practice to be private; teachers are unlikely to have opportunities to see into each other's classrooms. And, in cases where practice has been deprivatized, it is often for the purpose of evaluation (e.g., a principal observes a teacher's classroom as part of an assessment). Thus, in many educational systems, examining data for the purposes of improvement will require intentional work in establishing new ways of interacting around data regarding people's current practices.

For example, in the Literacy Improvement Partnership, a Regional Educational Lab–West project between WestEd and the Washoe County School District in Nevada, network leaders designed discussion protocols for teacher grade-level teams to use to examine data, home in on specific challenges, discuss potential solutions, and commit to next steps (Austin et al., 2019). Every grade level team included one teacher who was the facilitator for their team, who received additional support and coaching from the network leaders to support the work of the teams. These discussion protocols transitioned over time from teachers looking at data from their own classrooms, to eventually looking at data about certain instructional practices across all the classrooms in a grade level. Over time, a greater proportion of teachers reported that their team meetings were focused, that teachers were more equitably engaged across their team, and that the depth of conversation improved.

Another condition regards improvers' orientation toward the data and explanations of patterns that emerge in the data (Coburn & Turner, 2011). Engaging in continuous improvement should provide clear opportunities to articulate and critically analyze assumptions, revising theories to be more effective and just. However, a key challenge in using data to reflect on and adjust practices is that users tend to focus on patterns in data and ascribe meaning to those patterns that support their extant beliefs and theories (Coburn & Turner, 2011). This can be particularly problematic when these existing beliefs reflect bias and deficit-oriented beliefs about students or educators based on race, class, gender, or other characteristics.

To illustrate this challenge, we draw from work associated with the Practical Measures, Routines, and Representations (PMRR; https://www.pmr2.org//) project, which consists of three research-practice partnerships. In each partnership, a team of university-based researchers collaborated with a large school district to improve the quality of instruction in secondary mathematics classrooms, and thus student learning, with an emphasis on students who have been disadvantaged by the current educational system. While the details of each of the partner district's theory of improvement varies somewhat, the districts share a research-based vision of what counts as high-quality instruction, in which students are provided regular opportunities to solve cognitively demanding problems, communicate their thinking, and make sense of key mathematical ideas (e.g., National Council of Teachers of Mathematics, 2014). Further, each district's theory of improvement reflects the principles that improving instruction

requires ample support for teachers' learning, and that it requires building coherence between the support for teachers' learning and other aspects of the school and district system (e.g., principals' expectations for instruction, district leaders' expectations for principals) (Cobb et al., 2018).

To support the districts' improvement efforts, the PMRR team developed a set of measures for improvement that support practitioners to assess students' perceptions of key aspects of the mathematics classroom learning environment that research has linked to student learning. As one example, the *whole-class discussion measure* takes the form of a student-facing survey that can be completed in three minutes or less, which assesses students' perceptions of critical aspects of whole-class discussions that advance student learning (e.g., what students are held accountable for in the discussion, opportunities to make sense of others' ideas, extent to which students want to share their ideas and feel their ideas are valued). One issue that has arisen across partnerships concerns users' perspectives on their students' current capabilities. There have been cases in which teachers who do not currently view their students as capable of making sense of mathematics dismiss the resulting data, and, worse, view the data as further evidence of deficits in the students (Jackson et al., 2019). Unfortunately, this response is not entirely surprising. There is a tendency in educational systems to treat troubling data about instruction as reflective of deficits in individual students and in their families (Jackson et al., 2017).

In light of these findings, the PMRR team considered suggesting that the student-facing surveys only be used by teachers who view their students as capable of engaging in the intended forms of practice, and who see value in learning about students' perspectives. Coaches reported, however, that examining students' responses with teachers, alongside observational data about instruction and student work, provided them with insight into teachers' perspectives on their students' capabilities—perspectives they believe the teachers would not have discussed as openly had they not examined students' responses (Kochmanski, 2020). On the basis of this information, coaches could then adjust their goals for teachers' learning and the nature of the support they provided.

