CHAPTER 4

Measuring What Matters

Principles for a Balanced Data Suite That Prioritizes Problem Solving and Learning

Kate Bridges and Michael Woolcock

SUMMARY

Responding effectively and with professional integrity to public administration's many challenges requires recognizing that access to more and better quantitative data is necessary but insufficient. An overreliance on quantitative data comes with risks, of which public sector managers should be keenly aware. We focus on four such risks: first, that attaining easy-to-measure targets becomes a false standard of broader success; second, that measurement becomes conflated with what management is and does; third, that an emphasis on data inhibits a deeper understanding of the key policy problems and their constituent parts; and fourth, that political pressure to manipulate key indicators, if undetected, leads to falsification and unwarranted impact claims or, if exposed, jeopardizes the perceived integrity of many related (and otherwise worthy) measurement efforts. The cumulative concern is that these risks, if unattended to, will inhibit rather than promote public sector organizations' core problem-solving and implementation capabilities, an issue of high importance everywhere but especially in developing countries. We offer four cross-cutting principles for building an approach to the use of quantitative data—a "balanced data suite"—that strengthens problem solving and learning in public administration: (1) identify and manage the organizational capacity and power relations that shape data management; (2) focus quantitative measures of success on those aspects that are close to the problem; (3) embrace a role for qualitative data and a theory of change, especially for those aspects which require in-depth, context-specific knowledge; and (4) protect space for judgment, discretion, and deliberation because not everything that matters can be measured.

Kate Bridges is an independent consultant. Michael Woolcock is the lead social scientist in the World Bank's Development Research Department.

ANALYTICS IN PRACTICE

- Identify and manage the organizational capacity and power relations that shape data management. Make professional principles and standards for collecting, curating, analyzing, and interpreting data clear to all staff—from external consultants to senior managers—in order to affirm and enforce commitments to ensuring the integrity of the data themselves and the conclusions drawn from them. Make measurement accountable to advisory boards with relevant external members. Communicate measurement results to the public in a clear and compelling way, especially on contentious, complex issues.
- Focus quantitative measures of success on those aspects that are close to the problem. Ensure that the measurement approach itself is anchored to a specific performance problem. Target measurement investments at those performance problems that are prioritized by the administration. Ensure that any judgments on an intervention's success or failure are based on credible measures of the problem being fixed and not simply on output or process metrics. Where measures of success relate to whether the intervention is functioning, allow flexibility in the implementation of the intervention (where possible) and in the related measurement of its functioning. In this way, implementation strategies can shift if it becomes clear from the collected data that they are not making progress toward fixing the problem.
- Embrace an active role for qualitative data and a theory of change. Include qualitative data collection as a complement to quantitative data. This may be a prelude to future large-scale quantitative instruments or perhaps the only available data option for some aspects of public administration in some settings (such as those experiencing sustained violence or natural disasters). Draw on qualitative methods as a basis for eliciting novel or "unobserved" factors driving variation in outcomes. Tie measurement (both qualitative and quantitative) back to a theory of change. If the implementation of an intervention is not having its intended impact on the problem, assess whether there are mistaken assumptions regarding the theory of change.
- Protect space for judgment, discretion, and deliberation because not everything that matters can be measured. Consider carefully what you choose to measure, recognizing that whatever you choose will inevitably create incentives to neglect processes and outcomes that cannot be measured. Actively identify what you cannot (readily) measure that matters and take it seriously, developing strategies to manage that as well. Identify those aspects of implementation in the public sector that require inherently discretionary decisions. Employ strategies that value reasoned judgment and allow meaningful space for qualitative data inputs and the practical experience of embedded individuals, treating such inputs as having value alongside more quantitative ones.
- In the longer term, develop organizational systems that foster "navigation by judgment"—for example, a management structure that delegates high levels of discretion to allow those on the ground the space to navigate complex situations, recruitment strategies that foster high numbers of staff with extensive context-specific knowledge, and systems of monitoring and learning that encourage the routine evaluation of theory against practice.

INTRODUCTION

"What gets measured gets managed, and what gets measured gets done" is one of those ubiquitous (even clichéd) management phrases that hardly require explanation; it seems immediately obvious that the data generated by regular measurement and monitoring make possible the improvement of results. Less well known than the phrase itself is the fact that, although it is commonly attributed to the acclaimed management

theorist Peter Drucker, Drucker himself never actually said it (Zak 2013). In fact, Drucker's views on the subject were reportedly far more nuanced, along the lines of those of V. F. Ridgway, who argued over 65 years ago that not everything that matters can be measured and that not everything that can be measured matters (Ridgway 1956). Simon Caulkin (2008), a contemporary business management columnist, neatly summarizes Ridgway's argument, in the process expanding the truncated "to measure is to manage" phrase to "What gets measured gets managed—even when it's pointless to measure and manage it, and even if it harms the purpose of the organisation to do so."

Ridgway's and Caulkin's warnings—repeated in various guises by many since—remind us that the indiscriminate use of quantitative measures and undue confidence in what they can tell us may be highly problematic in certain situations, sometimes derailing the very performance improvements that data are intended to support (Merry, Davis, and Kingsbury 2015).[1] We hasten to add, of course, that seeking more and better quantitative data is a worthy aim in public administration (and elsewhere). Many important gains in human welfare (for example, recognizing and responding to learning disabilities) can be directly attributed to interventions conceived of and prioritized on the basis of empirical documentation of the reality, scale, and consequences of the underlying problem. The wonders of modern insurance are possible because actuaries can quantify all manner of risks over time, space, and groups. What we will argue in the following sections, however, is that access to quantitative data alone is not a sufficient condition for achieving many of the objectives that are central to public administration and economic development.

This chapter has five sections. Following this introduction, we lay out in section two how the collection, curation, analysis, and interpretation of data are embedded in contexts: no aspect takes place on a blank slate. On one hand, the institutional embeddedness of the data collection and usage cycle—in rich and poor countries alike—leaves subsequent delivery efforts susceptible to a host of possible compromises, stemming from an organization's inability to manage and deploy data in a consistently professional manner. At the same time, the task's inherent political and social embeddedness ensures it will be susceptible to influence by existing power dynamics and the normative expectations of those leading and conducting the work, especially when the political and financial stakes are high. In contexts where much of everyday life transpires in the informal sector—rendering it "illegible" to, or enabling it to actively avoid engagement with, most standard measurement tools deployed by public administrators—sole reliance on formal quantitative measures will inherently only capture a slice of the full picture.

In section 3, we highlight four specific ways in which an indiscriminate increase in the collection of what is thought to be "good data" can lead to unintended and unwanted (potentially even harmful) consequences. The risks are that (1) the easy-to-measure can become a misleading or false measure of broader reality, (2) measurement can become conflated with what management is and does, (3) an emphasis on what is readily quantified can inhibit a fuller and more accurate understanding of the underlying policy problem(s) and their constituent elements, and (4) political pressure to manipulate selected indicators, if undetected, can lead to falsification and unwarranted expectations—or, if exposed, can compromise the perceived integrity of otherwise worthy measurement endeavors.

Thankfully, there are ways to anticipate and mitigate these risks and their unintended consequences. Having flagged how unwanted outcomes can emerge, we proceed to highlight, in section 4, some practical ways in which public administrators might thoughtfully anticipate, identify, and guard against them. We discuss what a balanced suite of data tools might look like in public administration and suggest four principles that can help us apply these tools to the greatest effect, thereby enabling the important larger purposes of data to be served. For further methodological guidance, practitioners should consult appendix A, which provides a checklist titled "Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration." We stress from the outset that our concerns are not with methodological issues per se, or with the quality or comprehensiveness of quantitative data; these concerns are addressed elsewhere in *The Government Analytics Handbook* and in every econometrics textbook, and they should always be considered as part of doing "normal social science." The concerns we articulate are salient even in a best-case scenario, in which analysts have access to great data acquired from a robust methodology, although they are obviously compounded when the available data are of poor quality—as is often the case, especially in low-income countries—and when too much is asked of them.

HOW DATA ARE IMPACTED BY THE INSTITUTIONAL AND SOCIOPOLITICAL ENVIRONMENT IN WHICH THEY ARE COLLECTED

For all administrative tasks, but especially those entailing high-stakes decision-making, the collection and use of data is a human process inherently subject to human foibles (Porter 1995). This is widely accepted and understood: for example, key conceptual constructs in development (such as "exclusion," "household," and "fairness") can mean different things to different people and translate awkwardly into different languages. With this in mind, professional data collectors will always give serious attention to "construct validity" concerns to ensure there is close alignment between the questions they ask and the questions their informants hear.[2] For present purposes, we draw attention to issues given less attention, but which are critical nonetheless—namely, the institutional and political factors that comprise the context shaping which data are (and are not) collected, how and from whom they are collected, how well they are curated over time, and how carefully conclusions and policy implications are drawn from analyses of them. We briefly address each item in turn.

INSTITUTIONAL EMBEDDEDNESS OF DATA

Beyond the purposes to which they are put, the careful collection, curation, analysis, and interpretation of public data are themselves complex technical and administrative tasks, requiring broad, deep, and sustained levels of organizational capability. In this section, we briefly explore three institutional considerations shaping these factors: the dynamics shaping the (limited) "supply" and refinement of technical skills, the forging of a professional culture that is a credible mediator of complex (and potentially heated) policy issues yet sufficiently robust to political pressure, and the related capacity to infer what even the best data analysis "means" for policy, practice, and problem solving.

These issues apply in every country but are especially salient in low-income countries, where the prevailing level of implementation capability in the public sector is likely to be low, and where the corresponding expectations of those seeking to improve it by expanding the collection and use of quantitative data may be high. At the individual level, staff with the requisite quantitative analytical skills are likely to be in short supply because acquiring such skills requires considerable training, while those who do have them are likely to be offered much higher pay in the private sector. (One could in principle outsource some data collection and analysis tasks to external consultants, but doing so would be enormously expensive and potentially compromise the integrity and privacy of unique public data.)

