What standardized testing is and all the things it is not

John Tanner

Feb 2024

What follows in the next nine pages is an attempt to right more than a century of misunderstanding when it comes to America's long held obsession with what is commonly referred to as standardized testing. I cannot fathom any parallel in which time has been so ungenerous in preserving misunderstandings, with ever increasing negative consequences as the years have gone by.

What follows is not an anti-standardized testing screed. Neither myself, nor the organization I founded and lead, have ever taken an anti-test position. Standardized testing is based upon a novel approach for analyzing human traits that cannot be directly observed or measured and is something of a remarkable accomplishment in that sense, but from its invention it has been rife with misuse and misunderstanding.

What follows is my clearest explanation yet of how this research tool works. It is imperative that the world come to terms with what it is, so we can finally stop asking it to be what it is not. For more than a century this very complex research methodology has been presumed to say things it cannot, with surprisingly few opportunities or even efforts at correction. Perhaps 2024 will be different.

To get started, please put on hold everything you think you know about standardized testing. Pretend that you know nothing and that you've never even heard of the thing. Starting with a blank slate, with no preconceived notions, is helpful.

Next, realize that calling these things standardized tests is a bit of a misnomer. Standardization just means that you've created the same conditions for all test takers that allow the results to be compared. You couldn't compare results on a chemistry test if one set of students had access to scientific calculators and one did not, because in one case you were testing the ability to compute by hand, and in the other you were not. Leveling the starting point (i.e., either everyone has a calculator, or no one does), which standardization does, is a prerequisite to comparisons.

Any test, quiz, or assessment can be administered under standardized conditions, so we need to be clear about which tests administered under standardized conditions we're talking about here. Specifically, we're talking about any test that produces what I'll call *anticipatory results*. That just means that we can expect the results to be consistent over time and across administrations. Researchers are then able to use the results, alongside other research, to gather a sense of patterns among tested populations and whether those patterns are changing over time.

For the sake of clarity, I'm going to use the more descriptive label of *predictive testing* to make it clear that we're talking about the narrow range of standardized tests designed for this predictive or anticipatory purpose. That includes your state testing program (by Federal law) as well as any norm-referenced commercially available test (e.g., the Iowa Test of Basic Skills, NWEA's MAP test, or the Stanford Achievement Test).

Finally, realize that one of the greatest errors in educational policy is thinking the methodology can serve as the basis for school accountability.

Imagine a scale

Imagine that you'd like to analyze a human trait that is impossible to measure directly, like funniness. To do that, envision a scale, or a number line, but without any numbers on it, that stretches out in either direction further than the eye can see. Imagine a continuum of funniness that stretches from the least funny person to the funniest. That is a remarkable thing to imagine, since you are imagining a continuum of a thing for which no measurement is possible. There is, after all, no device that a person can blow into into that will register how funny anybody is.

Imagining such a thing is possible because of our capacity as humans to observe the relative differences between people and declare, in a reasonably reliable fashion, when one person has a bit more or less of a trait, like funniness, than another.

Let's start just such a scale from scratch to see how it would work. We would start by finding two random people and determining who is the funnier one. Once that is established, we have enough information to position the funnier person at a point on the scale and the less funny person at a relative point, in this case a lower point.

Consider how arbitrary that is and how little information was required to position those two people. We don't know if in the end both people will wind up towards the top or the bottom of the scale. We don't know how far apart they are. We don't know if in the final scaling they will land right next to each other, or if one will occupy the position furthest above and one the position furthest below average. We have no clue how much of the funny trait either of them has. All that can be known is that the two people are different from each other.

Imagine a third person comes along, and we observe that this person seems to be between the first two when it comes to being funny, so that's where we put them. The fourth person is truly hilarious, so we put them above the others, and on and on we go. As we add more and more people to the mix the scale of funniness would start to fill itself out. We'd have to adjust as we went, making space for new people where we may have squeezed too many in at first, or sliding some closer to others to narrow out the gaps.