In addition, power relations matter in routines for making sense of data, just as they do in the design of measures for improvement. One issue concerns who is invited to make sense of what data, at what times, and for what purposes. For example, school and district leaders across a number of improvement efforts have shared data about students with students, to have students inform the educators' understandings of what the data mean to them, and what next steps might be taken. A second issue concerns the authority relations between various participants (e.g., principal reviewing data with a grade-level team of teachers) in a data sense-making structure (Coburn & Turner, 2011). Discussion protocols can be used to prompt wider distribution of participation. Such protocols might include "equity pauses" (e.g., High Tech High, https://hthgse.edu/research-center/protocol-library/) to promote reflection on assumptions being made. Norms can be explicitly set at the outset to address prejudices that may be aired in the conversation, such as taking an asset rather than deficit orientation about students and their families.

Establishing Validity Arguments in Measurement for Improvement

As described earlier, measurement for improvement hinges on a unique combination of factors around context, design, and social processes. But how do we know if a measure, or set of measures, can be used effectively in the context of a continuous improvement effort? How can we attend to the intended and unintended consequences of the use of these measures to advance issues of equity? These are essential technical and ethical questions that implicate issues of validity when engaging in measurement for the purpose of improvement.

Foundational to the work of measurement validity is the idea of "evidentiary arguments" (Haertel, 2013; Kane, 2016; Mislevy, 2007). Mislevy, drawing on Toulmin (1958), describes evidentiary arguments as follows:

> We wish to support a claim with data. A warrant is a generalization that justifies the inference from the particular data to the particular claim. Theory and experience back the warrant. Alternative explanations can weaken the argument, which in turn may be supported or weakened by rebuttal evidence. Validity accrues as the warrant better fits the circumstances at hand, as its backing is stronger and comes from different sources, and as more alternative explanations can be countered more convincingly. (Mislevy, 2007, p. 464)

From the perspective of argumentation, traditional issues of validity are recast not as properties of measures themselves but as evolving interpretations and uses of scores (a term we use here to refer to results of a measurement process) in specific contexts. Building on research in educational measurement (e.g., Moss et al., 2006), we suggest that addressing issues of validity in the interpretation and use of measures for improvement requires generating evidentiary arguments that take into consideration the unique purposes and characteristics of improvement efforts.

In what follows, we discuss establishing evidentiary arguments for measures used in continuous improvement projects. The first part of our approach concerns investigating the technical rigor of a measure for improvement: Does it assess what it claims to assess? The second part of our approach concerns investigating validity in the context of using the measure for improvement (Moss et al., 2006): What interpretations do users make with respect to the data, and how, if at all, are those in service of improvement? How does the context of use shape the kinds of interpretations people make, and their subsequent action?

TECHNICAL RIGOR OF IMPROVEMENT MEASURES

Earlier we described five essential features of measures for improvement (see table 18.1) that should guide the selection and design of measures. While these features are foundational to the validity of these measures (e.g., measures are more likely to

be used for their intended purposes if they are meaningful to users), they are not the criteria for establishing the technical rigor of a measure. How an improvement team builds confidence about the technical rigor of an improvement measure will vary by the nature of the measure.

One central factor in establishing whether the measure assesses what it claims to assess concerns the extent to which the measure requires low or high inferences to interpret resulting data. Low-inference measures are often measures of easily observable processes or behaviors, such as the BTEN example regarding the frequency of feedback given to teacher candidates. It is easier to be confident that the measure is capturing the intended object of measurement if there is minimal ambiguity between what is being measured and what is being inferred. The higher inference the measure is, the more challenging it is to connect the measure to the object of measurement, and therefore the more work involved in confirming that an object of measurement is adequately captured.

An example of establishing the technical rigor of a higher-inference measure occurred in the development of the PMRR measures (Jackson et al., in press). To develop the whole-class discussion survey measure, improvers conducted five iterative cycles of design, analysis, and revision to the survey. If the measure was to support various improvers to assess the success of their current improvement strategies and to set goals for future work, it was important that the measure distinguish between whole-class discussions that prioritized sensemaking and were likely to support students to advance their conceptual understanding of key mathematical ideas, and those that did not. Thus, in each cycle, improvers, who included district math specialists, coaches, and researchers, observed a range of lessons that included whole-class discussions that varied in the quality of student learning opportunities, and administered the then-current version of the survey to students at the end of the whole-class discussion. Improvers then conducted cognitive interviews with a range of students (i.e., students who had more and less fully participated in the whole-class discussion) immediately following the lesson, in which students were asked to explain why they chose certain responses and to indicate any language that was unfamiliar. The team then analyzed the student interviews in relation to their classroom observations to determine whether students' responses reflected outside observers' assessment of the lesson, including whether the student responses in the aggregate appeared to distinguish discussions of varied learning opportunities, and to clarify issues of language. The team then revised the survey on the basis of the analysis. By the end of the five cycles, which included observations and administration of the survey in 15 classrooms, across two districts, and cognitive interviews with 81 students, the team felt that they had solid evidence to suggest that the items communicated as intended to students, and could be used to meaningfully discriminate in the quality of students' learning opportunities in whole-class discussions. In addition, observations indicated that students typically completed the survey in 1–3 minutes, which suggested that at least the administration of the survey was minimally disruptive to class instruction.