So understood, it would be unreasonable to expect the performance of data-centric public agencies to be superior to other service delivery agencies in the same context (for example, public health). Numerous studies suggest the prevailing levels of implementation capability in many (if not most) low-income countries are far from stellar (Andrews, Pritchett, and Woolcock 2017).[3] For example, Jerven's (2013) important work in Africa on the numerous challenges associated with maintaining the System of National Accounts—the longest-standing economic data collection task asked of all countries, from which their respective gross domestic products (GDPs) are determined—portends the difficulties facing less high-profile metrics (see also Sandefur and Glassman 2015).[4] Put differently: if many developing countries struggle to curate the single longest-standing, universally endorsed, most important measure asked of them, on what basis do we expect these countries to manage lesser, lower-stakes measures?

To be sure, building quantitative analytical skills in public agencies is highly desirable; for present purposes, our initial point is a slight variation on the old adage that the quality of outcomes derived from quantitative data is only as good as the quality of the "raw material" and the competence with which it is analyzed and interpreted.[5] Fulfilling an otherwise noble ambition to build a professional public sector whose decisions are informed by evidence requires a prior and companion effort to build the requisite

skills and sensibilities. Put differently, precisely because effective data management is itself such a complex and difficult task, in contexts where agencies struggle to implement even basic policy measures at a satisfactory level (for example, delivering mail and ensuring attendance at work), it is unlikely that, ceteris paribus, asking these agencies to also take a more "data-driven" approach will elicit substantive improvement. More and better "data" will not fix a problem if the absence of data is not itself the key problem or the "binding constraint"; the priority issue is discerning what *is* the key policy problem and its constituent elements. From this starting point, more and better data can be part of, but not a substitute for, strategies for enhancing the effectiveness of public sector agencies.

Even if both data management and broad institutional capability are functioning at high and complementary levels, there remains the structural necessity of interpreting what the data *mean*. Policy inference from even the best data and most rigorous methodology is never self-evident; it must always be undertaken in light of theory. This might sound like an abstract academic concern, but it is especially important when seeking to draw lessons from, or to make big decisions regarding the fate of, complex interventions. This is so because a defining characteristic of a complex problem is that it generates highly variable outcomes across time, space, and groups.

Promoting gender equality, for example, is a task that rarely generates rapid change: it can take a generation (or several, or centuries) for rules requiring equal participation in community meetings, or equal pay for equal work, to become the "new normal."[6] So, assessed over a five-year time frame, a "rigorous" methodology and detailed data may yield the empirical finding that a given gender empowerment project (GEP) has had "no impact"; taken at face value, this is precisely what "the data" would show and is the type of policy conclusion ("the GEP doesn't work") that would be drawn. However, interpreted in the light of a general theory of change incorporating the likely impact trajectory that GEP-type interventions follow—that is, a long period of stasis eventually leading to a gradual but sustained takeoff—a "doesn't work" conclusion would be unwarranted; five years is simply too soon to draw a firm conclusion (Woolcock 2018).[7] High-quality data and a sound methodology alone cannot solve this problem: a GEP may well be fabulous, neutral, useless, or a mixture of all three, but discerning which of these it is—and why, where, and for whom it functions in the way it does—will require the incorporation of different kinds of data into a close dialogue with a practical theory of change fitted for the sector, the context, and the development problem being addressed.

SOCIOPOLITICAL EMBEDDEDNESS OF DATA

Beyond these institutional concerns, a second important form of embeddedness shaping data collection, curation, and interpretation is the manner in which all three are shaped by sociopolitical processes and imperatives. All data are compiled for a purpose; in public administration, the scale and sophistication of the required data are costly and complex (requiring significant financial outlay and, thus, competition with rival claimants). Data are frequently called upon to adjudicate both the merits of policy proposals ex ante (for example, the Congressional Budget Office in the United States) and the effectiveness of programmatic achievements ex post (for example, the World Bank's Independent Evaluation Group), which frequently entails entering into high-stakes political gambits—for example, achieving signature campaign proposals in the early days of an administration and proclaiming their subsequent widespread success (or failure) as election time beckons again. (See more on this below.)

Beyond the intense political pressure "data" are asked to bear in such situations, a broader institutional consideration is the role large-scale numerical information plays in "rendering legible" (Scott 1998) complex and inherently heterogeneous realities, such that they can be managed, mapped, and manipulated for explicit policy purposes. We hasten to add that such "thin simplifications" (Scott's term) of reality can be both benign and widely beneficial: comprehensive health insurance programs and pension systems have largely tamed the otherwise debilitating historical risks of, respectively, disease and old age by generating premiums based on

general demographic characteristics and the likelihood of experiencing different kinds of risks (for example, injuries or cancer) over the course of one's life.

A less happy aspect of apprehending deep contextual variation via simplified (often categorical) data, however, is the corresponding shift it can generate in the political status and salience of social groups. The deployment of the census in colonial India, for example, is one graphic demonstration of how the very act of "counting" certain social characteristics—such as the incidence of caste, ethnicity, and religion—can end up changing these characteristics themselves, rendering what had heretofore been relatively fluid and continuous categories as fixed and discrete. In the case of India, this massive exercise in data collection on identity led to "caste" being created, targeted, and mobilized as a politically salient characteristic that had (and continues to have) deep repercussions (for example, at independence, when Pakistan split from India, and more recently within the rise of Hindu nationalism; see Dirks 2011).[8] Similarly, influential scholars have argued that the infamous Hutu/Tutsi massacre in Rwanda was possible at the scale at which it was enacted because ethnic categories were formalized and fixed via public documents whose origins lie in colonial rule (for example, Mamdani 2002).

For Scott (1998), public administration can only function to the extent its measurement tools successfully turn widespread anthropological variation, such as in languages spoken, into singular modern categories and policy responses—for instance, to ensure that education is conducted in one national language, in a school, and on the basis of a single curriculum.[9] The net welfare gains to society might be unambiguous, but poorer, isolated, marginalized, and less numerous groups are likely to bear disproportionately the costs of this trade-off. If official "data" themselves constitute an alien or distrusted medium by which certain citizens are asked to discern the performance of public agencies, merely providing (or requiring) "more" is unlikely to bring about positive change. In such circumstances, much antecedent work may need to be undertaken to earn the trust of citizens and to help them more confidently engage with their administrative systems.[10] By way of reciprocity, it may also require such systems to interact with citizens themselves in ways that more readily comport with citizens' own everyday (but probably rather different) vernacular for apprehending the world and interpreting and responding to events. Either way, it is critical that officials be wary of the potentially negative or unintended effects of data collection, even when it may begin with a benign intention to facilitate social inclusion and more equitable policy "targeting."[11]

THE UNINTENDED CONSEQUENCES OF AN INDISCRIMINATE PURSUIT OF "MORE DATA"

There is a sense in which it is axiomatic that more and better data are always a good thing. But the institutional and sociopolitical embeddedness of data generation and the use of data in public administration (as discussed in the preceding section) means we need to qualify this assertion by focusing on where and how challenges arise. With this in mind, we turn our attention to instances where the increased collection of what is thought to be "good data" leads to perverse outcomes. Here, we highlight four such outcomes that may materialize as the result of an undue focus on issues, concepts, inputs, or outcomes that happen to be most amenable to being quantified.

Outcome 1: The Easy-to-Measure Becomes a False Standard of Success

What may start as a well-intentioned managerial effort to better quantify meaningful success can instead generate a blinkered emphasis on that which is simply easiest to quantify. The result can be a skewed or false sense of what a project has (or has not) achieved, and how, where, and for whom outcomes have been achieved.

In a recent study, we demonstrate how a variety of institutional incentives align across the government of Malawi and the World Bank in such a way that both government and World Bank officials consistently favor

easy-to-measure indicators (inputs and outputs, or what we refer to as "changes in form rather than function") as the yardstick of project success (Bridges and Woolcock 2017). This is a quintessential example of what strategy writer Igor Ansoff describes as a situation in which "managers start off trying to manage what they want, and finish up wanting what they can measure" (quoted in Cahill 2017, 152). As a result of evaluating public financial management (PFM) projects that were implemented over the course of 20 years in Malawi, we show that almost 70 percent of what projects measure or aim for is "change in terms of whether institutions look like their functioning counterparts (that is, have the requisite structures, policies, systems, and laws in place)," whereas only 30 percent of what is measured can be said to be "functional"—that is, focused on "purposeful changes to budget institutions aimed at improving their quality and outcomes" (Andrews 2013, 7). What is more, we find that World Bank PFM projects have considerably more success in achieving "formal" results than "functional" ones. Unsurprisingly, demonstrable improvement in actual performance is far harder to achieve than change that is primarily regulative, procedural, or systems oriented. Unfortunately, an emphasis on what is easy-to-measure obfuscates this reality and allows reform "success" to be claimed.

In practice, Malawi's history of PFM reform is littered with projects that claim "success" based on hardware procured, software installed, legislation developed, and people trained, whereas even a basic analysis reveals stagnation or even regression in terms of more affordable spending decisions, spending that reflects budgeted promises, greater ability to track the flow of funds, or reduction in corruption. As long as the World Bank and the Malawian government focus on "formal" measures, they are able to maintain the illusion of success: that is, until something like Malawi's 2013 "Cashgate" crisis—in which it was revealed that about US$32 million in government funds had been misappropriated between April and September 2013—lifts the lid on the deep-rooted financial management problems that have remained largely unaffected by millions of dollars of reform efforts. In this sense, Malawi is a microcosm of many institutional reform efforts globally. Although similar financial reforms have been globally implemented in a manner that suggests some level of consensus about "what works," the outcomes of those reforms are varied at best and often considerably lower than anticipated (Andrews 2013).

