EDUCATIONAL THEORY

IN PRAISE OF ORDINARY MEASURES: THE PRESENT LIMITS AND FUTURE POSSIBILITIES OF EDUCATIONAL ACCOUNTABILITY

Jack Schneider

School of Education
University of Massachusetts – Lowell

Derek Gottlieb

School of Teacher Education
University of Northern Colorado

Abstract. State and federal policymakers "see" school performance via formal measures — data collected with attendance sheets and standardized tests. Such an approach, though not without its merits, is extremely limited and inherently exposed to the threat of systematic misperception and unintended consequences, especially as policymakers try to use data to leverage on-the-ground change. In this essay, Jack Schneider and Derek Gottlieb discuss the limitations of present accountability systems and advocate for the inclusion of what they call "ordinary measures." Long positioned as inferior to their formal counterparts, such measures offer much to clarify the picture of schools that good governance depends on. According to Schneider and Gottlieb, using ordinary measures, paired with deliberative evaluation processes, will improve the validity and utility of educational accountability systems.

Key Words. accountability; governance; measurement; policy

This essay advocates for an approach to educational measurement and decision making that has long been dismissed based on concerns about malfeasance and imprecision. Specifically, we offer a conceptual argument for including not only multiple measures, but multiple *approaches* to measurement in the process of educational accountability.

State- and federal-led efforts to gauge school performance at a massive scale — across nearly 100,000 schools in the United States — currently navigate a series of measurement tensions by relying heavily upon what we call "formal" measuring. Such an approach, which takes both "hard" and "soft" forms, is characterized by the creation of a measurement apparatus that officials can apply to all schools. Hard measuring is characterized by counting, and it centers around the production of metrics, like test scores and graduation rates, that align with broader constructs of interest, notably school quality. Soft measuring operates on similar assumptions, but it emphasizes representations of educational quality that are not fundamentally countable — things like student and teacher perceptions — by using instruments such as surveys and observation rubrics to produce data commensurable with hard measures. In both cases, formal measurement involves the construction of quantification tools (tests, surveys, and the like) that correspond to a particular aspect of school quality (academic achievement, school climate, and so on). These measurement devices work in concert with automated evaluative systems to produce official school ratings, which in turn serve to meet

state obligations, inform parents and communities, and guide school improvement efforts.

"Ordinary" measuring, by contrast, draws on the nontechnical capacities we use in all aspects of our public lives, including education, as we attempt to do the right thing, to be good neighbors, and to realize the aspirations of our vocations. In their attempts to understand school quality, stakeholders regularly rely on ordinary measuring; in fact, such an approach precedes its formal counterpart. Centralized accountability systems, however, allow no room for ordinary measurement. Consequently, when gaps open between the results of formal and ordinary measurement, stakeholders tend to question the legitimacy of the former rather than the latter.[1] And, as we suggest, they are right to do so.

Formal measures, particularly in their hard form, were a natural choice for state- and federal-led accountability; as we explain in this essay, the tilt is not without reason. Yet, as we also explain, the exclusion of ordinary measuring renders such systems vulnerable to critique on three grounds: narrowness, uniformity, and rigidity.[2] Our aim in this essay, then, is to demonstrate the need for the explicit inclusion of ordinary measuring in order to enhance the breadth, particularity, and adaptability of accountability systems.

Formal versus Ordinary Measurement

Formal measuring is defined by a modern-science view of logical and procedural rigor, both of which are geared toward assuring users — policymakers, practitioners, and the public — that these measures accurately reflect reality. The idea in formal measuring is to create an external representation of the construct of interest — a literal or figurative meter stick — against which anyone can hold up

1. When we use "legitimacy" in this paper, we will consistently use it in this sense. We are not concerned with the legitimacy bestowed by democratically established procedures for issuing judgments, where Rawls centers his attention (see John Rawls, *A Theory of Justice* [Cambridge, MA: Harvard University Press, 1971]); rather, here we are concerned with the legitimacy of the results of judgment and their ability to attract public consent. What "construct validity" is to measurement, our use of "legitimacy" is to accountability decisions. For a more thorough discussion of the distinctions, see Stanley Cavell, *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism* (Chicago: University of Chicago Press, 1990), chap. 3; and Katrina Forrester, *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy* (Princeton, NJ: Princeton University Press, 2019), chap. 2.

2. Kathleen Knight Abowitz, *Publics for Public Schools: Legitimacy, Democracy, and Leadership* (London: Routledge, 2016), 39–43; and Kenneth Strike, "Liberty, Democracy, and Community," *Yearbook of the National Society for the Study of Education* 102, no. 1 (2003).

JACK SCHNEIDER is Associate Professor in the School of Education at the University of Massachusetts – Lowell; e-mail <jack_schneider@uml.edu>. He is Director of Research for the Massachusetts Consortium for Innovative Education Assessment and the co-host of the education policy podcast *Have You Heard*.

DEREK GOTTLIEB is Assistant Professor in the School of Teacher Education at the University of Northern Colorado; e-mail <derek.gottlieb@unco.edu>. His primary areas of scholarship are education reform, accountability policy, and democratic theory.

a particular real-world object to gauge the amount of a given property that exists. Once that quantity is known, a decision about whether or not it is sufficient can be made.

Formal measuring rests on a few key assumptions. It requires that the construct being measured is external to us, like a natural property. The concept of length, for instance, formally measurable with a meter stick, does not depend on or involve anything human: as of 1983, the "meter" is defined in terms of the speed of light, a universal constant.[3] Formal measuring further requires that the construct is singular and stable, making it reasonably amenable to uncontroversial definition. We can use a meter stick to measure the length of an object, and we can count the result as knowledge, because "length" picks out a single dimension of the physical universe that remains conceptually consistent. Even when the natural object of comparison used to define the length of a meter changes, as from the "the arc of a meridian" to the speed of light, this only affects the precision of the measuring device, not the concept of length.[4] Finally, formal measuring requires a mechanical procedure for conducting measurement — another formal element. One must take care not to hold the meter stick up to a wall at an angle if one wants to know how long the wall is; there is a proper way of comparing the wall to a standard of length in order to produce an accurate measurement.