As we continued to add people, we would quickly reach a point where new people were so similar in being funny (or not) to people already in the ordering that new steps aren't meaningful, so multiple people would start to appear at various points, with more towards the middle than the extremes. Once we get to several thousand random people, what we would be able to see is a scale of the relative differences of funniness without having to do the impossible and determine how much funny anyone possesses. That's the level of ingenuity required if you want to analyze something that cannot be directly measured.

Before we could begin any sort of analyses using this scale, however, we would need to acknowledge the imperfections in what we just did so we don't mis or over-interpret things. I'll mention two big ones at this point.

The first is that we aren't observing the underlying reality, but rather, our best guess at what that reality might look like. If a breathalyzer test for the level of funniness did exist, and we gave it to all the people in our ordering based on *relative* differences, the *actual* differences between people would produce a different ordering. That shouldn't be a surprise. Our ordering based on relative differences was limited to a moment in time, it would involve human judgments, some of those judgments would represent close calls that could go one way or another, and we're human and not infallible.

But as the breathalyzer for funniness does not exist, we will always and forever only have the scale based on relative differences. That will require an admission on our part that what we have will be different than a scale based on reality, with the added caveat that we will never know exactly where the errors in our scale might be since it would be impossible to compare our scale with one that does not and will never exist.

A second big imperfection is that if you repeat the scaling exercise the very next day with the exact same people and processes you will get a different result. Observable expressions of funniness aren't going to be perfectly consistent from one day to the next, and the capacity of observers to order people from one day to the next will shift as well. You can be sure that some of the arguments about close calls from day one will end up with a different outcome, and that issue of not being infallible will again be there. In the end, the best you can hope for is to be close.

These imperfections don't invalidate a scale based on relative differences. Rather, acknowledging them helps remind us that what is being observed is a thing that is all but guaranteed to be different than the underlying reality that we cannot directly access.

Now, why would we bother creating that sort of imperfect scale, knowing it would be flawed at the outset? Because studying funniness and other human traits that cannot be directly measured is often a worthwhile thing to do. Imperfect scales never lose their imperfections, but they can still create opportunities to observe things about traits we cannot measure.

We could, for example, bring gender, socioeconomics, geography, and life expectancy to the funniness scale, which will enable us to search for broad patterns and opportunities for further investigation. Perhaps we see that men tend to be less funny than women. Perhaps we discover that those in poverty tend be less funny, and yet the funnier you are the longer you live. Perhaps we observe that some states are seriously behind when it comes to being funny.

Once we get a sense of the patterns, we could then verify them with other research, so we don't make bad assumptions, and then begin the search for causes. Why is it that the patterns might exist? What might have caused pockets of funniness, or a lack of it, and are there lessons to be learned? If a pattern is bothersome or negative and a cause can be identified, are there policies that can be put into place that might be able to correct for the negative effects?

Let's imagine we observed some patterns that we needed to make right and set about righting them. How might we see progress over time? By repeating the scaling exercise. If we conduct a similar experiment a few years out and see a similar patterning, we add that to our body of research and perhaps presume that whatever we tried to do had little to no effect. But if some of the negative patterns dissipated, we may be able to infer that our policies had an effect, but we'll still need to investigate to be sure that is the case.

And we could do all that without knowing how much funniness any person possesses. That's remarkable when you think about it. The smallest piece of information that we can use to analyze complex things about humans that are impossible to measure is to know how each human stacks up against all the other humans. The resulting scale will be imperfect—nothing you can do about that—but so long as the imperfections are noted the scale can be useful when no other options are available.