In the same way that an emphasis on the easy-to-measure can lead to overestimation of success, it can also contribute to underestimation. Reforms can sometimes yield meaningful change via what McDonnell (2017) calls "the animating spirit of daily practice" but end up being missed because managers do not have good means of measuring, attributing, and enhancing these *kinds* of shifts. For example, when researching the impact of technical assistance on a large government health program in Nigeria, we found that there were strong indications that important innovations and shifts took place at the local level, including in aspects as difficult to shift as cultural practices regarding contraceptives (Bridges and Woolcock 2019). These shifts in practice and their impact on contraceptive uptake could not be apprehended by aggregated statewide indicators, however, and since no measurement was being done below this level, the progress and valuable lessons of such interventions were being missed.

Another example of the importance of having access to a broader suite of data comes from an assessment of a program in rural India seeking to promote participatory democracy in poor communities, where the curation of such a data suite enabled more nuanced and constructive lessons to be drawn (see Rao, Ananthpur, and Malik 2017). The results of the initial randomized controlled trial (RCT) deemed the program to have had no mean impact—and if these were the only data available, that would have been the sole conclusion reached.[12] Upon closer inspection, however, it was learned that there was considerable variation in the program's impact. The average of this variation may have been close to zero, but for certain groups, the program had worked quite well, for others it had had no impact, while for still others it had been detrimental. Who were these different groups, and what was it about them that led to such variable outcomes? A companion qualitative process evaluation was able to discern that the key differences were the quality of implementation received by different groups, the level of support provided to them by managers and political leaders, and variations in the nature and extent of local-level inequalities (which, in turn, shaped which groups were able to participate and on what terms).[13] The administrative rules and implementation guidelines provided to all groups were identical, but in this case, a qualitative process evaluation was able to document the ways

and places in which variable fidelity to them yielded widely different outcomes (albeit with no net impact). Moreover, the qualitative data were able to discern subtle positive effects from the program that the quantitative survey instrument alone would have missed.

Outcome 2: Measurement Becomes Conflated with Management

An extension of the above point is that an undue emphasis on quantitative data can lead measurement to become a *substitute for* rather than a *complement to* management. This is evident when only that which is quantifiable receives any significant form of managerial attention, an outcome made possible when the easily quantifiable becomes the measure of success, becoming, in turn, the object of management's focus, typically to the exclusion of all else. As Wilson (1989, 161) famously intoned in a classic study of bureaucratic life, "Work that produces measurable outcomes tends to drive out work that produces immeasurable outcomes."

In one sense this is hardly surprising: the need for managers to make decisions on the basis of partial information is difficult and feels risky, so anything that claims to fill that gap and bypass the perceived uncertainty of subjective judgment will be readily welcomed. "The result," Simon Caulkin (2016) argues, "both practically and theoretically, is to turn today's management into a technology of control that attempts to minimise rather than capitalise on the pesky human element." And in public administration, a managing-it-by-measuring-it bias can mean that, over time, the bulk of organizational resources end up neglecting the "pesky human element" of change processes, even though it is this element that is often central to attaining the transformational outcomes managers are seeking.

This dynamic characterizes key aspects of the Saving One Million Lives (SOML) initiative, an ambitious health sector reform program launched by the government of Nigeria. The original goal of SOML was to save the lives of one million mothers and children by 2015; to this end, SOML gave priority to a package of health interventions known as "the six pillars."[14] The World Bank actively supported SOML, using its Program-for-Results (PforR) instrument to reward Nigerian states financially based on improvements from their previous best performance on the six key indicators.[15] Improvements were to be measured through yearly household surveys providing robust estimates at the state level.

In practice, of course, these six pillars (or intervention areas) were wildly different in their drivers and complexity; improvement within them was therefore destined to move at different trajectories and different speeds for different groups in different places. State actors, keen to raise their aggregate measure of success and get paid for it, soon realized that there was some gaming to be done. Our field research documents how the emphasis on singular measures of success introduced a perverse incentive for states to focus on the easier metrics at the expense of the harder ones (Bridges and Woolcock 2019). Interviews with state officials revealed that frontline staff increasingly focused their time and energies on those constituent variables that they discerned were easiest to accomplish (for example, dispensing vitamin supplements) over those that were harder or slower—typically those that involved a plethora of "pesky human elements," such as lowering maternal mortality or increasing contraceptive use. In selecting certain outcomes for measurement and managing these alone, others inevitably end up being sidelined.

Likewise, a recent report on results-based financing (RBF) in the education sector (Dom et al. 2020) finds evidence of a "diversion risk" associated with the signposting effect of certain reward indicators, with important areas deprioritized because of the RBF incentive. For example, in Mozambique, they find that an emphasis on simple process indicators and a focus on targets appears to have led officials to divert resources and attention away from "more fundamental and complex issues," such as power dynamics in the school council, the political appointment of school directors, and teachers' use of training. Dom et al. also report evidence of "cherry-picking risks," in which less costly or politically favored subgroups or regions see greater resources, in part because they are more likely to reach a target. For example, in Tanzania, they find evidence that the implementation of school rankings based on exam results was correlated with weaker students not sitting, presumably in an effort by the schools to raise average exam pass rates.

This tendency becomes a particular issue when the sidelined outcomes end up being the ones we care most about. Andrew Natsios (2011), the former administrator of the United States Agency for International

Development (USAID, an organization charged with "demonstrating the impact of every aid cent that Congress approves"), argues compellingly that the tendency in aid and development toward what he calls "obsessive measurement disorder" is a manifestation of a core dictum among field-based development practitioners—namely, "that those development programs that are most precisely and easily measured are the least transformational, and those programs that are most transformational are the least measurable." The change we often desire most is in very difficult-to-measure aspects, such as people's habits, cultural norms, leadership characteristics, and mindsets.

This reality is also aptly illustrated in many anticorruption efforts, where imported solutions have managed to change the easy-to-measure—new legislation approved, more cases brought, new financial systems installed, more training sessions held—but have failed to shift cultural norms regarding the unacceptability of whistleblowing or the social pressures for nepotism (Andrews 2013). Failure to measure and therefore manage these informal drivers of the problem ensures that any apparent reduction in fund abuses tends to be short-lived or illusory.

This phenomenon is hardly limited to poor countries. A more brutal example of how what cannot be measured does not get managed, with disastrous results, can be found in the United Kingdom's National Health System (NHS). While investigating the effects of competition in the NHS, Propper, Burgess, and Gossage (2008) discovered that the introduction of interhospital competition improved waiting times while also substantially *increasing* the death rate following emergency heart attacks. The reason for this was that waiting times were being measured (and therefore managed), while emergency heart-attack deaths were not tracked and were thus neglected by management. The result was shorter waiting times but more deaths as a result of the choice of measure. The authors note that the issue here was not intent but the extent to which one target consumed managerial attention to the detriment of all else; as they note, it "seems unlikely that hospitals deliberately set out to decrease survival rates. What is more likely is that in response to competitive pressures on costs, hospitals cut services that affected [heart-attack] mortality rates, which were unobserved, in order to increase other activities which buyers could better observe" (Propper, Burgess, and Gossage 2008).

More recently, in October 2019, the Global Health Security Index sought to assess which countries were "most prepared" for a pandemic, using a model that gave the highest ranking to the United States and the United Kingdom, largely on the basis of these countries' venerable medical expertise and technical infrastructure, factors which are readily measurable (McCarthy 2019). Alas, the model did not fare so well when an actual pandemic arrived soon thereafter: a subsequent analysis, published in *The Lancet* on the basis of pandemic data from 177 countries between January 2020 and September 2022, found that "pandemic-preparedness indices . . . were not meaningfully associated with standardised infection rates or IFRs [infection/fatality ratios]. Measures of trust in the government and interpersonal trust, as well as less government corruption, had larger, statistically significant associations with lower standardised infection rates" (Bollyky et al. 2022, 1).

Needless to say, variables such as "trust" and "government corruption" are hard to measure, are hard to incorporate into a single theory anticipating or informing a response to a pandemic, and map awkwardly onto any corresponding policy instrument. For present purposes, the inference we draw from these findings is not that global indexes have no place; rather, they suggest the need, from the outset, for curating a broad suite of data when anticipating and responding to complex policy challenges, the better to promote real-time learning. Doubling down on what can be readily measured limits the space for eliciting those "unobserved" (and perhaps unobservable) factors that may turn out to be deeply consequential.

Outcome 3: An Undue Emphasis on Data Inhibits Understanding of the Foundational Problem(s)

An indiscriminate emphasis on aggregated, quantitative data can erode important nuances about the root causes of the problems we want to fix, thereby hampering our ability to craft appropriate solutions and

undermining the longer-term problem-solving capabilities of an organization. All too often, the designation of indicators and targets has the effect of causing people to become highly simplistic about the problems they are trying to address. In such circumstances, what should be organizational meetings held to promote learning and reflection on what is working and what is not instead become efforts in accounting and compliance (Honig and Pritchett 2019). Reporting, rather than learning, is incentivized, and management increasingly focuses on meeting target numbers rather than solving the problem. Our concern here is that, over time, this tendency progressively erodes an organization's problem-solving capabilities.

The education sector is perhaps the best illustration of this: time and again practitioners have sought to codify "learning," and time and again this has resulted in an obfuscation of the actual causes underlying the problem. In a well-intentioned effort to raise academic performance, "the standards movement" in education promoted efforts hinged on quantitative measurement, as reported in the league tables of the Program for International Student Assessment (PISA).[16] PISA runs tests in mathematics, reading, and science every three years with groups of 15-year-olds in countries around the world. Testing on such a scale requires a level of simplicity and "standardization," thus the emphasis is on written examinations and the extensive use of multiple-choice tests so that students' answers can be easily codified and processed (Robinson and Aronica 2015). Demonstrating competence in fundamental learning tasks certainly has its place, but critics have increasingly argued that such tests are based on the incorrect assumption that what drives successful career and life outcomes is the kind of learning that is capable of being codified via a standardized test (Claxton and Lucas 2015; Khan 2012).