These assumptions do not hold, however, with respect to the human or social world. Length is a property of the physical universe in which we live, not a property produced by the *way* that we live; it would exist whether we existed or not. Our social realities are fundamentally different from naturally occurring features in this regard. What it means to be a "good citizen," for example, cannot be explained in terms of any absolute standard of goodness or any absolute definition of citizenship, but only by reference to current definitions and actual examples. This is because our social concepts are ultimately grounded not in nature, but in what Hubert Dreyfus describes as "our sense of what we *are*, which is … something we can never explicitly *know*."[5] Our social concepts, which we learn from, share with, and teach to others, are rooted in what Ludwig Wittgenstein calls "agreement in form of life," which comes prior to, and is the necessary condition for, the derivation of any formal definitions or rules.[6] For that reason, too, our concepts are neither readily isolable from one another, nor stable across time and space, as the concept of "citizenship" clearly demonstrates.[7] It is, of

3. Robert Crease, *World in the Balance: The Historic Quest for an Absolute System of Measurement* (New York: W. W. Norton, 2011), 251.

4. Ibid., 190.

5. Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992), 56.

6. Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte (London: Wiley-Blackwell, 2009), §§241-242

7. Henry Louis Gates Jr., *Stony the Road: Reconstruction, White Supremacy, and the Rise of Jim Crow* (New York: Penguin Press, 2019); Eric Foner, *The Second Founding: How the Civil War and*

course, possible to devise a mechanical procedure for measuring "citizenship" or similar concepts; but, because the concepts themselves have none of the solidity that formal measurement requires, any such method for producing measurements would fall victim to a "garbage-in, garbage-out" problem. Taken together, these facts suggest that social concepts do not naturally lend themselves to formal measurement.

This discrepancy between the requirements of formal measurement and the basic reality of our social concepts is not a new discovery. Wittgenstein's *Tractatus* attempted to formalize language on the basis of logic, which proved possible only when the job of "language" was reduced to describing the world.[8] Dreyfus made a career of tracing the repeated failures of twentieth-century artificial intelligence research to the faulty assumption that all of our understanding, including of social practices, was formalizable.[9] When we insist on formalizing such things, it leads either to reductiveness, as Wittgenstein found, or to conceptual distortion, as Dreyfus shows. In each case, we produce something that we can measure by formal means, but it is not exactly the thing we wanted to measure.

Ordinary measuring, however, allows us to account for the fact that we measure social things all the time. Whatever the challenges of formally measuring social concepts, we must remember that we can point to examples of good citizens and readily enumerate the definitional features that make them exemplary. Ordinary measuring is therefore a species of our everyday capacity for judgment. It remains a form of *measurement* because it involves comparing particulars against a standard, like comparing a particular wall against a meter stick in an assessment of length.[10] But because none of the assumptions that undergird formal measuring hold for social concepts, both the nature of the standard that we use, and the practice of comparing particulars against it, must be different.

Ordinary measuring specifically involves three features that formal measuring lacks: the negotiation of fitness, the implication of the self, and the openness to corrigibility by others. In an everyday example of ordinary measuring, a professor might ask himself, in front of the mirror, "What is the best tie to wear with this shirt?" As he scans the ties on the rack in his closet, he is looking for the one that will fit on this particular occasion — a high-profile lecture. "Fit" is not an absolute construct, but is internally related to the occasion itself, his role in it, and a broader and nonformalizable sense of how things are done. He considers the type of function, the likely audience, the message he is trying to send, his own gender identity, his aesthetic preferences, and the ties he has available. The list of possible considerations is open-ended, and the way the considerations come

Reconstruction Remade the Constitution (New York: W. W. Norton, 2019); and Sam Erman, *Almost Citizens: Puerto Rico, the US Constitution, and Empire* (Cambridge: Cambridge University Press, 2018).

8. Wittgenstein, *Philosophical Investigations*, §114.

9. Dreyfus, *What Computers Still Can't Do*, 190.

10. Crease, *World in the Balance*, 270.

together in a final decision is unpredictable: this openness and unpredictability explains ordinary measurement's characteristic need to negotiate fitness.

This open-endedness and unpredictability is also why we find ourselves implicated in the results of ordinary measurement. The ultimate choice of tie, precisely because it cannot be made in terms of absolute criteria, reveals something about the professor himself, namely his understanding of the situation and the way he has valued or failed to value various possible considerations. All aspects of this measurement process, because they derive from public practices, are open to view and critique. Was the message the professor sent, via his accessory, appropriate to the classroom setting? Is the professor aware that wearing this particular color on this particular day is significant, or was it merely an accident? One's way of assessing an overall situation and factoring in relevant features is openly visible in the process of ordinary measurement: one's stance on the world is revealed.

Lastly, ordinary measurement is open to correction by others and augmentation by new information. The tie-wearing professor can discover from his friend — via a skeptical facial expression, or a gentle comment — that his measuring procedure has gone awry. Perhaps he has been unaware of emerging trends in neckwear. Or perhaps he is reminded of the typical dress code for lectures like the one he is to give. Such new information would shift his perspective, leading him to see the situation as a whole, and his decision within it, from a larger point of view — his own *and* that of another.