It can also be useful to establish an evidentiary argument that a measure predicts what one intends to predict, in relation to an evolving theory of improvement

(see the chapter "Data-Intensive Improvement" in this volume for new advances in establishing predictive validity). The ultimate goal of improvement efforts is to improve key educational outcomes. Examining this relationship between measures of focal areas and key outcomes can build confidence that the measures of the focal area do, in fact, capture something meaningful in an improvement effort. An example of establishing a predictive relationship occurred in the development and use of the Student Electronic Exit Ticket (SEET) (Penuel & Watkins, 2019). The SEET captures students' experience of project-based science instruction by asking them to report on the perceived coherence of the lesson, the relevance of the lesson, and the degree to which they contributed in class and whether those contributions were recognized by other students. The instrument is easy to administer and is intended to be administered multiple times throughout a unit to inform both teachers and curriculum developers working to improve science units. The SEET has been used in several projects as an improvement measure, and during its development, researchers undertook confirmatory work to investigate whether the instrument could be administered easily, whether data collected through the SEET were usable and valuable, and that scores from the SEET were internally coherent as well as predictive of external measures like student assessment scores. For example, Penuel and colleagues (2018) held informal discussions with teachers implementing a new science curriculum, monitored overall use of the SEET, asked teachers to complete a survey on the value of the SEET, conducted observations of teachers' classrooms related to SEET items, and correlated SEET items with scores from robust unit-specific assessments.

Evidence that the SEET assessed what it was intended to assess and was correlated with learning outcomes supported its subsequent use in OpenSciEd, a consortium of materials developers and state science leaders that provides free, open, and high-quality resources and implementation supports for project-based science in grades K–12 (Edleson et al., 2021). Krumm and colleagues (2020) used the SEET as one of several data sources for improving OpenSciEd instructional and professional learning materials. In particular, the SEET was used to examine the ways in which students from different demographic backgrounds experienced science instruction across units being developed and tested, in order to support the consortium's aim that the materials advance equity.

Issues of equity are central to understanding the interpretation and use of any measurement process, including measures for the purposes of improvement. It is incumbent upon designers of measures to critically reflect on how their measures may be misrepresenting the experiences of certain subgroups of individuals, particularly students who have been historically underserved by the educational system. Correcting for bias in the measurement tool is essential for developing measures that support inquiry about educational inequities. Ultimately, the actions that are taken in an improvement effort should not incur negative impacts on minoritized groups due to poorly designed measures. Continual confirmatory work leveraging participation of a broad coalition of stakeholders in improvement efforts can surface potential biases and unintended consequences.

VALIDITY IN THE USE OF IMPROVEMENT MEASURES

Because the goal of continuous improvement work is to actually make improvements in the system, having confidence about improvement-related decisions based on the measures means seeing its journey from design through use. Building on the data-driven decision-making literature, Moss (2016) argues that validity is not only a technical issue about design and intended use but also an interpretative issue about actual use, or the ways in which improvement teams interact with the measure and the resulting data. When use of a measure is expanded to new contexts, even if there is initial consistency between the intended use as defined by the developers and as articulated by the improvers, arguments need to be made for using the measure in a new context (Ing et al., 2021).

Improvement teams can gather ongoing evidence to build an argument about validity-in-use in a variety of ways to answer fundamental questions related to how data are connected to claims. To examine validity-in-use, Moss (2016) argues for the importance of

> an ongoing research agenda that examines (a) the way professionals in different roles and contexts interact with and use data; (b) the organisational resources at different levels of the system that support or constrain their practice of data use; and (c) the ways in which different approaches to data use impact the practices of education professionals and organisations as well as the learning of their students. (p. 248)