In reality, the gap between the skills that children learn and are tested for and the skills that they need to excel in the 21st century is becoming more obvious. The World Economic Forum noted in 2016 that the traditional learning captured by standardized tests falls short of equipping students with the knowledge they need to thrive.[17] Yong Zhao (2012), the presidential chair and director of the Institute for Global and Online Education in the College of Education at the University of Oregon, points out that there is an inverse relationship between those countries that excel in PISA tests and those that excel in aspects like entrepreneurism, for example (see figure 4.1).

While a focus on assessing learning is laudable—and a vast improvement over past practices (for example, in the Millennium Development Goals) of merely measuring attendance (World Bank 2018)—for present purposes the issue is that the drivers of learning outcomes are far more complex than a quantifiable content deficit in a set of subjects. This is increasingly the case in the 21st century, which has brought a need for new skills and mindsets that go well beyond the foundational numeracy and literacy skills required during the Industrial Revolution (Robinson and Aronica 2015). A survey of chief human resources and strategy officers by the World Economic Forum (2016) finds a significant shift between 2015 and 2020 in the top skills future workers will need, with "habits of mind" like critical thinking, creativity, emotional intelligence, and problem solving ranking well ahead of any specific content acquisition. None of this is to say that data do not have a role to play in measuring the success of an educational endeavor. Rather, the data task in this case needs to be informed by the complexity of the problem and the extent to which holistic learning resists easy quantification.[18]

Finally, relying exclusively on high-level aggregate data can result in presuming uniformity in underlying problems and thus lead to the promotion of simplistic and correspondingly generic solutions. McDonnell (2020) notes, for example, that because many developing countries have relatively high corruption scores, an unwelcome outcome has been that *all* the institutions in these countries tend to be regarded by would-be reformers as similarly corrupt and uniformly ineffectual. In her impressive research on "clusters of effectiveness," however, she offers evidence of the variation in public-sector performance within states, noting how the aggregated data on "corruption" masks the fact that the difference in corruption scores between Ghana's best- and worst-rated state agencies approximates the difference between Belgium (WGI = 1.50) and Mozambique (WGI = −0.396), in effect "spanning the chasm of so-called developed and developing worlds." The tendency of reform actors to be guided by simplistic aggregate indicators—such as those that are used to determine a poor country's "fragility" status and eligibility for International Development Association (IDA) funding—has prevented a more

FIGURE 4.1 Country Scores on Program for International Student Assessment Tests and Perceived Entrepreneurial Capability

Source: Based on Zhao 2012.
Note: PISA = Program for International Student Assessment.

detailed and context-specific understanding of the lessons that could be drawn from positive outlier cases, or what McDonnell refers to as "the thousand small revolutions quietly blooming in rugged and unruly meadows."[19]

Outcome 4: Pressure to Manipulate Key Indicators Leads to Falsification and Unwarranted Impact Claims

As an extension of our previous point regarding how the easy-to-measure can become the yardstick for success, it is important to acknowledge that public officials are often under extreme pressure to demonstrate success in selected indicators. Once data themselves, rather than the more complex underlying reality, become the primary objective by which governments publicly assess (and manage) their "progress," it is inevitable that vast political pressure will be placed on these numbers to bring them into alignment with expectations, imperatives, and interests. Similar logic can be expected at lower units of analysis (for example, field offices), where it tends to be even more straightforward to manipulate data entry and analysis. This, in turn, contributes to a perverse incentive to falsify or skew data, to aggregate numbers across wildly different variables into single indexes, and to draw unwarranted inferences from them.

This risk is particularly acute, for instance, when annual global rankings are publicly released (assessing, for example, a country's "investment climate," "governance," and gender equity), thereby shaping major investment decisions, credit ratings, eligibility for funding from international agencies, and the fate of senior officials charged with "improving" their country's place in these global league tables. Readers will surely be aware of the case at the World Bank in September 2021, when an external review revealed that the *Doing Business* indicators had been subject to such pressure, with alterations made to certain indicators from certain countries (WilmerHale 2021). Such rankings are now

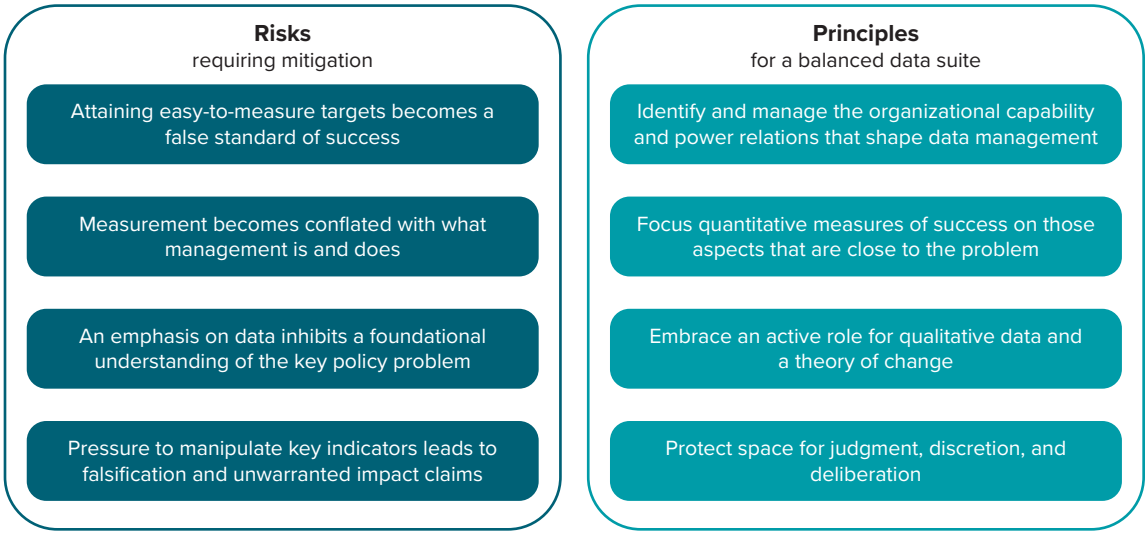
omnipresent, and if they are not done by one organization, they will inevitably be done by another. Even so, as the *Economist* (2021) magazine concluded, some might regard the *Doing Business* episode as "proof of 'Goodhart's law,' which states that when a measure becomes a target, it ceases to be a good measure." At the same time, it pointed out that there is a delicate dance to be done here, since "the Doing Business rankings were always intended to motivate as well as measure, to change the world, not merely describe it," and "if these rankings had never captured the imagination of world leaders, if they had remained an obscure technical exercise, they might have been better as measures of red tape. But they would have been worse at cutting it."

Such are the wrenching trade-offs at stake in such exercises, and astute public administrators need to engage in them with their eyes wide open. Even (or especially) at lower units of analysis, where there are perhaps fewer prying eyes or quality-control checks, the potential is rife for undue influence to be exerted on data used for political and budgetary-allocation purposes. Fully protecting the integrity of data collection, collation, and curation (in all its forms) should be a first-order priority, but so, too, is the need for deploying what should be standard "risk diversification" strategies on the part of managers—namely, not relying on single numbers or methods to assess inherently complex realities.

PRINCIPLES FOR AN EXPANSIVE, QUALIFIED DATA SUITE THAT FOSTERS PROBLEM SOLVING AND ORGANIZATIONAL LEARNING

In response to the four risks identified above, we offer a corresponding set of cross-cutting principles for addressing them. Figure 4.2 summarizes the four risks in the left-hand column and presents the principles as vertical text on the right, illustrating the extent to which the principles, when applied in combination, can serve to produce a more balanced data suite that prioritizes problem solving and learning.

FIGURE 4.2 Four Risks with Corresponding Principles for Mitigating Them to Ensure a Balanced Data Suite



Source: Original figure for this publication.
Note: The principles are "cross-cutting," in the sense that they apply in some measure to all the risks; they are not one-to-one.

Principle 1: Identify and Manage the Organizational Capacity and Power Relations That Shape Data Management

The data collection and curation process takes place not in isolation but in a densely populated political and institutional ecosystem. It is difficult, expensive, and fraught work; building a professional team capable of reliably and consistently doing this work—from field-level collection and curation at headquarters to technical analysis and policy interpretation—will be as challenging as it is in every other public sector organization. Breakdowns can happen at any point, potentially compromising the integrity of the entire endeavor. For this reason, it is important for managers not just to hire those with the requisite skills but to cultivate, recognize, and reward a professional ethos wherein staff members can do their work in good faith, shielded from political pressure. Such practices, in turn, need to be protected by clear, open, and safe procedures for staff to report undue pressure, complemented by accountability to oversight or advisory boards comprising external members selected for their technical expertise and professional integrity. In the absence of such mechanisms, noble aspirations for pursuing an "evidence-based policy" agenda risk being perceived as means of providing merely "policy-based evidence."

The contexts within or from which data are collected are also likely to be infused with their own sociopolitical characteristics. Collecting data on the incidence of crime and violence, for example, requires police to faithfully record such matters and their response to them, but they must do so in an environment where there may be strong pressure to underreport, whether because of personal safety concerns, lack of administrative resources, or pressure to show that a given unit's performance is improving (where this is measured by showing a "lower incidence" of crime). In this respect, good diagnostic work will reveal the contours of the institutional and political ecosystem wherein the data work will be conducted and the necessary authorization, financing, and protection sought; it will also help managers learn how to understand and successfully navigate this space.[20] The inherent challenges of engaging with such issues might be eased somewhat if those closest to them see data deployment not as an end in itself or an instrument of compliance but as a means to higher ends—namely, learning, practical problem solving, and enhancing the quality of policy options, choices, and implementation capability.[21]

A related point is that corresponding efforts need to be made to clearly and accurately communicate to the general public those findings that are derived from data managed by public administrators, especially when these findings are contentious or speak to inherently complex issues. This has been readily apparent during the COVID-19 pandemic, with high-stakes policy decisions (for example, requiring vulnerable populations to forgo income) needing to be made on the basis of limited but evolving evidence. Countries such as Vietnam have been praised for the clear and consistent manner in which they issued COVID-19 response guidelines to citizens (Ravallion 2020), but the broader point is that even when the most supported decisions are based on the best evidence generated by the most committed work environments, it remains important for administrators to appreciate that the very act of large-scale measurement and empirical interpretation, especially when enacted by large public organizations, can potentially be threatening to or misunderstood by the very populations they are seeking to assist.