That means that we need to constantly remind ourselves about the limitations in a methodology designed to scale relative differences, as well as its imperfections, to avoid over or mis-interpreting things:

1. The point is to organize a trait along a scale based on observations of relative differences between people.
2. It is most useful when the trait cannot be directly measured. If a thing can be directly measured (e.g., height, weight, near-sightedness), the direct measure is preferred.
3. We will never know the amount of the trait possessed by anybody. That is not a part of a scale based on relative differences.
4. We'll need to acknowledge the imperfections. We'll need to acknowledge that repeating the scaling exercise on consecutive days will produce different results, and that both of those orderings will differ from an ordering that would be produced if we could directly measure the thing, which we cannot. As a result, we are guaranteed to have some amount of error in our work, but we will never be able to say where along the scale the error occurred.
5. We can't make a judgment based only on the patterns in the scale. We don't have the information to do that. Judgments are only possible after lots more research.
6. We can't draw a line in the sand anywhere along the scale and say, there it is, there's the funny line. Everyone above this line is hilarious, and everyone below is dull. Any line will be invalid and meaningless because of how often the additional research would prove us wrong.
7. The scale is about the trait, not the people on it. Each point along the scale contains some number of people who appear to have a similar amount of the trait, but the point is to understand the trait.
8. New people can be analyzed using the scale to determine the relative amount of the trait they possess. But what that means must be determined elsewhere.

Predictive Testing

We are now ready to tackle *predictive testing*.

One day in the late 1800s a researcher in France named Alfred Binet thought it would be worthwhile to analyze the relative differences of intelligence in children. (Quick note: intelligence for Binet was a way to summarize or state how intellectually capable someone was at a moment in time, but would, like being funny, most definitely change over time, and one of Binet's goals was to change it for the better.) He wanted to order students from the least intelligent student to the most intelligent student to investigate patterns in learning and remediation.

You can imagine the challenges that would present. The equivalent of a breathalyzer does not and has never existed that can signal the amount of intelligence or learning possessed by anyone, which is why Binet knew the best he could do was observe relative differences. This presented another problem: trying to ascertain relative differences across thousands of children on something that cannot be easily or reliably observed would present a logistical challenge here in 2024, let alone at the end of the nineteenth century. Binet needed some way to observe relative differences and scale them that was far more efficient than interviewing one student at a time and then trying to figure out who had a bit more and who a bit less of the trait, and how to position that relative amount of the trait along a scale.

Binet conceived a way to do just that. He found if he could ask students a series of increasingly difficult questions, starting with the easiest and progressing towards the hardest, eventually students would not be able to answer any more correctly, and that point, or step, could stand in as his observation.

For making the test easy to administer on paper to thousands of students, Binet could take the thirty or so questions that represented the scale, mix them up, and give them to all students. Doing so would still create an observation, but it would be a bit circuitous to get to it. When the test was scored it would be like unwinding the questions into their original ordering to see what step the student landed on.

It wasn't perfect by any means. It was an estimate. Students will guess and answer some questions correctly, or slip up where they know something, Students will perform a bit differently from one day to the next, and the underlying realities that cannot be accessed will most certainly be different than what is being observed via the test. But for a researcher trying to analyze something that cannot be directly measured, it was more than a bit helpful.

How do you come up with that sort of test? How do you find a set of questions in which each step along the scale is signaled by the questions to that point that should have been answered correctly, and those past that point that should be missed? That is the question from which the field of psychometrics was born. Binet was one of the first, but then statisticians and psychologists put his work on steroids (and, I think Binet would say, in many cases corrupted his initial intent from what he was trying to accomplish), and they've never looked back.

By the 1920s the Stanford Achievement Test was available for use in schools, college admissions were being determined more and more by the methodology, and the American IQ obsession was in full swing. Fast forward to our more recent history: in 1994 the Clinton administration required all schools to use it in three different grades. In 2001 the Bush administration extended that requirement to lots more grades and subjects, and the Obama administration later left that requirement largely intact.

Part of the answer to the questions about how you build such a test is that you impose a whole series of artificial contrivances that render this sort of test and the questions that comprise it useless for anything other than observing relative differences in the amount of a trait. This is a surprise to a great many, who attempt to use predictive test scores to judge schools and students, drive detailed curricular decisions, identify effective teachers, etc., none of which is there to be had.

