The Tests Are Lousy, So How Could the Scores Be Meaningful?

Alfie Kohn

February 2, 2021

Accountability and Testing Assessment High-Stakes Testing and Evaluation Standards-Based Reform

Way before Waze, I listened to traffic reports on my car radio whenever I had to drive to the airport. Over time I came to realize that the information they broadcast was frequently wrong: I'd be caught in a jam on a highway that the radio had just told me was in good shape or, conversely, I'd be speeding along on a road I'd been warned about. So why did I find myself tuning in again on my next trip? Similarly, why do I keep checking weather forecasts to help me decide when to take a walk with a friend…even though I've learned that these forecasts, too, are not very reliable[1]?

In short, am I really so addicted to data that I prefer misleading information to none at all?

This *mea culpa* is my way of reminding myself not to be judgmental about the following story, which I have told during lectures: It seems a group of teachers at a school in Florida were sitting in the faculty lounge, exchanging tales of woe about the state's standardized testing regimen, grimly reviewing the ways it demoralized their students and impeded learning. As they were commiserating, the P.A. speaker suddenly crackled to life with an announcement from the principal: The school's official test results from last year had just been released and they were much better than the prior year's.

Can you guess what the teachers did then? They burst into applause.

I emitted a deep guttural groan when I first heard that story. After all, these teachers presumably had seen just how dumb some of the test questions are. They could rattle off the names of impressive kids whose intellectual gifts don't show up on standardized tests. They had witnessed how test performance largely reflects circumstantial factors and children's socioeconomic status. Surely, then, they knew better than to treat the final scores as if they were meaningful….even if those scores happened to make them or their school look good this time.

Standardized tests are so poorly constructed that low scores are nothing to be ashamed of — and, just as important, high scores are nothing to be proud of. The fact that an evaluation is numerical and the scoring is done by a computer doesn't make the result "objective" or scientific. Nor should it privilege those results over a teacher's first-hand, up-close knowledge of which students are flourishing and which are struggling.

Sadly, though, some educators have indeed come to trust test scores more than their own judgment. One hears about parents who ask a teacher about problems their child is having in school, only to have the teacher reach into a desk and fish out the student's test results. Somewhere along the way such teachers have come to discount their own impressions of students, formed and reformed through months of observation and interaction. Instead, they defer to the results of a one-shot, high-pressure, machine-scored exam, attributing almost magical properties to the official numbers even when they know those exams are terrible.

I understand the cognitive dissonance that might lead someone to divorce the output from the input, the impulse to treat whatever the machine spits out as if it were somehow imbued with significance regardless of the dubious process on which it's based. But however understandable that impulse is, we have a duty to resist it, at least when it can do real harm. For example, we should think twice about citing education studies in which standardized test results are used as markers for achievement — even when a study seems to justify practices we like or to indict those we oppose.

I say this, first, because these tests measure what matters least about learning. Research has found a statistically significant *negative* correlation between deep thinking and high scores on several such tests.[2] The implication is that a high aggregate score (for a school, district, state, or country) may not be simply meaningless but a reason for concern. Add to that the pressure to raise scores, which often truncates, flattens, or otherwise dumbs down the curriculum, particularly for our most vulnerable students.

Second, every time a study that relies on test scores as the primary dependent variable is published or cited, those tests gain further legitimacy. If we're not keen on bolstering their reputation and perpetuating their use, we would want to avoid relying on them. If we treat these scores as if they *were* meaningful — which, of course, is also the implication of cheering when they make our schools look good — we help to confer respectability on them and thereby contribute to dooming more students to their damaging effects.

For anyone who's unfamiliar with what I've reported here about the inherent unreliability of standardized tests, allow me to suggest dipping into some additional resources on the topic — as well as the considerable literature on more informative and less destructive alternatives for evaluating teaching and learning. (Anyone who claims that standardized tests are necessary for those purposes is just revealing his or her ignorance of the whole field of authentic assessment.)

And for those who already know all this, well, the question is why these folks continue to invoke or celebrate test scores. Maybe the informed part of us needs to keep a watchful eye on the part of us that's a sucker for an easy-to-summarize, precise-sounding indicator of educational success.

NOTES

1. Of course there are multiple variables to be considered in evaluating the reliability of weather forecasts, starting with the time period in question. Forecasts more than three or four days in advance are virtually worthless; they get better as the target day approaches, but it's striking how often even those for tomorrow turn out to be wrong. Accuracy also varies depending on whether we're talking about temperature or precipitation. If the former, how broad a range are we looking for? (A forecast of "in the 50s" obviously is more likely to prove accurate than one specifying a high of "56 degrees.") With precipitation, it's harder to predict how much will fall than whether there will be any at all. Then there's the size of the area: Are we interested in whether it will rain somewhere in your county or whether (let alone when) it will rain in your neighborhood? (Two things I learned from recent reading: The day's low temperature is harder to predict than the high, and there are significant regional variations in forecast accuracy.)

During the pandemic I've amused myself by paying closer attention to weather forecasts, and I came to see just how unreliable they are. On Tuesday, we're told rain is likely on Wednesday afternoon; by Wednesday morning, the story has changed. And if you really want to become a skeptic, try comparing forecasts from multiple services. A couple of weeks ago, I wanted to know what the probability of rain was for various times the following day in my town. Here's what I found online at the same moment:

	10 am	11 am	Noon	4 pm
Weather.com	25	35	25	75
Accuweather	16	51	47	39
NWS/NOAA	50	42	60	31

Notice that the services can't even agree on whether rain becomes more or less likely as the day goes on, much less on approximate probabilities. Again, this is for the very next day.

Are such results just a function of my limited sample size or perhaps a sign of my cognitive biases (a tendency to notice or remember inaccurate forecasts and forget the ones that turned out right)? Apparently not. Systematically recorded data from people who do this for a living corroborate my observations, at least for the National Weather Service and at least for my city: Over the course of an entire year, high-temperature forecasts for one to three days in advance were accurate within three degrees only about half the time. And even binary forecasts of whether or not there would be any precipitation were wrong one out of five times, so you can only imagine how poor the record is for *amount* of precipitation. (I happen to be writing this during a week when six to twelve inches of snow was predicted for the following day and we ended up with an inch or two.)

2. One such study classified elementary school students as "actively" engaged in learning if they went back over things they didn't understand, asked questions of themselves as they read, and tried to connect what they were doing to what they had already learned. Students were classified as "superficially" engaged if they just copied down answers, guessed a lot, and skipped the hard parts. It turned out that the superficial style was positively correlated with high scores on the Comprehensive Tests of Basic Skills (CTBS) and the Metropolitan Achievement Test (MAT). Similar findings have emerged from studies of middle school and

high school students, using different tests. Of course these correlations, while statistically significant, are not absolute correspondences. Many students think deeply and score well on tests, while many others do neither. But, as a rule, better standardized exam results are more likely to go hand-in-hand with a shallow approach to learning than with deep understanding. (For more details and citations to the original studies, please see my book *The Case Against Standardized Testing*.)