Principle 2: Focus Quantitative Measures of Success on Those Aspects That Are Close to the Problem

If we wish to guard against the tendency to falsely ascribe success based on the achievement of poorly selected indicators, then we should ensure that any indicators used to claim or deny reform success are as readily operational and close to the service delivery problem as possible. Output and process indicators are useful in their own ways, but we should not make the mistake of conflating their achievement with "problem fixed." There are often strong institutional incentives to claim reform success based on whether a new mechanism or oversight structure has been created, a new law has been passed, or a percentage of participation has been achieved, but if meaningful change is sought, these incentives need to be countered. All of these

measures are changes in *form* that, while useful as indicators of outputs being met, can be achieved (and have been in the past) without any attendant functional shifts in the underlying quality of service delivery.

Officials can guard against this tendency by taking time to ensure that an intervention is focused on specific problems, including those that matter at a local level, and that the intervention's success and its attendant metrics are accurate measures of the problems being fixed. Tools like the Problem-Driven Iterative Adaptation (PDIA) Toolkit, designed by members of the Building State Capability program at the Center for International Development (CID) at Harvard University, can help guide practitioners in this process.[22] The PDIA approach is designed to help practitioners break down their problems into root causes, identify entry points, search for possible solutions, take action, reflect upon what they have learned, adapt, and then act again. By embedding any intervention in such a framework, practitioners can ensure that success metrics are well linked to functional responses to locally felt problems.

Whatever tool is applied, the goal should be to arrive at metrics of success that represent a compelling picture of the root performance problem being addressed (and hopefully solved). So, for example, in our education example, metrics such as the number of teachers hired, the percentage of the budget dedicated to education, and the number of schools built are all output measures that say nothing about actual learning. Of course, there are assumptions that these outputs *lead* to children's learning, but as many recent studies now show, such assumptions are routinely mistaken; these indicators can be achieved even as actual learning regresses (Pritchett 2013; World Bank 2018). By contrast, when a robust measure of learning—in this case, literacy acquisition—was applied in India, it allowed implementers to gain valuable insights about which interventions actually made a difference, revealing that teaching to a child's actual level of learning, not their age or grade, led to marked and sustained improvements. Crucially, such outcomes are the result of carefully integrated qualitative and quantitative approaches to measurement (Banerjee et al. 2016).

Going further, various cross-national assessments around the world are trying to tackle the complex challenge of finding indicators that measure learning not just in the acquisition of numeracy, science, and literacy skills but in competencies that are increasingly valuable in the 21st century: grit, curiosity, communication, leadership, and compassion. PISA, for example, has included an "innovative domain" in each of its recent rounds, including creative problem solving in 2012, collaborative problem solving in 2015, and global competence in 2018. In Latin America, the Latin American Laboratory for Assessment of the Quality of Education (LLECE) included a module on socioemotional skills for the first time in its assessment of sixth-grade students in 2019, focusing on the concepts of conscience, valuing others, self-regulation, and self-management (Global Partnership for Education 2020). Much tinkering remains to be done, but the increase in assessments that include skills and competencies such as citizenship (local and global), socioemotional skills, information and communication technology literacy, and problem solving is a clear indication of willingness to have functional measures of success, capturing outcomes that matter.

In summary, then, public administrators who wish to guard against unwarranted impact claims and ensure metrics of success are credible can begin by making sure that an intervention itself is focused on a specific performance problem that is locally prioritized and thereafter ensure that any judgment of that intervention's success or failure is based not on output or process metrics but on measures of the problems being fixed. And having ensured that measures of success are functional, practitioners must allow for flexibility of implementation where possible so strategies can shift if it becomes clear from the collected data that they are not making progress toward fixing the problem, possibly due to mistaken assumptions regarding their theory of change.

Principle 3: Embrace an Active Role for Qualitative Data and a Theory of Change

The issues we have raised thus far, we argue, imply that public administrators should adopt a far more expansive concept of what constitutes "good data"—namely, one that includes insights from theory and qualitative research. Apprehending complex problems requires different forms and sources of data; correctly interpreting empirical findings requires active dialogue with reasoned expectations about what outcomes should be attained by when. Doing so helps avoid creating distortions that can generate (potentially wildly) misleading claims regarding "what's going on and why" and "what should be done."

Specifically, we advocate for the adoption of a complementary *suite* of data forms and sources that favors flexibility, is focused on problem solving (as opposed to being an end in itself), and values insights derived from seasoned experience. In the examples we have explored above, reliance on a single form of data (sometimes even a single number) rendered projects vulnerable to political manipulation, unwarranted conclusions, and an inability to bear the decision-making burdens thrust upon them. More constructively, it was the incorporation of alternative methods and data in dialogue with a reasoned theory of change that enabled decision-makers to anticipate and address many of these same concerns.

To this end, we have sought to get beyond the familiar presumption that the primary role of qualitative data and methods in public administration research (and elsewhere) is to provide distinctive insights into the idiosyncrasies of an organization's "context" and "culture" (and thus infuse some "color" and "anecdotes" for accompanying boxes).[23] Qualitative approaches can potentially yield unique and useful material that contributes to claims about *whether* policy goals are being met and delivery processes duly upheld (Cartwright 2017); they can be especially helpful when the realization of policy goals requires integrating both adaptive and technical approaches to implementation—for example, responding to COVID-19. But perhaps the more salient contributions of qualitative approaches, we suggest, are to explore *how*, *for whom*, and *from whom* data of all kinds are deployed as part of broader imperatives to meet political requirements and administrative logics in a professional manner and to elicit novel or previously "unobserved" variables shaping policy outcomes.

Principle 4: Protect Space for Judgment, Discretion, and Deliberation

Our caution is against using data reductively: as a replacement or substitute for managing. Management *must* be about more than measuring. A good manager needs to be able to accommodate the immeasurable because so much that is important to human thriving is in this category; dashboards etc. certainly have their place, but if these were all that was needed, then "managing" could be conducted by machines. We all know from personal experience that the best managers and leaders take a holistic interest in their staff, taking the time and making the effort to understand the subtle, often intangible processes that connect their respective talents. As organizational management theorist Henry Mintzberg (2015) wisely puts it,

> Measuring as a complement to managing is a fine idea: measure what you can; take seriously what you can't; and manage both thoughtfully. In other words: If you can't measure it, you'll have to manage it. If you can measure it, you'll especially have to manage it. Have we not had enough of leadership by remote control: sitting in executive offices and running the numbers—all that deeming and downsizing?[24]

Contrary to the "what can't be measured can't be managed" idea, we *can* manage the less measurable if we embrace a wider set of tools and leave space for judgment. The key for practitioners is to begin with a recognition that measurability is not an indicator of significance and that professional management involves far more than simply "running the numbers," as Mintzberg puts it. Perhaps the most compelling empirical case for the importance of "navigating by judgment" in public administration has been made by Honig (2018), in which he shows—using a mix of quantitative data and case study analysis—that the more complex the policy intervention, the more necessary it becomes to grant discretionary space to frontline managers, and the more necessary such discretion is to achieving project success. Having ready access to relevant, high-quality quantitative data can aid in this "navigation," but true navigation requires access to a broader suite of empirical inputs.

In a similar vein, Ladner (2015, 3) points out that "standard performance monitoring tools are not suitable for highly flexible, entrepreneurial programs as they assume that how a program will be implemented follows its original design." To avoid "locking in" a theory of change that prevents exploration or responsive adaptation, some practitioners have provided helpful suggestions for how to use various planning frameworks in ways that support program learning.[25] The Building State Capability team highlights lighter-touch methods, such as their PDIA "check-ins," which include a series of probing questions

to assist teams in capturing learning and maximizing adaptation. Teskey and Tyrrel (2017) recommend participating in regularized formal and informal Review and Reflection (R&R) points, during which a contractor can demonstrate how politics, interests, incentives, and institutions were systematically considered in problem selection and design and, in turn, justify why certain choices were made to stop, drop, halt, or expand any activity or budget during implementation. The common connection across all these tools is that they seek to carve out meaningful space for qualitative data and the hard-won insights born out of practical experience.

In summary, then, public administrators can embed the recognition that management must be about more than measuring by first recognizing that whatever they choose to measure will inevitably create incentives to neglect processes and outcomes that cannot be measured (or are hard to measure) but are nonetheless crucial for discerning whether, how, where, and for whom policies are working. Following that recognition, they need to be very careful about what they choose to measure. Second, they can actively identify what they cannot (readily) measure that matters and take it seriously, developing strategies to manage that as well. A key part of those strategies will be that they create space for judgment, qualitative data inputs, and the practical experience of embedded individuals (focus group discussions, case studies, semi-structured interviews, review and reflection points, etc.) and treat these inputs as equally valid alongside more quantitative ones. As far as longer-term strategies to manage the immeasurable, administrations can work toward developing organizational systems that foster navigation. Such systems might include, for example, a management structure that delegates high levels of discretion to allow those on the ground the ability to navigate complex situations, recruitment strategies that foster high numbers of staff with extensive context-specific knowledge, and systems of monitoring and learning that encourage the routine evaluation of theory with practice.