The misunderstandings stem from the fact that teachers assess learning, and predictive test makers assess patterns, via questions. So, why not, the prevailing logic suggests, use the questions from a predictive test to guide classroom efforts? Why not use them to judge what teachers and schools do? Why not use the points on the scale as a judgment tool? After all, aren't all questions just questions and available to be interpreted in a variety of ways?

No they are not. Not even close. And it has to do with how a predictive test selects and limits itself to the questions that can create that scale.

Let's imagine creating a scale based on literacy. We would have to rely on relative differences to do that as no breathalyzer exists that can measure the amount of literacy possessed by any student. And since the logistical challenge of staring at two students and deciding who has more of the trait that the other would prove impossible to carry out at scale, we would in turn need to follow Binet's lead and use the predictive testing methodology.

We can imagine starting such a scale very similar to how we started the scale of funniness. If we imagine three students and use our eyes to initially observe their literacy skills that would work. Let's say student #1 shows minimal literacy ability, student #2 shows a fair amount, and student #3 three shows a great deal.

In your mind, position them along a scale relative to each other, which would order them 1, 2, and 3. We can now begin to imagine the characteristics of some questions that could be asked that would also position them in that same order. Let's start with students #1 and #3, those at the two extremes. A question that would reflect their positions is one that student #1 would get wrong, and student #3 would get right. That would correspond to our sense that student #3 has more of the trait of literacy than student #1. But, if the question we chose showed the opposite pattern, which could occur for a ton of reasons, it could not be used here. Maybe student #1 has a passion for what the question asked, but student #3 had never heard of it. Maybe it's due to curricular differences, or great teaching on the part of someone or poor teaching on the part of another.

Maybe, maybe, maybe. It doesn't matter. We couldn't use that question. It would be telling us the opposite of what we need to see to get the ordering right. It isn't going to aide us in the identification of an obvious difference. It may be a perfectly fine question for lots of other reasons, like understanding effective teaching or the degree to which students learned, but here, when building a predictive test, only one thing matters: questions must reveal the relative differences between students and not send conflicting messages. For that reason it must go.

Now let's bring in student #2, who according to our eyeball observations should be slotted between students #1 and #3. We would need two questions to slot this student appropriately. We would first ask student #2 the same question as the first two students. The student will answer it right or wrong. Let's say for our example that student #2 answered it correctly. Students #2 and #3 now are at the upper point.

Student #2 would at that moment be wrongly placed. Student #2 should be *between* students #1 and #3 based on our initial observations but isn't yet. At present, we can't observe any difference between student #2 and #3 and yet we know one is there.

Adding a second question and having all three students answer it would help us identify the difference, but *only* if it places student #2 between the other two students. To do that, it needs to be a question that *both* student #1 and student #2 will get wrong. That would leave student #1 with none correct at the lowest point, which is accurate, Student #2 with one correct at the middle point which is accurate, and student #2 with two correct at the upper point, also accurate.

If the second question failed to position student #2 where we expect them to be we couldn't use it. If student #2 answered the second question correctly, student #2 and student #3 will both still be at the top with two correct answers, which means student #2 is still misclassified. So too if student #2 missed the second question as we expected them to do, but student #1 answered it correctly. In that case, #1 and #2 would both be at the lowest point with one correct answer, again, misclassifying student #2. That question would not be contributing to our ability to observe obvious relative differences between the students and so we would need to exclude it.

We could continue this with a third question and a fourth student, but if you think that adding a second question was potentially confusing, the third question is exponentially more so. And then there's the twentieth, and the thirtieth, and…

Most predictive tests used today have 40-50 questions, and in the case of state testing they reference a set of state content. The process of creating predictive tests is so complex that test makers know they will never get the steps perfect, but they need to get as close as possible. If they can observe that each point along the scale is reasonably well identified by looking in one direction and seeing that students tended to answer those questions correctly, and the other way and seeing that those questions tended to be mostly missed, that is as good as it is likely to get.