These three features of ordinary measuring are necessary precisely because of the absence of an external, unchanging standard against which to compare our social concepts. This fact means that our judgments regarding social things will regularly differ. And when we differ in the measurements or judgments we produce, it is not because we individually hold radically different concepts in our heads — something that we could call upon a formal definition to adjudicate. Rather, it is because we are examining common concepts, produced in relation to our form of life, from different vantage points. Our individual perspectives are not limitations here, as they would be in formal measuring; there is no analogously external vantage we could take up with respect to our social things. Attempting to claim such a vantage — disowning our ordinary embeddedness in the world — is what leads to the reductiveness or distortion discussed earlier.[11]

Our individual perspectives on social concerns are not private beliefs. Instead, as Dreyfus observes, they are "local elaborations of a whole which they presuppose."[12] Bringing these perspectives together in combination achieves the kind of objectivity proper to social things: we ordinarily do this not by achieving an

11. Derek Gottlieb, *A Democratic Theory of Educational Accountability: From Text-Based Assessment to Interpersonal Accountability* (London: Routledge Press, 2020), 62–67.

12. Dreyfus, *What Computers Still Can't Do*, 14.

absolute perspective on a social matter, but by seeing a topic from various sides.[13] Settling differences in value judgments of social concepts, therefore, cannot be done adequately by appealing to formal measures. What we need in these instances is an approach to measurement that acknowledges the necessity of our differently situated perspectives on a common matter. Such an approach would also acknowledge, as Linda Zerilli writes, that our "perspectives are corrigible not by something that is extraperspectival or neutral, but by other perspectives themselves."[14] Unfortunately, a century's worth of so-called "good-governance" traditions makes such an approach look counterintuitive at best, especially where the state's responsibility for education is concerned.

How State and Federal Offices "See" Schools

Like their peers in other sectors of government, state and federal education officials have worked to produce formal measures to facilitate performance management systems. As James Scott argues, administrative officials typically work without firsthand, on-the-ground knowledge because of the scale and heterogeneity of their charges.[15] This is a major obstacle for centralized offices: they cannot act upon schools if they cannot "see" them; and offices working at the state or federal scale of management do not have enough eyes for the task.

In response, state and federal leaders rely on quantitative systems designed to measure performance — in essence, "seeing" schools through the technology of formal measurement. Scott calls this the "synoptic" view, which transforms a disorganized social reality into a legible and administratively convenient format. Like a map, which leaves out most of an environment's complexity, the synoptic view captures only what it is programmed to include.[16]

Formal measuring is also a natural fit because it helps state and federal officials avoid politicization. Already operating at some distance from any local context, centralized offices base their credibility on their disinterestedness. As Theodore Porter puts it: "The appeal of numbers is especially compelling to bureaucratic officials who lack the mandate of a popular election, or divine right. Arbitrariness and bias are the most usual grounds upon which such officials are criticized."[17] Explicit rules of measurement and indisputable numbers offer the appearance of fairness and impartiality.

State and federal offices have compelling reasons for involving themselves in the governance of schools, as well as for relying on formal measurement in

13. Hannah Arendt, *The Human Condition*, 2d ed. (Chicago: University of Chicago Press, 1958), 52; and Linda Zerilli, *A Democratic Theory of Judgment* (Chicago: University of Chicago Press, 2016), 32.

14. Zerilli, *A Democratic Theory of Judgment*, 8.

15. James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven, CT: Yale University Press, 1998).

16. Ibid., 55–59.

17. Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, NJ: Princeton University Press, 1996), 23.

such work. The aim of equity, for instance, is impossible to pursue without the affordance of what Thomas Green calls "aggregation" — a process that requires establishing commensurability across schools and compiling cumulative data.[18] Over the past two decades, however, the unintended consequences of formal measurement systems, particularly in high-stakes applications, have steadily revealed themselves. Daniel Koretz, for example, has reinvigorated attention to Campbell's Law, pointing to the tendency of strict accountability regimes to incentivize "score inflation" rather than real gains.[19] At the same time, some of the staunchest supporters of performance management have recognized the perverse consequences — like teaching-to-the-test and narrowing of the school curriculum — that they failed to anticipate.[20] This is to say nothing of several notorious cheating scandals that rocked major school districts, perhaps most notably in Atlanta.[21] It seems, then, that the same limitation requiring formal measuring in the first place — the inability to perceive local activity except through objectification — makes the system vulnerable to gaming.

Formal measuring, of course, boasts undeniable strengths. Attributes like accuracy, comparability, and automation allow for a level of oversight that would otherwise be unimaginable. And policy's focus on formal measuring has certainly yielded some large-scale successes, particularly for historically underserved students whose marginalization would remain relatively indiscernible if not for standardized forms of measurement.[22] Nevertheless, formal measuring comes at the expense of other attributes, such as breadth, particularity, and adaptability, that are essential for seeing well enough to act reasonably. These attributes are primarily associated with local perspectives and standpoints. When formal measures produce judgments that contradict the lived reality of local stakeholders, the content of those judgments — if not their consequences — is easily dismissed. The absence of these attributes, then, rather than anything about formal measuring itself, has imperiled the legitimacy of state accountability systems.[23]

18. Thomas F. Green, "Excellence, Equity, and Equality," in *The Handbook on Teaching and Educational Policy*, ed. Lee S. Shulman and Gary Sykes (New York: Longman, 1982).

19. Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008); and Daniel Koretz, *The Testing Charade: Pretending to Make Schools Better* (Chicago: Chicago University Press, 2017).

20. Diane Ravitch, *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education* (New York: Basic Books, 2010), 12–13.

21. Rachel Aviv, "Wrong Answer: In an Era of High-Stakes Tests, a Struggling School Made a Shocking Choice," *New Yorker*, July 24, 2014, https://www.newyorker.com/magazine/2014/07/21/wrong-answer.