CONCLUSION

Quantitative measurement in public administration is undoubtedly a critical arrow in the quiver of any attempt to improve the delivery of public services. And yet, since not everything that matters can be measured and not everything that can be measured matters, a managerial emphasis on measurement alone can quickly and inadvertently generate unwanted outcomes and unwarranted conclusions. In the everyday practices of public administration, effective and professional action requires forging greater complementarity between different epistemological approaches to collecting, curating, analyzing, and interpreting data. We fully recognize that this is easier said than done. The risks of reductive approaches to measurement are not unknown, and yet simplified appeals to "what gets measured gets managed" persist because they offer managers a form of escape from those "pesky human elements" that are difficult to understand and even more so to shift.

Most public administrators might agree in principle that a more balanced data suite is necessary to navigate their professional terrain, yet such aspirations are too often honored in the breach: under sufficient pressure to "deliver results," staff from the top to the bottom of an organization are readily tempted to reverse engineer their behavior in accordance with what "the data" say (or can be made to say). Management as measurement is tempting for individuals and organizations that fear the vulnerability of their domain to unfavorable comparison with other (more readily measurable and "legible") domains, as well as the complexity of problem solving and the necessity of subjective navigation that it often entails. But given how heavily institutional and sociopolitical factors shape how data are collected, how well they are collected and curated, and how they can be manipulated for unwarranted purposes, a simplistic approach to data as an easy fix is virtually guaranteed to obscure learning and hamper change efforts. If administrations genuinely wish to build their problem-solving capabilities, then access to more and better quantitative data will be necessary, but it will not be sufficient.

Beginning with an appreciation that much of what matters cannot be (formally) measured, public administration must routinely remind itself that promoting and accessing data is not an end in itself: data's primary purpose is not just monitoring processes, compliance, and outcomes, but contributing to problem solving and organizational learning. More and better data will not fix a problem if the absence of such data is not itself the key problem or the "binding constraint." Administrations that are committed to problem solving, therefore, will need to embed their measurement task in a broader problem-driven framework, integrate complementary qualitative data, and value embedded experience in order to apprehend and interpret complex realities more accurately. Their priority in undertaking good diagnostic work should be to identify and deconstruct key problems, using varied sources of data, and then to track and learn from potential solutions authorized and enacted in response to the diagnosis. Accurate inferences for policy and practice are not derived from data alone; close interaction is required between data (in various forms), theory, and experience. In doing all this, public administrators will help mitigate the distortionary (and ultimately self-defeating) effects of managing only that which is measured.

NOTES

Our thanks to Galileu Kim, Daniel Rogger, Christian Schuster, and participants at an authors' workshop for helpful comments and constructive suggestions. More than 20 years of collaboration with Vijayendra Rao have also deeply shaped the views expressed herein. Remaining errors of fact or interpretation are solely ours.

1. See, for example, former USAID Administrator Andrew Natsios (2011), citing Lord Wellington in 1812 on the insidious manner in which measures of "accountability" can compromise rather than enable central policy objectives (in Wellington's case, winning a war). For his part, Stiglitz has argued that "what you measure affects what you do. If you don't measure the right thing, you don't do the right thing" (quoted in Goodman 2009). Pritchett (2014), exemplifying this point, notes (at least at the time of his writing) that the Indian state of Tamil Nadu had 817 indicators for measuring the delivery of public education but none that actually assessed whether students were learning. In this instance, an abundance of "measurement" and "data" was entirely disconnected from (what should have been) the policy's central objective. In many cases, however, it is not always obvious, especially ex ante, what constitutes the "right thing" to measure—hence the need for alternative methodological entry points to elicit what this might be.
2. Social science methodology courses classically distinguish between four key issues that are at the heart of efforts to make empirical claims in applied research: *construct validity* (the extent to which any concept, such as "corruption" or "poverty," matches particular indicators), *internal validity* (the extent to which causal claims have controlled for potential confounding factors, such as sample selection bias), *external validity* (the likelihood that claims are generalizable at larger scales and to more diverse populations or novel contexts), and *reliability* (the extent to which similar findings would be reported if repeated or replicated by others). See, among many others, Johnson, Reynolds, and Mycoff (2019). Of these four issues, qualitative methods are especially helpful in ensuring construct validity, since certain terms may mean different things to different people in different places, complicating matters if one seeks to draw comparisons across different linguistic, cultural, or national contexts. In survey research, for example, it is increasingly common to include what is called an "anchoring vignette"—a short, real-world example of the phenomenon in question, such as an instance of corruption by a government official at a port—before asking the formal survey question so that cross-context variations in interpretation can be calibrated accordingly (see, among others, King and Wand 2007). Qualitative methods can also contribute to considerations pertaining to internal validity (Cartwright 2017) and external validity—helping to identify the conditions under which findings "there" might apply "here" (Woolcock 2018; see also Cartwright and Hardie 2012).
3. If such agencies or departments do in fact happen to perform especially strongly—in the spirit of the "positive deviance" cases of government performance in Ghana provided in McDonnell (2020)—then it would be useful to understand how and why this has been attained. For present purposes, our point is that, perhaps paradoxically, we should not expect, ex ante, that agencies or departments in the business of collecting and curating data for guiding policy and performance should themselves be exemplary exponents of the deployment of that data to guide their *own* performance—because doing this is a separate ontological task, requiring distinct professional capabilities. Like the proverbial doctors, if data analysts cannot "heal themselves," we should not expect other public agencies to be able to do so merely by "infusing them with more and better data."
4. A special issue of *The Journal of Development Studies*, 51.2, was dedicated to this problem. For example, on the enduring challenges associated with agricultural data—another sector with a long history of data collection experience—see Carletto, Jolliffe, and Banerjee (2015).

5. The adage is popularly known as GIGO: garbage in, garbage out.
6. See the evolution in early work on gender inclusion in rural India and subsequent work (Ban and Rao 2008; Duflo 2012; Sanyal and Rao 2018).
7. This does not mean, of course, that nothing can be said about GEPs after five years—managers and funders would surely want to know by this point whether the apparent "no net impact" claim is a result of poor technical design, weak implementation, contextual incompatibility, countervailing political pressures, or insufficient time having elapsed. Moreover, they would likely be interested in learning whether the GEP's zero "average treatment effect" is nonetheless a process of offsetting outcomes manifest in a high standard deviation (meaning the GEP works wonderfully for some groups in some places but disastrously for others) and/or is yielding unanticipated or unmeasured outcomes (whether positive or negative). For present purposes, our point is that reliance on a single form and methodological source of data is unlikely to be able to answer these crucial administrative questions; with a diverse suite of methods and data, however, such questions become both askable and answerable. (See Rao, Ananthpur, and Malik 2017 for an instructive example, discussed below.)
8. One could say that this is a social scientific version of the Heisenberg uncertainty principle, in which the very act of measuring something changes it. See also Breckenridge (2014) on the politics and legacy of identity measurement in pre- and postcolonial South Africa and Hostetler (2021) on the broader manner in which imposing singular (but often alien) measures of time, space, and knowledge enabled colonial administration. More generally, Sheila Jasanoff's voluminous scholarship shows how science is a powerful representation of reality, which, when harnessed to technology, can reduce "individuals to standard classifications that demarcate the normal from the deviant and authorize varieties of social control" (Jasanoff 2004, 13).
9. Among the classic historical texts on this issue are *Peasants into Frenchman* (Weber 1976) and *Imagined Communities* (Anderson 1983). For more recent discussions, see Lewis (2015) on "the politics and consequences of performance measurement" and Beraldo and Milan (2019) on the politics of big data.
10. This is the finding, for example, from a major empirical assessment of cross-country differences regarding COVID-19 (Bollyky et al. 2022), wherein—controlling for a host of potential confounding variables—those countries with both high infections and high fatalities are characterized by low levels of trust between citizens and their governments and between each other. See further discussion of this study and its implications below.
11. The British movie *I, Daniel Blake* provides a compelling example of how even the literate in rich countries can be excluded by administrative systems and procedures that are completely alien to them—for example, filling out forms for unemployment benefits on the internet that require users to first "log on" and then "upload" a "CV." The limits of formal measurement to bring about positive policy change has long been recognized; when the Victorian-era writer George Eliot was asked why she wrote novels about the lives of the downtrodden rather than contributing to official government reports more formally documenting their plight, she astutely explained that "appeals founded on generalizations and statistics require a sympathy ready-made, a moral sentiment already in activity" (quoted in Gill 1970, 10). Forging such Smithian "sympathy" and "moral sentiment" is part of the important antecedent work that renders "generalizations and statistics" legible and credible to those who might otherwise have no reason for engaging with, or experience interpreting, such encapsulations of reality.
12. We fully recognize that, in principle, econometricians have methods available to identify both outcome heterogeneity and the factors driving it. Even so, if local average treatment effects are reported as zero, the "no impact" conclusion is highly likely to be the (only) key takeaway message. The primary benefit of incorporating both qualitative and econometric methods is the capacity of the former to identify factors that were not anticipated in the original design (see Rao 2022). In either case, Ravallion's (2001) injunction to "look beyond averages" when engaging with complex phenomena is worth being heeded by all researchers (and those that interpret researchers' findings), no matter their disciplinary or methodological orientations.
13. On the use of mixed methods in process evaluations, see Rogers and Woolcock (2023).
14. The six pillars were: maternal, newborn, and child health; childhood essential medicines and increasing treatment of important childhood diseases; improving child nutrition; immunization; malaria control; and the elimination of mother-to-child transmission of human immunodeficiency virus (HIV).
15. A PforR is one of the World Bank's three financing instruments. Its unique features are that it uses a country's own institutions and processes and links disbursement of funds directly to the achievement of specific program results. Where "traditional" development interventions proceed on the basis of ex ante commitments (for example, to designated "policy reforms" or to the adoption of procedures compliant with international standards), PforR-type interventions instead reward the attainment of predetermined targets, typically set by extrapolating from what recent historical trajectories have attained. According to the Project Appraisal Document for SOML, "each state would be eligible for a grant worth $325,000 per the percentage point gain they made above average annual gain in the sum of six indicators of health service coverage." The six indicators were: vitamin A, Pentavalent3 immunization, use of insecticide-treated nets (ITNs) by children under five, skilled birth attendance, contraceptive prevalence rate, and the prevention of mother-to-child transmission of HIV.
16. These tables are based on student performance in standardized tests in mathematics, reading, and science, which are administered by the Paris-based Organisation for Economic Co-operation and Development (OECD).