Doing that well requires an extensive capacity with statistics and psychometrics (and just for the record, I am dramatically over-simplifying what they do for the purpose of creating an understanding of a highly complicated undertaking). To answer 10 questions correctly on a 40-question predictive test means that the test maker will need to limit the number of avenues to get 10 correct so that the corresponding point on the scale can be said to suggest that those that answer 10 correctly most likely have a bit more of the trait than those to their left who answered 9, and a bit less than those to their right who answered 11. That doesn't seem overwhelming until the math kicks in—there are 847 million paths from within those 40 questions to get ten correct!

(If that sounds unrealistic, just remember that in the Powerball lottery, which picks five white balls numbered 1-69, and 1 red ball numbered 1-26, 292,201,338 combinations of winning numbers are possible each time a drawing occurs. Or that of the 52 cards in a deck, there are 8 x 10^{67} (that's 10 with 67 zeros to follow) possible outcomes for a shuffle. These sorts of numbers get huge very quickly.)

If there are even a million ways for a student to arrive at a score of 10, that point on the scale would be meaningless. It would not say anything about what it means to be at a 10 relative to any other point on the scale. No one at that point on the scale could be said to be like the others at that point, and as a result, that point could not be declared to be relatively different than any other point along the scale. In the end, you would have an unanalyzable, unusable data set that would need to be scrapped.

What predictive test makers have become adept at doing is finding those 40-50 questions that when answered, suggest a point along a scale that is at least somewhat unique from the points on either side. Which in turn makes for a decent pattern recognition tool.

Let's stop for a moment and recognize what an accomplishment all this is. Remember, there is no breathalyzer that can signal the amount of learning any student possesses. Let's repeat that so we never lose sight of it: *there is no such thing as a measuring device for how much students know, for how literate or numerate they may be, or for how much anyone has learned*. None.

And yet someone invented two methodologies that would enable an analysis anyway. One of those had to do with how to create a scale based on relative differences, and the other had to do with making those observations via a predictive test.

While that is remarkable, it comes with a host of limitations and imperfections we'd best not forget.

At best, we will only ever be able to observe the trait being assessed via the relative differences between students. That introduces one set of imperfections because it means we will always and only be viewing something a bit askew from the underlying reality.

Additional imperfections are then introduced because we are relying on test questions as the means to our observations of relative differences. Some imperfections will occur as an artifact of selecting one set of questions over another. Some occur because of the technical processes required to build a predictive test and the fact that such an instrument is always less than perfect given the challenges in building it. And some are due to fact that students must answer questions that comprise the observations, and students will never do so in a perfectly consistent fashion.

All that can be said is that the scale of the relative amounts of literacy possessed at each point along it will be an imperfect approximation at best of an underlying reality that will most certainly be different from what we can observe, so be careful.

Barometers

The question selection process, based as it is in the relative amounts of the trait at each point, and though less than perfect, nevertheless produces an instrument that is responsive to shifts in the trait, rather like a barometer regarding air pressure.

If, for example, the United States experienced a surge in literacy, odds are it would be reflected as a pattern change and detected by predictive test scores. Why the change occurred, how it occurred, and what it means would have to be investigated, but if a predictive test showed that sort of shift it would be worth investigating.

Or think about a school. If a school adopts programs and practices designed to prioritize literacy which are embraced by the faculty and the students, the result is likely to show that the students will have an increased amount of the trait. Given the uncertainty in a predicative test and its distance from the underlying realities, the results should be interpreted carefully and alongside other information, but as part of a body of research that information can be useful.