22. Thomas S. Dee and Brian A. Jacob, "The Achievement Consequences of the No Child Left Behind Act" (paper presented at the NCLB: Emerging Findings Research Conference at the Urban Institute, Washington, DC, August 12, 2009); and Douglas Lauen and S. Michael Gaddis, "Shining a Light or Fumbling in the Dark? The Effects of NCLB's Subgroup-Specific Accountability on Student Achievement," *Educational Evaluation and Policy Analysis* 34, no. 2 (2012): 185–208

23. Dana Mitra, Bryan Mann, and Mark Hlavacik, "Opting Out: Parents Creating Contested Spaces to Challenge Standardized Tests," *Education Policy Analysis Archives* 24 (2016): 1–23; Nicholas Tampio,

Must state and federal offices accept such limitations if they wish to play an active role in education? Must we choose between formal measuring and nothing at all? The ability to align policy intentions with policy outcomes depends upon resisting such false choices. If state and federal offices are to play an active role — as they should, given their power to advance the essential aim of equity — they must resolve the problems generated by their particular approach to measuring. In service of that aim, we consider three central tensions that should demand the attention of scholars and policymakers.

Tension 1: Breadth versus Accuracy

One tension between formal and ordinary measuring pertains to the different goals of seeing a concept in its full breadth and seeing it in a manner free from human error. On the surface, these aims may not seem contradictory. Yet seeing broadly depends on methodological flexibility — it depends on taking in as much as possible from a contextualized vantage point and making sense of those various inputs.[24] Seeing a well-defined thing accurately, by contrast — that is, perceiving it from a universal, decontextualized perspective — requires methodological rigidity. Tilting toward one aim requires a shift away from the other.

State and federal leaders have leaned toward formal measuring and its claim on accuracy. Insofar as the legitimacy of state and federal offices is rooted in objectivity, rather than politics, they depend on systems that align with or enhance that fundamental concern. As Porter explains: "A decision made by the numbers (or by explicit rules of some other sort) has at least the appearance of being fair and impersonal. Scientific objectivity thus provides an answer to a moral demand for impartiality and fairness. Quantification is a way of making decisions without seeming to decide. Objectivity lends authority to officials who have very little of their own."[25]

By contrast, consider how educators develop knowledge about student performance. Leveraging their proximity to students and using their powers of perception, teachers develop comprehensive understandings of what students know and can do. The problem, from the perspective of the state or federal government, is that individual teachers have their own values, definitions, experiences, and orientations. Consequently, any firsthand measuring by teachers will inevitably introduce

Common Core: National Education Standards and the Threat to Democracy (Baltimore, MD: Johns Hopkins University Press, 2018); Terri S. Wilson, "Refusing the Test: Youth Activism and the Right to Opt Out of State Assessments," *Philosophy of Education 2018*, ed. Megan Laverty (Urbana, IL: Philosophy of Education Society, 2018), 575–587; and Terri S. Wilson, Matthew Hastings, and Michele S. Moses, "Opting Out as Democratic Engagement? The Public Dimensions and Challenges of Education Activism," *The Good Society* 25, no. 2–3 (2017): 231–255.

24. Hubert L. Dreyfus and Stuart E. Dreyfus, "Peripheral Vision: Expertise in Real World Contexts," *Organization Studies* 26, no. 5 (2005): 779–792; and Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. Donald Landes (London: Routledge, 2013), 17.

25. Porter, *Trust in Numbers*, 23.

variability.[26] To centralized offices, such variability is intolerable because it opens the door to claims of arbitrariness and bias.[27] Rather than grapple with the case-by-case complexity, then, state and federal officials have opted for uniform definitions and clear benchmarks. As definitions and forms of measuring tighten, accuracy improves. At the same time, the comprehensiveness of the perspective suffers.

The shift away from teacher judgments about student performance provides an illustration of this tension. Letter grades emerged in the nineteenth century as a way of comparing students to standards of expectation, and these grades were assigned by teachers. In fact, they initially reflected not merely the academic dimension, but also moral and behavioral dimensions — a clear indicator of a tendency toward breadth in evaluation.[28] Grades were an initial attempt to capture the entire mission of schooling in a single, easily communicable rating system. But the effort at total description required the use of individual teachers' moral judgments in generating a grade, and those judgments could vary substantially. Eliminating moral and behavioral elements from the concept of grading was only a temporary solution, as teacher judgments of academic quality also vary. Rather than rely on teachers' perceptions of students, then, states pursued a different path, adopting standardized tests in "an effort to replace impressionistic evaluations of students with 'hard facts.'"[29] By the time No Child Left Behind (NCLB) became law in the United States, standardized tests — mandated by the federal government in all fifty states — were the chief mechanism for measuring student academic performance.

State and federal offices have paid a price for this panoptic view of student learning. Presently, they must ignore the conceptual slippage between achievement on math and reading exams and the far larger construct of "student performance" that such tests seek to measure.[30] Because their preferred metrics can only capture a tiny slice of what "student performance" means, they must choose between relying on their narrow set of perceptual tools and sorting through a wild proliferation of incommensurable perspectives. State and federal leaders understandably opt for using perfectly standard definitions and measures. Yet much is lost in that process. A broad concept like "student performance" is reduced to a metric as accurate as it is narrow. It is both right and wrong at the same time.

26. Hannah Arendt, *The Life of the Mind* (New York: Houghton Mifflin Harcourt, 1981), 83; and Zerilli, *A Democratic Theory of Judgment*, 4–6.

27. Theodore M. Porter, "Thin Description: Surface and Depth in Science and Science Studies," *Osiris* 27, no. 1 (2012): 210.