17. A summary of the report explains that

> whereas negotiation and flexibility are high on the list of skills for 2015, in 2020 they will begin to drop from the top 10 as machines, using masses of data, begin to make our decisions for us. A survey done by the World Economic Forum's Global Agenda Council on the Future of Software and Society shows people expect artificial intelligence machines to be part of a company's board of directors by 2026. Similarly, active listening, considered a core skill today, will disappear completely from the top 10. Emotional intelligence, which doesn't feature in the top 10 today, will become one of the top skills needed by all. (Gray 2016)

See also Soffel (2016).

18. Many companies and tertiary institutions are ahead of the curve in this regard. Recently, over 150 of the top private high schools in the US, including Phillips Exeter Academy and the Dalton School—storied institutions that have long relied on the status conveyed by student ranking—have pledged to shift to new transcripts that provide more comprehensive, qualitative feedback on students while ruling out any mention of credit hours, GPAs, or A–F grades. And colleges—the final arbiters of high school performance—are signaling a surprising willingness to depart from traditional assessments that have been in place since the early 19th century. From Harvard and Dartmouth to small community colleges, more than 70 US institutions of higher learning have weighed in, signing formal statements asserting that competency-based transcripts will not hurt students in the admissions process. See the "College Admissions" page on the New England Secondary School Consortium website: http://www.newenglandssc.org/resources/college-admissions/.
19. See Milante and Woolcock (2017) for a complementary set of dynamic quantitative and qualitative measures by which a given country might be declared a "fragile" state.
20. For development-oriented organizations, a set of tools and guidelines for guiding this initial assessment according to a political economy analysis (PEA) framework—crafted by USAID and ODI (London) and adopted by certain parts of the World Bank—is *Thinking and Working Politically through Applied Political Economy Analysis: A Guide for Practitioners* (Rocha Menocal et al. 2018). Its key observations include the following. First, a well-designed process of policy implementation should answer not only the technical question of what needs to be done but also how it should be done. Second, in-depth understanding of the political, economic, social, and cultural forces needs to supplement technical analysis to achieve successful policy implementation. Third, PEA should incorporate three pillars: the foundational factors (geography, natural resource occurrence, national borders), the "rules of the game" (institutions at the formal [political system, administrative structure, and law] and the informal [social and cultural norms] levels), and the "here and now" (current leaders, geopolitical situation, and natural hazards). Fourth, it is crucial to pay attention to the institutions, the structure of incentives, and the constraints, as well as the gains and losses of all the actors involved in policy implementation, including those outside of the traditional purview of development organizations. Fifth, policy solutions should be adjusted to political realities encountered on the ground in an iterative and incremental fashion. And finally, the evaluation of policy success should be extended to incorporate "process-based indicators," including trust and quality of relationship. Hudson, Marquette, and Waldock (2016) offer a guide for "everyday political analysis," which introduces a stripped-back political-analysis framework designed to help frontline practitioners make quick but politically informed decisions. It aims to complement more in-depth political analysis by helping programming staff to develop the "craft" of political thinking in a way that fits their everyday working practices.
21. On the application of such efforts to the case of policing in particular, see Sparrow (2018).
22. The *PDIA*toolkit*: A DIY Approach to Solving Complex Problems* (Samji et al. 2018) was designed by members of Harvard's Building State Capability program to guide government teams through the process of identifying, deconstructing, and solving complex problems. See in particular the section "Constructing your problem," which guides practitioners through the process of defining a problem that matters and building a credible, measurable vision of what success would look like.
23. As anthropologist Mike McGovern (2011, 353) powerfully argues, taking context seriously

> is neither a luxury nor the result of a kind of methodological altruism to be extended by the soft-hearted. It is, in purely positivist terms, the epistemological due diligence work required before one can talk meaningfully about other people's intentions, motivations, or desires. The risk in foregoing it is not simply that one might miss some of the local color of individual "cases." It is one of misrecognition. Analysis based on such misrecognition may mistake symptoms for causes, or two formally similar situations as being comparable despite their different etiologies. To extend the medical metaphor one step further, misdiagnosis is unfortunate, but a flawed prescription based on such a misrecognition can be deadly.

More generally, see Hoag and Hull (2017) for a summary of the anthropological literature on the civil service. Bailey (2017) provides a compelling example of how insights from qualitative fieldwork help explain the strong preference among civil servants in Tanzania for providing new water infrastructure projects over maintaining existing ones. Though a basic benefit-cost analysis favored prioritizing maintenance, collective action problems among civil servants themselves, the prosaic challenges of mediating local water management disputes overseen by customary institutions, and the performance targets set by the government all conspired to create suboptimal outcomes.

24. Says Mintzberg (2015): "Someone I know once asked a most senior British civil servant why his department had to do so much measuring. His reply: 'What else can we do when we don't know what's going on?' Did he ever try getting on the ground to find out what's going on? And then using judgment to assess that?"
25. Teskey (2017) and Wild, Booth, and Valters (2017) give examples of an adaptive logframe, drawn from Department for International Development experiences, that sets out clear objectives at the outcome level and focuses monitoring of outputs on the quality of the agreed rapid-cycle learning process. Strategy Testing (ST) is a monitoring system that the Asia Foundation developed specifically to track programs that are addressing complex development problems through a highly iterative, adaptive approach.

REFERENCES

Anderson, Benedict. 1983. *Imagined Communities: Reflections on the Origins and Spread of Nationalism.* London: Verso.

Andrews, Matt. 2013. *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions.* New York: Cambridge University Press.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action.* New York: Oxford University Press.

Bailey, Julia. 2017. "Bureaucratic Blockages: Water, Civil Servants and Community in Tanzania." Policy Research Working Paper 8101, World Bank, Washington, DC.

Ban, Radu, and Vijayendra Rao. 2008. "Tokenism or Agency? The Impact of Women's Reservations on Village Democracies in South India." *Economic Development and Cultural Change* 56 (3): 501–30. https://doi.org/10.1086/533551.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." NBER Working Paper 22746, National Bureau of Economic Research, Cambridge, MA.

Beraldo, Davide, and Stefania Milan. 2019. "From Data Politics to the Contentious Politics of Data." *Big Data & Society* 6 (2): 1–11. https://doi.org/10.1177/2053951719885967.

Bollyky, Thomas J., Erin N. Hulland, Ryan M. Barber, James K. Collins, Samantha Kiernan, Mark Moses, David M. Pigott, et al. 2022. "Pandemic Preparedness and Covid-19: An Exploratory Analysis of Infection and Fatality Rates, and Contextual Factors Associated with Preparedness in 177 Countries, from Jan 1, 2020, to Sept 30, 2021." *The Lancet* 399 (10334): 1489–512. https://doi.org/10.1016/S0140-6736(22)00172-6.

Breckenridge, Keith. 2014. *Biometric State: The Global Politics of Identification and Surveillance in South Africa, 1850 to the Present*. Cambridge, UK: Cambridge University Press.

Bridges, Kate, and Michael Woolcock. 2017. "How (Not) to Fix Problems That Matter: Assessing and Responding to Malawi's History of Institutional Reform." Policy Research Working Paper 8289, World Bank, Washington, DC.

Bridges, Kate, and Michael Woolcock. 2019. "Implementing Adaptive Approaches in Real World Scenarios: A Nigeria Case Study, with Lessons for Theory and Practice." Policy Research Working Paper 8904, World Bank, Washington, DC.

Cahill, Jonathan. 2017. *Making a Difference in Marketing: The Foundation of Competitive Advantage.* London: Routledge.

Carletto, Calogero, Dean Jolliffe, and Raka Banerjee. 2015. "From Tragedy to Renaissance: Improving Agricultural Data for Better Policies." *The Journal of Development Studies* 51 (2): 133–48. https://doi.org/10.1080/00220388.2014.968140.

Cartwright, Nancy. 2017. "Single Case Causes: What Is Evidence and Why." *In Philosophy of Science in Practice*, edited by Hsiang-Ke Chao and Julian Reiss, 11–24. New York: Springer.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better.* New York: Oxford University Press.

Caulkin, Simon. 2008. "The Rule Is Simple: Be Careful What You Measure." *Guardian* (US edition), February 9, 2008. https://www.theguardian.com/business/2008/feb/10/businesscomment1.

Caulkin, Simon. 2016. "Decision-Making: How to Make the Most of Data." *The Treasurer*, January 4, 2016. https://www.treasurers.org/hub/treasurer-magazine/decision-making-how-make-most-data.

Claxton, Guy, and Bill Lucas. 2015. *Educating Ruby: What Our Children Really Need to Learn.* New York: Crown House Publishing.

Dirks, Nicholas. 2011. *Castes of Mind.* Princeton, NJ: Princeton University Press.

Dom, Catherine, Alasdair Fraser, Joseph Holden, and John Patch. 2020. "Results-Based Financing in the Education Sector: Country-Level Analysis. Final Synthesis Report." Report submitted to the REACH Program at the World Bank by Mokoro Ltd.

Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature* 50 (4): 1051–79. https://doi.org/10.1257/jel.50.4.1051.

Economist. 2021. "How World Bank Leaders Put Pressure on Staff to Alter a Global Index." September 17, 2021. https://www.economist.com/finance-and-economics/2021/09/17/how-world-bank-leaders-put-pressure-on-staff-to-alter-a-global-index.

Gill, Stephen. 1970. "Introduction to Elizabeth Gaskell." In *Mary Barton: A Tale of Manchester Life.* London: Penguin.

Global Partnership for Education. 2020. *21st-Century Skills: What Potential Role for the Global Partnership for Education?* Washington, DC: Global Partnership for Education Secretariat. https://www.globalpartnership.org/sites/default/files/document/file/2020-01-GPE-21-century-skills-report.pdf.

