The False and Damaging Premise of School Accountability

By John Tanner

One who is "accountable" provides a complete, objective accounting and then accepts responsibility for the findings. Educators accountable for schools owe such an accounting to their communities and the students they serve, and sound educational policy should support them in doing so. We are, after all, talking about a key institution for the preservation of American democracy. Judging it accurately should be a priority.

But educational accountability occupies an odd space in the American political psyche. Our schools are subjected to a policy approach typically reserved for dangerous industries, whereby we "hold their feet to the fire" because they cannot be trusted to be accountable on their own.[1] The primary objective of such policies is to punish those who fail to comply.

In 1983, the bombshell report *A Nation at Risk* made headlines with its claim that American education was in such a sorry state that, had a foreign country been involved, it would be tantamount to an "act of war."[2] That claim helped institutionalize the idea that education in the U.S. was an abysmal failure. Despite the basis for the report's claims being thoroughly debunked a few years later,[3] the avalanche of distrust it triggered has persevered, justifying the policy approach used since. That approach presumes educators' feet must be held to the fire, along with serial polluters and dangerous industries, and each generation of accountability since has presented an iteration of that thinking.[4]

The tool selected for holding schools accountable is primarily end-of-year test scores in key subject areas, the results of which are then used to pass judgments on schools. Those judged negatively are deemed non-compliers and subject to sanctions.

Policymakers committed two critical errors in creating that system, based on two serious misunderstandings. The first is a misunderstanding regarding test scores. They noticed that averages from standardized tests matched their bias regarding what they considered "good" and "bad" schools, and then codified that misunderstanding into laws that insist no schools should have low test scores.[5] The second is that they presumed it acceptable for a judgment made about a portion of an organization to serve as the basis for judging the entire organization.

Gaping Holes in Understanding

Researchers developed standardized tests to help analyze human characteristics they could not easily observe. Literacy or numeracy within a population, for example, is much trickier to analyze than, say, height. To analyze height a researcher could simply measure the height of a group of people. Or, if no measuring stick existed, a researcher could line people up from the shortest to the tallest and use the relative differences between them as the basis for the analysis.[6] What a lot of people don't realize is that statistics doesn't require knowing how tall anybody is to do its work; it just needs to know the relative positons in the data set.

But some traits, such as literacy and numeracy, lack both a measuring stick *and* the ability to directly observe differences. Test items selected against a very narrow, specific set of statistical criteria form a test that shows the relative positions of test takers to each other: The results will show a continuum from the student who is furthest below average to the student who is furthest above average.

Such information is useful for analysis so long as it is presented alongside other data, i.e., "multiple measures." It is especially useful for signaling patterns. But like any tool its usefulness is limited by its

design, which is why a carpenter has both a hammer and a saw. For example, at no point along the continuum of students do we measure what was learned, so the amount of learning at any point can't be known.[7] Knowing the cause for a student's score also can't be known: Was it good or bad teaching, a good or challenging home life, or something else entirely? None of that can be gleaned from simply knowing a student's position relative to others: Those things need to be understood through tools specific to those purposes. Nevertheless, enamored with the underlying statistics, policymakers selected the methodology as the basis for educational accountability.

Policymakers then compounded that mistake with the presumption that somewhere along the continuum of test scores exists a special score that can signal whether a school is doing a good job. While the student at the top of the continuum can be said to possess a greater amount of the thing being analyzed than the student at the bottom, what cannot be known without looking is what caused either to possess what they do. Perhaps the student at the top acquired what he or she has entirely from educated parents, and the school had nothing to do with it. Or perhaps the student at the bottom came to school from very difficult circumstances and has made significant progress entirely due to a dedicated teacher.

Literacy or numeracy levels can only be understood for complex reasons that lie outside the test score. The same is true for measuring school quality.