28. Ethan L. Hutt and Jack Schneider, "A Thin Line Between Love and Hate: Educational Measurement in the United States," in *Assessment Cultures: Historical Perspectives*, ed. Cristina Alarcón López and Martin Lawn (New York: Peter Lang, 2018), 237.

29. Ibid., 240.

30. Derek Gottlieb, *Education Reform and the Concept of Good Teaching* (New York: Routledge, 2015).

Tension 2: Particularity versus Comparability

State and federal officials are charged with overseeing more districts and schools than any single person could know well or describe. Leaders of centralized offices face this conundrum: they are responsible for ensuring that each school provides a high-quality education, but they cannot see specifically what each school is actually giving to its students. Such leaders cannot, in other words, render a holistic assessment. Were all schools identical, this might not be so troubling. Yet schools operate in different contexts and take different approaches, even while pursuing similar aims or standards. This is only to say that what certain levels of test performance *mean* in terms of a school's quality will vary with the school's mission. The perceptual capacities of centralized offices, however, cannot account for such differences, much less formulate accountability mechanisms tailored to each setting.

As in the history of letter grades, state and federal responsibility for ensuring that each school was providing its students with an adequate education, regardless of special emphasis, led such offices to value comparability over particularity. In order to ensure an adequate education for each student, centralized officials needed to "see" every school, and to see only those elements that every school would have in common with every other school. To facilitate this kind of sight, state and federal offices have defined the "core" aims of public education and adopted uniform metrics to gauge performance against those aims.[31]

Such an approach succeeds in capturing all schools and districts within a uniform administrative gaze, facilitating both criterion referencing and norm referencing. All schools, in other words, can be compared against some established benchmark, as well as against each other. The tradeoff, obviously, is that state and federal offices sacrifice the means to see and value the particularities that make schools distinct from one another. Whether a school focuses on the performing arts or career training, only common-denominator competencies can matter for state and federal purposes.

But the sheer inability to include important context in state and federal systems threatens a number of the purposes that accountability serves. Perhaps most seriously, when state and federal accountability systems ignore particular school missions, they can produce evaluations that diverge from stakeholders' direct experience with those schools. And that, in turn, can undermine the perceived legitimacy of the accountability system.[32] Thus, while it may be unwise to treat each individual school as if it is completely unique, entirely ignoring particularity appears to sacrifice too much.

Looking at a state's actual rating of a particular school will help to illustrate our point. Let us call it School X. School X is situated in a relatively poor area of South Boston, Massachusetts. According to the two primary accountability aggregations

31. Tampio, *Common Core*, 17–27.

32. Koretz, *The Testing Charade*; and Knight Abowitz, *Publics for Public Schools*, 23–27.

that the state uses, School X looks weak: the school has achieved a "target percentage" of 35 percent, when 75 percent would indicate acceptable performance. And on its "accountability percentile," School X only outscores 17 percent of other schools "that administer similar MCAS [Massachusetts Comprehensive Assessment System] tests" — the annual exams issued by the state of Massachusetts in accordance with federal law. According to the 2018 School Report Card produced by the state, School X is only "partially meeting" its accountability targets, barely avoiding mandatory interventions.[33] This is primarily what the state can see, and it does not look good. If there were ever a school in need of more oversight, School X would be an obvious candidate.

Ought it make a difference that School X is Boston Arts Academy (BAA), which aspires to be "a laboratory and a beacon for artistic and academic innovation," that it is widely recognized as a leading school in the area, and that it prides itself on offering "high-level training for students in dance, music, theater, or the visual arts in the context of a college preparatory curriculum"?[34] The state cannot see the particularity of BAA's mission, nor can it see the reasons parents would fight to enroll their children there. Certainly, we recognize that not every school with faltering test scores will turn out to be BAA. Still, the point of the example is that every school serves its students and families in ways that go beyond what the state's distant gaze can perceive.

In short, central offices of education can only "see" educational effects that might be found in *any* school — the lowest common denominator across these often unique, locally influenced organizations. In so doing, such offices fail to capture distinctive approaches and context-appropriate practices, thereby ensuring that their view of schools will be, at best, incomplete. Indeed, to the extent that state and federal agencies enact policies designed to give "teeth" to accountability systems, they positively discourage the kinds of excellence and innovation that schools such as Boston Arts Academy pursue. Once more, the state is as wrong as it is right.

Tension 3: Adaptability versus Automation

The aforementioned tensions pertain to difficult choices among the kinds of metrics that a state or federal office might use to determine educational quality. Our third tension, however, pertains to the *process* of determining educational quality — that is, of arriving at evaluative judgments that will drive consequential local decisions.

Under the current Every Student Succeeds Act (ESSA), state- and federal-led measurement feeds directly into an accountability calculation — an algorithm combining test scores, graduation rates, and a small handful of other measures.

33. Massachusetts Department of Elementary and Secondary Education, "School Report Card, Boston Arts Academy" (2018). Retrieved from http://reportcards.doe.mass.edu/2018/00350546.

34. Boston Public Schools, "Boston Arts Academy" (2019). Retrieved from https://www.bostonpublicschools.org/school/boston-arts-academy.