For example, although the PMRR team invested substantial effort to ensure the technical rigor of the measure of whole-class discussion as part of the design process described above, the team continues to investigate "validity-in use" of the measure (Ing et al., 2021; Jackson et al., in press). This investigation includes gathering evidence of the interpretations and actions various users make and take, respectively, in relation to specific improvement strategies in a range of contexts, and for a range of specific purposes. For instance, in one district whose theory of improvement focused on the implementation of high-quality coaching cycles, team members gathered and analyzed data (e.g., audio-recordings of co-planning and debrief conversations between coaches and teachers, field notes of classroom observations) to investigate whether changes in students' responses to the survey appeared to correspond to changes in the teacher's instruction; and when coaches and teachers used classroom measure data productively to identify and negotiate goals for instruction, as well as cases when the coaches and teachers did not (Kochmanski, 2020). Analyses of the use of the classroom measures in different districts supported the specification of "conditions of use," for example, what appears to be important in terms of coaches' expertise and practice in reviewing data with teachers.

Concerns of equity are central in the work of building an argument for validity-in-use. A key question to investigate is how measures are prompting critical reflection on individual or institutional bias, and in service of shifts toward more equitable practices. The PMRR example described in the "Social Processes" section earlier in this chapter

regarding how the use of the measure brought to the fore teachers' deficit-oriented beliefs about some of their students is an example of the type of work that researchers have done to examine the validity-in-use of a measure in relation to equity.

ACCRUAL OF EVIDENCE OF VALIDITY ACROSS USE CASES

Attending to the validity of measures used in continuous improvement projects can, at times, be complex, labor-intensive work. As of yet, there are few continuous improvement projects in education in which the validity arguments for measurement are thorough, well-established, and documented. Often, in the early phases of an improvement project, it is not yet clear what should be measured, let alone whether those measures have strong validity evidence. Improvement teams may be conducting root cause analyses and developing their early theories of improvement, and testing changes in some areas only to discover that they need to refocus their work. Furthermore, the levels of evidence of validity needed when one teacher is considering recommending a practice to a colleague to try in her class differs from the level of evidence needed when a district is deciding whether to implement a practice across all of its schools. The work of building the validity argument, therefore, is on a developmental path itself, alongside the other aspects of the continuous improvement work.

The building of a validity argument also extends beyond any individual continuous improvement effort in a specific context. It is reasonable to expect that evidence for the validity of a measure can accumulate across use cases and contexts, just as it does with measures for purposes other than improvement. For this to happen, it is crucial that the field establishes an infrastructure for collecting and sharing findings regarding conditions of use for particular measures across cases. Currently, one of the authors, Sola Takahashi, is engaged in a project at WestEd, funded by the Bill & Melinda Gates Foundation, to collect measures that inform continuous improvement in grades 6–9 math teaching and learning. As part of this work, the team is collecting information about how each measurement instrument has been used in prior continuous improvement work. This includes information about any evidence of usefulness, the measurement infrastructure and process that enabled its use, the frequency of use, analytic decisions, supporting factors, and challenges of use. The project aims to advance the analytic work of math-focused improvement efforts of educators who may otherwise each be in the situation of developing their own measurement tools. The building of the field of measurement for improvement in education will depend on pursuing the establishment of validity arguments for measures over time and across improvement efforts.

Conclusion

Measurement is an essential part of a continuous improvement approach to system transformation; however, how to engage in measurement for improvement, in practice, is often elusive. We have not yet, as a field, fully invested in identifying,

developing, and enacting measurement of key processes in our systems or other leading indicators for valued outcomes. As a result, educators and researchers embarking on continuous improvement work are often in the position of developing measures to fill the gap. We offer this chapter as guidance for those aiming to do this work, and to make clear that the work of selecting or developing measurement instruments is only part of the effort of measurement for improvement. Because the concerns in this work are ultimately about the sense that system members make of the data, and what decisions and actions result from this sensemaking, attending to the analytic infrastructure, social processes of data use, and validity in the design and use of the measures are also essential aspects of measurement for improvement.

There is the potential of great harm in using improvement measures for evaluative or judgment purposes. The technical requirements of measurement for evaluation and accountability differ from the requirements we have outlined in this chapter for measurement for improvement, and thus, many improvement measures will be ill suited for evaluation. Furthermore, the data culture surrounding measurement for accountability tends to be in opposition to that supporting measurement for improvement, where transparency, trust, and a sense of safety are essential.