Compounding the issue is that such tests reveal a highly specific pattern: Literacy and numeracy, at any age, exist in greater degrees in wealthier populations than in socioeconomically disadvantaged populations.[8] That in no way means that students in the disadvantaged category can never catch up, but rather, that they need the time, support and resources to do so. It also does not automatically mean that schools serving poorer populations are worse than schools serving wealthier populations, which is a bias many people (and policymakers) share. Rather, a quality school is one that knows what its students need and then executes accordingly.

In that vein, high-quality schools that effectively serve low-scoring students, and low-quality schools that underserve high-scoring students surely exist, and a valid accountability program would identify them as such. But a very different result emerges if the patterns of literacy and numeracy are asked to serve an accountability role: The result will correlate with the socioeconomics of the school and ignore the actual work being done.

Policymakers nevertheless chose to assign quality labels to schools and students based on the levels of literacy and numeracy signaled by test scores, not on why that level of literacy or numeracy exists. The result of that decision was easily predictable: Teachers in poorer schools are now highly likely to be judged as ineffective and their schools as failing, while teachers in wealthier schools are now highly likely to be judged as effective and their schools successful, regardless of the underlying reality.[9]

A Little Knowledge is Dangerous

The judgments from end-of-year test scores in key subject areas are then asked to represent the entire school in terms of its quality. That such tests are incapable of signaling why any level of literacy or numeracy exists doesn't seem to matter: Those in a policy positon seem comfortable presuming that good schools cause high test scores, and bad schools cause low test scores, despite their selection of a testing methodology that can signal neither.

But the problem goes much deeper. A judgment of one part of an organization — even if valid — if extended to the entire organization must be viewed as invalid, because it is a judgment made without evidence.

Imagine rising one morning with a goal of tending to a large lawn. To know the proper actions to take, you need an understanding based on a complete accounting: Does it need to be weeded, fertilized, mowed, watered, reseeded or some combination of all the above? Do different spots require a different treatment? Do you need expert advice? With organizations, let alone most lawns, the proper treatment is complex, with different areas requiring different types of attention. Only a complete accounting can address that complexity. Extending judgment from a partial accounting to an entire organization — whether a lawn, a business or a school — risks vastly inappropriate judgments.

In its current form, educational accountability is a partial accounting extended to the entire organization. In fact, end-of-year test scores designed to signal the distribution of literacy or numeracy across a grade level are *always* a partial accounting of literacy and numeracy. When extended to the whole of each domain, and then to the whole of the school, it creates an educational accountability based on a partial accounting of a partial accounting.

Partial accountings create inefficiencies. If a school is judged as failing when it is not, acts accordingly, and changes what is working, it risks harm to the students it serves. If a school is judged as succeeding when it is not, acts accordingly, and does not change, a false message of success is sent at the expense of its students.

The current system of partial school accountability drives inefficiency in schools through "feet-to-the-fire" judgments made prior to having the evidence to pass a valid judgment. Schools are assigned a judgment based on the amount of literacy or numeracy in a school as of a date during the school year, regardless of cause, and then required to act as if that judgment is true.

All this has a particularly chilling effect on schools that serve impoverished students. Students in such schools tend to have developed fewer literacy and numeracy skills as of a moment in time than their wealthier peers, with the primary difference being that the poorer students have had a different experience. A complete and proper accounting would demand that the school demonstrate an understanding of the unique needs of its students, have a plan to serve them and then aggressively execute that plan to its fullest. Those that do should be applauded. Instead, poverty is equated with failure, the school is told to act accordingly, and the children suffer.

An impoverished school is highly likely to be judged negatively year after year.[10] Repeated judgments of that ilk cannot help but create inefficiencies, hurt the most vulnerable of our students and then insist that the process be repeated year after year.

A True Accounting

Lest anyone presume I'm arguing for an easier way, my argument is for *true* accountability. True accountability requires a greater, not a lessor, accounting. Virtually every educator I've had the chance to work with has said they would embrace such an accounting. It would allow them to avoid specious judgments, manage from the truth and more effectively attend to the challenge of educating each child.