The calculation renders judgments, or ratings, while circumventing the biases or prejudices of human intervention. Such automation meets a certain standard of procedural fairness; yet, it will regularly produce suboptimal results on the ground unless it can be paired with reasonable adaptability.[35]

This procedural fairness, rooted in aggregation, has been an important tool in helping state and federal offices meet their responsibility for ensuring equity. Any reasonable adaptability that complements automation must support rather than subvert this mission. Aggregative measures, including test scores, have been instrumental in revealing persistent inequities and have therefore provided civil rights groups with an important lever for demanding state action.[36] Automated procedures for issuing evaluative judgments, moreover, have made it impossible for local schools to avoid the consequences of what these aggregative measures reveal. Thus, automatic accountability processes can serve — and have served — the needs of the state and the nation, as well as of vulnerable constituencies. In support of this claim, one might point to the improved outcomes for students — particularly students from historically marginalized subgroups — that are frequently attributed to test-based accountability.[37]

But automating such processes, even in the pursuit of equity, can also cause as many problems as it solves. Automated evaluations of school quality in the interest of promoting social equity generally rest on the assumption that future outcomes can be predicted by a prescribed set of metrics.[38] Yet education is staggeringly complex. Students, for whom the educational process is supposed to work, come from different families, neighborhoods, and social contexts. Desirable outcomes are multiple and multifaceted. And mediating processes, like relationships, work differently depending on who the students are and what the school is trying to accomplish. In short, algorithmic judgment may ensure a kind of procedural fairness while, at the same time, failing to account for what students actually need.

35. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2016), 208.

36. Jesse Hessler Rhodes, "Progressive Policy Making in a Conservative Age? Civil Rights and the Politics of Federal Education Standards, Testing, and Accountability," *Perspectives on Politics* 9, no. 3 (2011): 519–544.

37. Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio, "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," Working Paper 13 (National Center for Analysis of Longitudinal Data in Education Research, 2007); and Brian A. Jacob, "Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments," NBER Working Paper No. 12817 (National Bureau of Economic Research, 2007).

38. Thomas J. Kane, Jonah E. Rockoff, and Douglas O. Staiger, "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27, no. 6 (2008): 615–631; and Eric A. Hanushek and Ludger Woessmann, *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes* (Paris, France: OECD Publishing, 2010).

The early years of schooling under NCLB provide an excellent example of this duality. On the theory that universal "proficiency" was a good proxy for the education system's overall mission, the law required that each school show "adequate yearly progress" toward 100 percent proficiency. In practice, however, this sometimes had the paradoxical effect of directing attention in an inequitable fashion.[39] Historically marginalized students, unlike their more privileged counterparts, received a narrowed curriculum and a test-driven pedagogy.[40] NCLB tethered inadequate and predetermined criteria to an algorithmic process for issuing high-stakes judgments. As a result, the law perversely incentivized the kinds of inequity that such legislation was explicitly intended to remedy.[41] The combination of formal measurement and automatic judgment, even when intended to advance equity, structurally exposes accountability systems to the possibility of radical error, up to and including the active frustration of the system's stated goals.

ESSA, which replaced NCLB in 2015, seems to have understood the previous law's failures to be remediable by including further measures within the same algorithmic process of judging: it took the narrowness of formal measures themselves, rather than the formal process of combining measures, to be the central problem. The new law has therefore expanded the set of measures allowed and encouraged by the federal government. Soft measures, for instance, can now be included in state accountability plans along with hard measures, as long as they meet the requirements for validity, reliability, and comparability set forth in §1111(c)(2)(H)(v) of ESSA. Unlike NCLB, in which accountability was entirely reliant upon hard measures, ESSA now requires test scores, graduation rates, English-language proficiency metrics, and an additional, state-developed measure. This is superficially a more nuanced approach to the measures used for accountability purposes, but it is not enough. As Koretz puts it, "ESSA only slightly broadens the focus from test scores, does nothing to confront Campbell's Law, doesn't allow for reasonable variations among students, doesn't take context into account, doesn't make use of professional judgment, and largely or entirely … continues to exclude the quality of educators' practice from the mandated accountability system."[42] Including

39. Derek Neal and Diane Whitmore Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics* 92, no. 2 (2010): 263–283; and Jennifer Booher-Jennings, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42, no. 2 (2005): 231–268.

40. Wayne Au, *Unequal by Design: High-Stakes Testing and the Standardization of Inequality* (London: Routledge, 2010); Wayne Au, "Meritocracy 2.0: High-Stakes, Standardized Testing as a Racial Project of Neoliberal Multiculturalism," *Educational Policy* 30, no. 3 (2016): 39–62; and John B. Diamond and James P. Spillane, "High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *Teachers College Record* 106, no. 6, (2004): 1145–1176.

41. Laura S. Hamilton, Mark Berends, and Brian M. Stecher, *Teachers' Responses to Standards-Based Accountability* (Santa Monica, CA: RAND, 2005); and Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118, no. 3 (2003): 843–877.

42. Koretz, *The Testing Charade*, 246.

soft measures alongside hard ones might appear to cast a wider net. But because soft measures differ from hard measures in degree rather than in kind — they are both formal forms of measurement — they lend themselves to the same kind of algorithmic judgment that plagued NCLB's rating system.[43]

An ordinary approach to producing judgments or ratings of educational institutions would not banish formal measures from consideration; rather, it would require the creation of deliberative bodies to put those measures in conversation with one another. Consider jury deliberations, in which the accepted aim is to consider the facts of a particular case, in light of the letter of the law, and then to render a judgment. ESSA has allowed — and required — a greater number of facts to enter into the conversation, but there is still no *conversation* to be had. The process of rendering judgment remains algorithmic because formal measurement is designed to facilitate automation. Take Colorado's ESSA plan, for example, which says that "each school will receive a summative index score … based on points assigned for each of the five ESSA indicators."[44] The relations among the facts of any given case are laid out in advance, and evaluative judgments are made algorithmically. This sort of automatic judgment from above, without further attention to the particularities of any given context, is precisely what aroused public ire under NCLB.