Throughout, we have called attention to questions about how the work of measurement for improvement can advance equity, understanding that the work of antiracism and redressing additional, and intersectional, inequities takes intentionality and deliberate action. This attention to equity is essential in each aspect of the work discussed in this chapter. Anytime we develop or identify a measure, we are necessarily making choices about what counts as important to attend to, and what is not. When we design and implement a measurement infrastructure, we make decisions about who is generating the data we gather and how to represent these data, including how to categorize and represent groups of people, often students. When improvers make sense of data, some interpretations gain more traction than others, and interpretations have the potential to push against or reinscribe the status quo. It is therefore necessary to ensure a multiplicity of actors are part of the work depicted in each of the facets of the learning loop, in which a theory of change is crafted, measures are created, and data are interpreted, with special attention to the voices and interpretations of actors who are often disenfranchised by educational systems. Future work in measurement for improvement must continue to critically examine both the challenges and possibilities in using measurement to advance educational equity.

In cases where measurement for improvement is happening in the context of a research-practice partnership, there may be conflicts between the research agenda for the researchers and the priorities of the system members. In some cases, the two groups may have different questions they want to answer. In other cases, the researchers' press for generalizable knowledge may come up against the system members' interest in solving a specific local problem in the present time. In our experience, it has been helpful to keep both sets of aims, interests, and questions on the table, but to start with the needs and interests of the local system members at the center. The work of continuous improvement is not possible without the interest and investment of system members, and prioritizing their needs and interests is helpful not only in understanding what is needed and what works at the ground level of education but

also in developing relationships of trust and collaboration across organizations at the outset of the work. This can help signal some of the shifts this work can bring about in knowledge hierarchies.

There is a tendency to suggest that measurement for improvement is easy—and therefore lacking rigor; however, our experience points to the contrary. The work of designing measures specific to processes that are the target of improvement and ensuring validity-in-use is resource-intensive and time-consuming work, and it takes members of a continuous improvement effort who are dedicated to the measurement aspects of the effort, while simultaneously integrated into the broader improvement efforts. Even considering these challenges, we see the endeavor of measurement for improvement in the context of continuous improvement to hold great promise, in providing useful information that supports educators in making and sustaining positive changes in their system. We urge educational researchers and system leaders to expand the field's capacity to provide and support the use of these measures in educational improvement efforts. As such work is pursued, we continue to need venues in which this work can be shared to enrich collective learning.

Notes

1. We are very grateful for the valuable comments provided by two reviewers. The contributions of Kara Jackson and Marsha Ing are based upon work supported by the National Science Foundation under grant numbers 1620851 and 1621238. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

2. For more examples of measurement for improvement, see Takahashi et al. (2020).

References

Ahn, J., Campos, F., Hays, M., & Digiacomo, D. (2019). Designing in context: reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics*, *6*(2), 70–85.

Allensworth, E. (2013). The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR), 18*(1), 68–83.

Austin, K., Bolte, J., & Takahashi, S. (2019, April 16–18). *Improvement on the front lines: Using learning huddles to shift instruction* [Conference presentation]. Carnegie Foundation Summit on Improvement in Education 2019 Conference, San Francisco, CA.

Bennett, B. (2018, September). Branching out: Use measurement trees to determine whether your improvement efforts are paying off. *Quality Progress*. https://qi.elft.nhs.uk/wp-content/uploads/2018/09/QP_Branching-Out_Measurement-Tree_20180901.pdf

Bennett, B., & Provost, L. (2015, July). What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress*. https://qi.elft.nhs.uk/resource/whats-your-theory-2/

Berwick, D. M. (1991). Controlling variation in health care: A consultation from Walter Shewhart. *Medical Care, 29*(12), 1212–1225.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 97–117). Springer.

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

Cobb, P., Jackson, K., Henrick, E., Smith, T., & MIST Team. (2018). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Harvard Education Press.

Coburn, C. E., Honig, M. I., & Stein, M. K. (2009). What's the evidence on districts' use of evidence? In J. D. Bransford, D. J. Stipek, N. J. Vye, L. M. Gomez, & D. Lam (Eds.), *The role of research in educational improvement* (pp. 67–87). Harvard Education Press.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement, 9*, 173–206.

Edelson, D. C., Reiser, B. J., McNeill, K., Mohan, A., Novak, M., Mohan, L., Affolter, R., McGill, T., Buck Bracey, Z. E., Deutch, N. J., Kowalski, S., Novak, D., Lo, A. S., Landel, C., Krumm, A., Penuel, W. R., Van Horne, K., González-Howard, M., & Suárez, E. (2021). Developing research-based instructional materials to support large-scale transformation of science teaching and learning: The approach of the OpenSciEd Middle School Program. *Journal of Science Teacher Education, 32*(7), 780–804.