Goodman, Peter S. 2009. "Emphasis on Growth Is Called Misguided." *New York Times*, October 4, 2009. https://www.nytimes.com/2009/09/23/business/economy/23gdp.html.

Gray, Alex. 2016. "The 10 Skills You Need to Thrive in the Fourth Industrial Revolution." World Economic Forum. https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution.

Hoag, Colin, and Matthew Hull. 2017. "A Review of the Anthropological Literature on the Civil Service." Policy Research Working Paper 8081, World Bank, Washington, DC.

Honig, Dan. 2018. *Navigation by Judgment: Why and When Top-Down Management of Foreign Aid Doesn't Work*. New York: Oxford University Press.

Honig, Dan, and Lant Pritchett. 2019. "The Limits of Accounting-based Accountability in Education (and Far Beyond): Why More Accounting Will Rarely Solve Accountability Problems." Working Paper No. 510, Center for Global Development, Washington, DC.

Hostetler, Laura. 2021. "Mapping, Registering, and Ordering: Time, Space and Knowledge." In *The Oxford World History of Empire: Volume One: The Imperial Experience*, edited by Peter Fibiger Bang, C. A. Bayly, and Walter Scheidel, 288–317. New York: Oxford University Press.

Hudson, David, Heather Marquette, and Sam Waldock. 2016. "Everyday Political Analysis." Working paper, Developmental Leadership Program, University of Birmingham, Birmingham, UK.

Jasanoff, Sheila. 2004. "Ordering Knowledge, Ordering Society." *In States of Knowledge: The Co-Production of Science and the Social Order,* edited by Sheila Jasanoff, 13–45. London: Routledge.

Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It.* Ithaca, NY: Cornell University Press.

Johnson, Janet Buttolph, Henry T. Reynolds, and Jason D. Mycoff. 2019. *Political Science Research Methods.* 9th ed. Thousand Oaks, CA: Sage.

Khan, Salman. 2012. *The One World Schoolhouse: Education Reimagined.* London: Hodder & Stoughton.

King, Gary, and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15 (1): 46–66. https://doi.org/10.1093/pan/mpl011.

Ladner, Debra. 2015. *Strategy Testing: An Innovative Approach to Monitoring Highly Flexible Aid Programs.* Working Politically in Practice Case Study 3. San Francisco: The Asia Foundation.

Lewis, Jenny M. 2015. "The Politics and Consequences of Performance Measurement." *Policy and Society* 34 (1): 1–12. https://doi.org/10.1016/j.polsoc.2015.03.001.

Mamdani, Mahmood. 2002. *When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda*. Princeton, NJ: Princeton University Press.

McCarthy, Niall. 2019. "The Countries Best Prepared to Deal With a Pandemic." *Statista*, October 28, 2019. https://www.statista.com/chart/19790/index-scores-by-level-of-preparation-to-respond-to-an-epidemic/.

McDonnell, Erin. 2017. "Patchwork Leviathan: How Pockets of Bureaucratic Governance Flourish within Institutionally Diverse Developing States." *American Sociological Review* 82 (3): 476–510. https://doi.org/10.1177/0003122417705874.

McDonnell, Erin. 2020. *Patchwork Leviathan: Pockets of Bureaucratic Effectiveness in Developing States.* Princeton, NJ: Princeton University Press.

McGovern, Mike. 2011. "Popular Development Economics: An Anthropologist among the Mandarins." *Perspectives on Politics* 9 (2): 345–55. https://doi.org/10.1017/S1537592711000594.

Merry, Sally Engle, Kevin E. Davis, and Benedict Kingsbury, eds. 2015. *The Quiet Power of Indicators: Measuring Governance, Corruption, and Rule of Law*. New York: Cambridge University Press.

Milante, Gary, and Michael Woolcock. 2017. "New Approaches to Identifying State Fragility." *Journal of Globalization and Development* 8 (1): 20170008. https://doi.org/10.1515/jgd-2017-0008.

Mintzberg, Henry. 2015. "If You Can't Measure It, You'd Better Manage It." *Henry Mintzberg* (blog). May 28, 2015. https://mintzberg.org/blog/measure-it-manage-it.

Natsios, Andrew. 2011. "The Clash of the Counter-Bureaucracy and Development." Essay, Center for Global Development, Washington, DC. https://www.cgdev.org/sites/default/files/1424271_file_Natsios_Counterbureaucracy.pdf.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton, NJ: Princeton University Press.
Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning.* Washington, DC: Center for Global Development.
Pritchett, Lant. 2014. "The Risks to Education Systems from Design Mismatch and Global Isomorphism: Concepts, with Examples from India." WIDER Working Paper 2014/039, United Nations University World Institute for Development Economics Research, Helsinki.
Propper, Carol, Simon Burgess, and Denise Gossage. 2008. "Competition and Quality: Evidence from the NHS Internal Market 1991–9." *The Economic Journal* 118 (525): 138–70. https://doi.org/10.1111/j.1468-0297.2007.02107.x.
Rao, Vijayendra. 2022. "Can Economics Become More Reflexive? Exploring the Potential of Mixed-Methods." Policy Research Working Paper 9918, World Bank, Washington, DC.
Rao, Vijayendra, Kripa Ananthpur, and Kabir Malik. 2017. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." *World Development* 99 (11): 481–97. https://doi.org/10.1016/j.worlddev.2017.05.037.
Ravallion, Martin. 2001. "Growth, Inequality and Poverty: Looking Beyond Averages." *World Development* 29 (11): 1803–15. https://doi.org/10.1016/S0305-750X(01)00072-9.
Ravallion, Martin. 2020. "Pandemic Policies in Poor Places." CGD Note, April 24, Center for Global Development, Washington, DC.
Ridgway, V. F. 1956. "Dysfunctional Consequences of Performance Measurements." *Administrative Science Quarterly* 1 (2): 240–7. https://doi.org/10.2307/2390989.
Robinson, Ken, and Lou Aronica. 2015. *Creative Schools: Revolutionizing Education from the Ground Up.* London: Penguin UK.
Rocha Menocal, Alina, Marc Cassidy, Sarah Swift, David Jacobstein, Corinne Rothblum, and Ilona Tservil. 2018. *Thinking and Working Politically through Applied Political Economy Analysis: A Guide for Practitioners.* Washington, DC: USAID, DCHA Bureau Center of Excellence on Democracy, Human Rights, and Governance. https://usaidlearninglab.org/sites/default/files/resource/files/pea_guide_final.pdf.
Rogers, Patricia, and Michael Woolcock. 2023. "Process and Implementation Evaluation Methods." In *Oxford Handbook of Program Design and Implementation*, edited by Anu Rangarajan, 294–316. New York: Oxford University Press.
Samji, Salimah, Matt Andrews, Lant Pritchett, and Michael Woolcock. 2018. *PDIAtoolkit: A DIY Approach to Solving Complex Problems*. Cambridge, MA: Building State Capability, Center for International Development, Harvard University. https://bsc.cid.harvard.edu/PDIAtoolkit.
Sandefur, Justin, and Amanda Glassman. 2015. "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics." *The Journal of Development Studies* 51 (2): 116–32. https://doi.org/10.1080/00220388.2014.968138.
Sanyal, Paromita, and Vijayendra Rao. 2018. *Oral Democracy: Deliberation in Indian Village Assemblies.* New York: Cambridge University Press.
Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* New Haven, CT: Yale University Press.
Soffel, Jenny. 2016. "The 21st-Century Skills Every Student Needs." World Economic Forum. https://www.weforum.org/agenda/2016/03/21st-century-skills-future-jobs-students/.
Sparrow, Malcolm. 2018. "Problem-Oriented Policing: Matching the Science to the Art." *Crime Science* 7 (1): 1–10. https://doi.org/10.1186/s40163-018-0088-2.
Teskey, Graham. 2017. "Thinking and Working Politically: Are We Seeing the Emergence of a Second Orthodoxy?" Governance Working Paper Series 1, ABT Associates, Canberra.
Teskey, Graham, and Lavinia Tyrrel. 2017. "Thinking and Working Politically in Large, Multi-Sector Facilities: Lessons to Date." Governance Working Paper Series 2, ABT Associates, Canberra.
Weber, Eugen. 1976. *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914.* Palo Alto, CA: Stanford University Press.
Wild, Leni, David Booth, and Craig Valters. 2017. *Putting Theory into Practice: How DFID Is Doing Development Differently.* London: ODI.
WilmerHale. 2021. *Investigation of Data Irregularities in "Doing Business 2018" and "Doing Business 2020": Investigation Findings and Report to the Board of Executive Directors.* https://thedocs.worldbank.org/en/doc/84a922cc9273b7b120d49ad3b9e9d3f9-0090012021/original/DB-Investigation-Findings-and-Report-to-the-Board-of-Executive-Directors-September-15-2021.pdf.
Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It.* New York: Basic Books.

Woolcock, Michael. 2018. "Reasons for Using Mixed Methods in the Evaluation of Complex Projects." In *Contemporary Philosophy and Social Science: An Interdisciplinary Dialogue*, edited by Michiru Nagatsu and Attilia Ruzzene, 149–71. London: Bloomsbury Academic.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise.* Washington, DC: World Bank.

World Economic Forum. 2016. *New Vision for Education: Fostering Social and Emotional Learning through Technology.* Industry Agenda. Geneva: World Economic Forum.

Zak, Paul. 2013. "Measurement Myopia." Drucker Institute, Claremont, CA. December 6, 2021. https://www.drucker.institute/thedx/measurement-myopia/.

Zhao, Yong. 2012. "Test Scores vs. Entrepreneurship: PISA, TIMSS, and Confidence." *Yong Zhao* (blog). http://zhaolearning.com/2012/06/06/test-scores-vs-entrepreneurship-pisa-timss-and-confidence/.