A process of determining the *meaning* of the facts in any given case, and what that meaning implies for the process of school improvement, remains necessary. In place of algorithmic approaches to producing index scores and ratings, deliberative bodies might triangulate among the facts emerging from a given school, state performance standards, and broader notions of education's mission — not only to determine whether the school is meeting its responsibilities, but also precisely where and how it needs to improve. This triangulation would facilitate objectivity in the Arendtian sense discussed earlier: a bringing-together of multiple perspectives on an object of common concern.[45] This sort of process would lend resulting judgments the power and purchase to make the meaningful reforms that all parties recognize as necessary. The shared responsibility for looking closely at a school and mapping a path forward would help generate the buy-in necessary to realize the school's aspirations.

Critics might worry that leaving room for adaptability in rendering judgments raises the specter of the kinds of bad discretion that has so often occurred at the

43. Under its ESSA plan, Iowa now includes a measure of "school climate." But this is called the "IS3 Index," and it algorithmically combines data derived from discipline and suspension reporting, student and family surveys, absenteeism figures, and so on, according to a predetermined weighting system. The index scores are then tied to a four-tiered scale from "healthy" to "unhealthy." Iowa Department of Education (IDE), *Every Student Succeeds Act in Iowa* (May 3, 2018), 204–224; https://educateiowa.gov/sites/files/ed/documents/2019-12-23ESSAPlan_508.pdf.

44. Colorado Department of Education (CDE), *Consolidated State Plan under the Every Student Succeeds Act (ESSA)* (2018), 75; http://www.cde.state.co.us/fedprograms/co-consolidatedstateplan-final-websitepdf.

45. Zerilli, *A Democratic Theory of Judgment*, 29.

expense of marginalized communities. But automated procedures for rendering judgments and handing down consequences have proven weaponizable, as well: the danger that critics point to is real, but it is not solved simply by replacing bad discretion with no discretion.[46] Far from guaranteeing the intended results, this displacement simply makes negative consequences, such as system-gaming, impossible to remedy. We cannot absolutely preclude human malfeasance in this or any other social process, but forbidding adaptability actively prevents those with contextualized knowledge from ever intervening on the side of the good in the process of measurement.

A Measured Plan

In the brief section that follows, we outline a plan for incorporating ordinary measuring into existing accountability systems. Given the limits of space, such a plan is only a sketch, but it should serve to illustrate the manner by which we envision present systems evolving.

Prior to presenting this sketch, it is important to address the possible concern that, in advocating for ordinary measuring, we are deviating from established measurement guidelines. This is an important matter to address, since the act of altering accountability systems, unlike the act of presenting a conceptual hypothetical, has an ethical dimension that demands due diligence.

With regard to the ideas discussed in this essay, the most relevant measurement concept to address is perhaps that of validity. How valid would ordinary measurement be? As discussed by measurement experts like Michael Kane and Saskia Wools, the validity of an instrument must account not only for the precision of its measurement, but also for its functional utility.[47] As already discussed, formal measurement systems are severely limited with regard to their breadth, particularity, and adaptability. They have other strengths, to be sure; yet those strengths come at the expense of features essential to effective functioning in the real world.

Present accountability systems, reliant as they are on hard measuring, and coupled to algorithmic procedures for aggregating summative index scores, also fail to maintain sufficient construct validity. While the measures themselves certainly have good validity where more bounded constructs such as English language arts (ELA) and math achievement are concerned, robustly valid measures of limited domains cannot simply combine into judgments of a larger concept like "school

46. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018); and Naomi Murakawa, *The First Civil Right: How Liberals Built Prison America* (Oxford: Oxford University Press, 2014).

47. Michael T. Kane and Saskia Wools, "Perspectives on the Validity of Classroom Assessments," in *Classroom Assessment and Educational Measurement*, ed. Susan M. Brookhart and James H. McMillan (New York: Routledge, 2020), 11–26.

quality."[48] Ordinary measuring is required in order to ensure the legitimacy of accountability judgments by making sense of the information produced by formal measuring procedures. Validation, of course, is not an activity that occurs once assessments are developed; rather, it is an ongoing process.[49] The approach we suggest is, therefore, one that would structurally incorporate the process of member checking, or participant validation, into the process of accountability.[50] While there is ample reason to be skeptical of relying exclusively on local judgment, there is presently, as Koretz notes, no "adequate substitute" for firsthand knowledge.[51]

In light of such concerns, we suggest a basic plan designed to draw on the complementary capabilities that ordinary measuring has to offer without sacrificing the existing strengths of formal measures. Such an outcome is essential for the purpose of assembling a better picture of school performance — not only for practical uses, but also for the purpose of garnering maximal public acceptance. The components in this rough outline are not intended to upend present accountability systems, but rather are offered in the spirit of enhancing existing systems.

Component 1: Hard Measurement

While it is all too easy to interpret reservations about hard measurement as radical refusal, we intend nothing of the kind. Graduation rates offer insight into the work of schools; so do measures of ELA and math performance. It is important to have consistent, accurate, and unbiased measures of each in order to make informed decisions about how well our schools are doing. The comparability and reliability of such measures allow stakeholders a wide-angle view on how any given school is performing on matters of general concern. Our argument throughout has simply been that these are not, in themselves, sufficient for rendering a comprehensive judgment of any school. Thus, we would continue collecting the hard data that presently constitute formal measurement systems, doing so in a manner, which we discuss, that would reduce the incentive to game these statistics.

Component 2: Soft Measurement

Though this essay has chiefly discussed the importance of multiple *kinds* of measuring, the use of multiple measures is equally important. This is hardly a controversial point — it was, for instance, a major feature of Arne Duncan's NCLB waiver proposals, and many districts now incorporate a wide range of hard and

48. AERA, APA, NCME, *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, 2014), 11–19, 225, https://www.apa.org/science/programs/testing/standards.

49. Samuel Messick, "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning," *American Psychologist* 50, no. 9 (1995): 741.