Haertel, E. (2013). Expanding views of interpretation/use arguments. *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 68–70.

Hannan, M., Russell, J. L., Takahashi, S., & Park, S. (2015). Using improvement science to better support beginning teachers: The case of the Building a Teaching Effectiveness Network. *Journal of Teacher Education, 66*(5), 494–508.

Hough, H. J., Willis, J., Grunow, A., Krausen, K., Kwon, S., Mulfnger, L., & Park, S. (2017). *Continuous improvement in practice.* Policy Analysis for California Education.

Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use this measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice, 40*(1), 92–100. https://doi.org/10.1111/emip.12393

Jackson, K., Cobb, P., Ing, M., Ahn, J., Smith, T., Kochmanski, N., Chinen, S., & Nieman, H. (in press). Developing and using practical measures to inform instructional improvement in mathematics at scale. In P. LeMahieu & P. Cobb (Eds.), *Practical measurement for improvement*. Harvard Education Press.

Jackson, K., Gibbons, L., & Sharpe, C. (2017). Teachers' views of students' mathematical capabilities: Challenges and possibilities for ambitious reform. *Teachers College Record, 119*(7).

Jackson, K., Nieman, H., & Kochmanski, N. (2019, April). *Making sense of teachers' varied responses to representations of practice* [Paper presentation]. National Council of Teachers of Mathematics Research Conference, San Diego, CA.

Kane, M. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211.

Kochmanski, N. (2020). *Aspects of high-quality mathematics coaching: What coaches need to know and be able to do to support individual teachers' learning* [Doctoral dissertation]. Vanderbilt University.

Kosovich, J., Hulleman, C., & Barron, K. (2019). Measuring motivation in educational settings: A case for pragmatic measurement. In K. Renninger & S. Hidi (Eds.), *The Cambridge handbook of motivation and learnin*g (pp. 713–738). Cambridge University Press. https://doi.org/10.1017/9781316823279.030

Krumm, A., Penuel, W. R., Pazera, C., & Landel, C. (2020). Measuring equitable science instruction at scale. In M. Gresalfi & I. S. Horn (Eds.), *The interdisciplinarity of the learning sciences* (Vol. 4, pp. 2461–2468). 14th International Conference of the Learning Sciences (ICLS) 2020, Nashville, TN.

Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). Jossey-Bass.

LeMahieu, P. G., Bryk, A. S., Grunow, A., & Gomez, L. M. (2017). Working to improve: Seven approaches to improvement science in education. *Quality Assurance in Education, 25*(1), 2–4.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463–469.

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236–251.

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

Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2009). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for Students Placed at Risk (JESPAR), 10*(3), 269–280.

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. https://www.nctm.org/PtA/

Penuel, W. R., Riedy, R., Barber, M. S., Peurach, D. J., LeBoeuf, W. A., & Clark, T. (2020). Principles of collaborative education research with stakeholders: Toward requirements for a new research and development infrastructure. *Review of Educational Research, 90*(5), 627–674. https://doi.org/10.3102/0034654320938126

Penuel, W. R., Van Horne, K., Jacobs, J., & Turner, M. (2018). *Developing a validity argument for practical measures of student experience in project-based science classrooms* [Paper presentation]. American Educational Research Association Annual Meeting, New York.

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *Annals of the American Academy of Political and Social Science, 683*(1), 201–216.

Provost, L. P., & Murray, S. (2011). *The health care data guide*. Jossey-Bass.

Sherer, D., Norman, J., Bryk, A. S., Peurach, D. J., Vasudeva, A., & McMahon, K. (2019). *Evidence for improvement: An integrated analytic approach for supporting networks*. Carnegie Foundation for the Advancement of Teaching.

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

Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. In P. A. Moss (Ed.), *Evidence and decision making. Yearbook of the National Society for the Study of Education, 106*(1), 46–73.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Struebing, L. (1996). Measuring for excellence. *Quality Progress, 29*(12), 25–30.

Takahashi, S., Norman, J., Jackson, K., Ing, M., & Chinen, S. (2020). Measurement for improvement in education. In D. Peurach & J. Russell (Eds.), *Oxford bibliographies in education: Scholarship on improvement*. Oxford University Press.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Yeager, D., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Carnegie Foundation for the Advancement of Teaching.