Complex issues are often best remedied through simple solutions. That is the case here. First, school leaders should make decisions only from a true and complete accounting and recognize the harm to a student and a school caused when a partial accounting pretends to represent the truth. Second, recognize that state test scores have a surprisingly limited amount of interpretive power, which never included judging school quality. Judgments of school quality must be made, but they should be made based on evidence capable of rendering that judgment.

Finally, don't wait for a better set of educational policies. You don't need them to do what is right, and students don't have that kind of time.

Footnotes:

---

[1] As but one example, the U.S. Department of Labor has a summary called "History of Mine Safety and Health Legislation" that can be viewed online at https://arlweb.msha.gov/mshainfo/mshainf2.htm. Similar histories exist for manufacturing, the steel industry, construction, power plant emissions, etc. In each case the idea is to force industries to protect citizens through compliance. It is strange that educational accountability policy continues to follow this pattern, which treats educators in a similar vein.

[2] *A Nation at Ris*k in 1983. (Gardner, Richard, et.al., *A Nation at Risk: The Imperative for Educational Reform* (Washington D.C.: The National Commission on Excellence in Education, 1983).)

[3] Most of the claims for *A Nation at Risk* were shown to be inaccurate in a report out of the Sandia Labs prepared in 1990, but the report was quashed given that it did not match the political rhetoric that schools were failing miserably. Thanks to technology the report has now seen the light of day. While highly technical, a highly readable summary can be found at https://www.edutopia.org/landmark-education-report-nation-risk.

[4] The most recent iteration is school grading systems. Started in Florida, the approach is now in seventeen states, with more considering it.

[5] While the idea of offering credit for "growth" is now the norm in state educational statutes, the underlying assumption is still that no students should have low test scores, which represents the end goal.

[6] Conducting this simple exercise is enlightening: ask a group of people to line themselves up from the shortest to the tallest, and having done so ask several questions: where is average? Who is the furthest below average? Above average? What patterns do you see in the data (e.g., men tend to be taller than women), and how stable do you think average will be? What none will be able to answer without looking beyond where each stands: why anyone is at that point in the scale. That question can only be answered from other information.

[7] This, of course, appears to the naked eye to be counterintuitive. After all, the items look like what was taught, and some number of them stacked into a pile look like they should be able to tell us something about the amount that was learned, like a measuring tape. But that is not the case with items selected for their ability to discriminate between a below and an above average student. To do that requires items that about half the students will answer incorrectly, so they can divide students into a below and an above average pile. An item selected to evaluate whether a child has learned something would be doing its job if everyone answered it correctly if all the students learned the material, but if they did it would not make it into a state test. A teacher trying to understand what was learned, and a researcher trying to understand the distribution of literacy or numeracy across a population require tests and items unique to their purpose, and substituting one for the other is always a mistake.

[8] The 2015 National Assessment of Educational Progress again confirmed this trend across the nation and in every state where data are available. The score tables at https://www.nationsreportcard.gov/reading_math_2015/ show this quite clearly.

[9] See, for example, Adams, Curt. M. et. al, J. "An empirical test of Oklahoma's A-F grades." Education Policy Analysis Archives, 24(4).

[10] In Georgia, a review of the Chronically Failing Schools list at https://gosa.georgia.gov/sites/gosa.georgia.gov/files/related_files/site_page/Chronically%20Failing%20

Schools%20List%2001052017.pdf compared with the levels of poverty in each school shows this trend to be alive and well.

John Tanner is a San Antonio, Texas-based educational writer and consultant specializing in educational structures and systems. His book, "The Pitfalls of Reform," outlines how school accountability efforts appear rational from a policy perspective but have little to do with educational excellence, student aspirations or equity. Tanner is the executive director of Test Sense, an educational consulting firm, and a co-director of the Texas Performance Assessment Consortium, a project in which 44 school districts have joined forces to build a community-based accountability system for their schools.