50. Yvonna S. Lincoln and Egon G. Guba, *Naturalistic Inquiry* (Thousand Oaks, CA: Sage, 1985).

51. Koretz, *The Testing Charade*, 203.

soft measures into their locally issued "scorecards."[52] The aim of doing this is to incorporate multiple measures in order to achieve a kind of multi-perspectivity. In practice, however, the multiplicity of measures is often overstated, reflecting only a single perspective in different forms. First, soft measures are generally excluded or minimized in state accountability systems; that is, they may be included in some form of state- or local-level reporting, but they are not used in the algorithms that render judgment. Second, soft measures are routinely validated by establishing statistical relationships with test score outcomes.[53] Selecting measures in this manner undermines ostensible multiplicity.[54] We endorse the expanded use of evidence like school climate surveys and recommend using them in a less mechanical and more deliberative approach to judgment.

Component 3: School Inspectorates

We are not the first to call for the introduction of observers into the process of school quality assessment.[55] In places like England, the Netherlands, New Zealand, and South Korea, school inspectorate teams are oriented not just toward breadth — the kind of broad view that hard measuring often cannot take — but also toward particularity. The orientation toward particularity requires that inspectorate teams' observational protocols go beyond rubrics or checklists: the very point is the possibility of perceiving evidence or aspects of a school's quality that do not generalize widely. Inspectorates, then, represent important and overlooked sources of information about schools. They are one mechanism by which the process of ordinary measuring might be incorporated into the process of accountability. Our inspectorates would consist of students, educators, administrators, and state officials. The multiplicity of perspectives might be even further enhanced were students and educators to join teams that inspect not only their own schools, but also other local schools. Drawing on their collective understanding, these teams would annually produce a report based on their observations, and that report would stand alongside other measures of school quality for consideration.

Component 4: Deliberative Bodies

Attempting to bring multiple measures together to render a judgment requires a deliberative, rather than an algorithmic, process. Deliberation would include information produced by formal measures, both hard and soft, along with the sort of ordinary measurement produced by an inspectorate's observations. And it

52. Duncan, "Partners in Reform."

53. Thomas J. Kane, Eric S. Taylor, John H. Tyler, and Amy L. Wooten, "Identifying Effective Classroom Practices Using Student Achievement Data," *Journal of Human Resources* 46, no. 3 (2011): 587–613; and Ronald F. Ferguson, "Can Student Surveys Measure Teaching Quality?," *Phi Delta Kappan* 94, no. 3 (2012): 24–28.

54. Koretz, *The Testing Charade*, 234.

55. Helen F. Ladd, "Education Inspectorate Systems in New Zealand and the Netherlands," *Education Finance and Policy* 5, no. 3 (2010): 378–392; and Richard Rothstein, Rebecca Jacobsen, and Tamara Wilder, *Grading Education: Getting Accountability Right* (New York: Teachers College Press, 2008).

would draw on the firsthand knowledge of involved local participants — district officials, building-level administrators, teachers and employees, community members, parents, and students. As in jury selection, the representatives charged with deliberating over school quality would be randomly selected, with membership turning over in a manner similar to that of the U.S. Senate: one-third of the body rotating off each year. To bring this to life, states would create forums at the local level, fostering procedural transparency and allowing for broader community input in decision making. Worth noting here is the fact that some structures for community deliberation already exist. The local school councils of Massachusetts are one example; similar local institutions in Chicago are another. Yet, though these bodies were designed to bring multiple stakeholders together in conversation, they have not yet been called upon to serve in this sort of judgment or planning capacity.

Procedure

In the system we envision, a public process of accountability might unfold over the period of two years. In the first year, accountability teams would work to make determinations about school performance and craft plans for improvement, doing so in a manner consistent with present public policy (that is, through public meetings, public documents, and so on). In the second year, although data would continue to be collected, teams would also work to see their plans enacted. This would address a critique of present accountability systems so far undiscussed in our essay: the fact that educational ratings rarely produce meaningful change in the operation of schools. A team with the mandate to determine school quality would be well-positioned to develop a school improvement plan, which might gain further public acceptance through something like a referendum process.

Resources

Accountability team members would be compensated by state and/or local agencies for their time, much as jurors are compensated. State and/or local agencies would also support the work of these teams by assigning a staff member to assist with procedural matters. Such staff might help arrange public meetings, secure necessary documents, and generally advance the aim of transparency. In short, we envision the state playing an important role in building capacity, ensuring procedural fairness, and generating structures that will support the underlying aim of evidence-based deliberation. Although this might seem a significant reorientation of the state's role, we believe that such work better aligns with the strengths of central offices.

Conclusion

The current preference for formal measures over ordinary ones in assessing school performance is easy to understand from a historical perspective. As education became increasingly important to economic success on both a societal and an individual level over the course of the twentieth century, and as state and federal offices assumed responsibility for ensuring access to high-quality education for all students, student performance and school quality needed to be legible to centralized actors. Formal measures offered ostensibly objective means for comparing

local contexts and for doing so in a rigidly fair manner. Ordinariness, by contrast, was associated with untrustworthy subjectivity. But the tensions that we have pointed out show that formal measures, hard and soft alike, are vulnerable to malfeasance and conceptual distortion, too.

Given all this, we advocate for the inclusion of ordinary measuring in official accountability processes precisely because of the way it complements the aspects of educational quality that formal measures already reveal. We are not suggesting that ordinary measures are a panacea. Rather, our message is that these two approaches to measurement, each of which has inherent limitations, might work best by working together. This combination should not be conducted via an algorithm that mechanically combines results. Instead, it should rely upon deliberative structures, such as local school councils, to take account of all available evidence and produce judgments of educational quality that will be equally useful to local and state-level stakeholders. The state might never see as humans do, but it might see more humanely.