Predictive testing is like a barometer in other ways as well. A barometer provides an indicator of atmospheric pressure, but what that means requires professional meteorologists to interpret alongside thermometers, anemometers, maps, computer models, etc. A barometer can only say whether atmospheric pressure is high or low or rising of falling. It has no interpretive capacity regarding what it reveals. It has nothing to offer about other aspects of the weather, like temperature, wind speed, humidity, etc. We could attempt to infer those things, but we would always be wrong. We can accurately say—without criticism—that a barometer is useless except for doing the one thing it was designed to do.

Predictive testing provides an estimate of the relative differences of a trait in a population such as numeracy and literacy. Like a barometer, it is useless for everything else. It has no interpretive capacity regarding the trait or those who possess some amount of it.

It cannot be used, for example, to make instructional decisions (though it often mistakenly is). Questions that might have been useful to guide instruction were tossed because they rarely react in predictable ways, and trying to glean instructional tidbits from the questions that remain would be foolish, misleading, and counterproductive.

If raising scores is considered important by a school, the most efficient way to do that is to focus on the trait, not the instrument. *The instrument was designed to reflect the trait.* It would be a silly and illogical thing to redesign the trait to reflect the instrument.

Just as the reading on a barometer offers a benign number, so too does a predictive test. A score from a predictive test does not contain any information from which a judgment can be rendered. The pressure shown by a barometer is high or low, or falling or rising, but that only obtains meaning in a larger interpretive context. Predictive test scores are relatively high or low, or rising or falling, and that too only obtains meaning in a larger interpretive context.

In fact, it is invalid even to say that a student got a good score or a bad score, or did well or poorly on a predictive test, as those are judgments being made prior to an understanding of what caused them. Perhaps something good or bad has happened that may need to be judged, or perhaps not, but that can only be known after additional research is done.

High and low scores certainly exist, and there are advantages and disadvantages given the amount of the trait students possess as of a moment in time, but it is the cause of those advantages and disadvantages that needs to make itself available for judgment, not the advantage or the disadvantage.

And like a barometer, predictive testing would make for an illogical accountability tool. Its imprecisions and its distance from the underlying realities of the trait being examined notwithstanding, predictive test scores don't offer any sort of chance at an accounting. They cannot account for what happened in a school or on a student's learning journey. At best they offer one reflection of a current state that, like a barometer reading, needs to be considered by those good at interpreting such things.

Drawing a line at some point in the scale and declaring that enough of the trait to indicate success exists at that point would be illogical at every level.

The amount of the trait at any point cannot be known, so declarations of "enough" would be deeply problematic. Students will be estimated to be at a position relative to their peers for any number of reasons that such a line ignores. Some of those reasons would be worthy of judgment, and some will not, so to mush them all together and make a declaration would be foolhardy.

And so many students and schools would be misclassified as effective or not that line would be meaningless. Success or failure are likely to be found across the entire scale but can only be identified through other means.

Accountability is supposed to tell the truth about an organization, about where it is effective and where it has challenges. Predictive testing, whether as of a moment in time or over time, was never designed to do that.

Predictive testing can signal patterns and shifts in underlying academic traits. That is all it was ever designed to be.

Of course, that isn't how we use it at all.

Errors and misunderstandings

Predictive testing is ubiquitous in American society and has been for more than a century, as have the misunderstandings that have almost always surrounded it.

Right from the start the interpretations from the results went wrong given that the results appeared so consistent over time. The consistency, however, was just an artifact of viewing a trait through a scale. If you build a scale that stretches from the least of a human trait to the most, somewhere in the middle you will observe average. And the nature of "average" in a scale is that lots of people will tend to cluster around that point, stabilizing it if you will.

So, when the ordering exercise is repeated a year later, odds are the average position will be about the same, and so too will people's relative positions to it. If a person is a few steps above average in terms of being funny this year and nothing changes, odds are, we can predict, with a reasonable degree of accuracy, the person will probably still be a few steps above average the next time we check.

That isn't magic at all. But it risks looking like magic. The consistency that happens when a single trait is scaled created the risk right at the outset that people might believe the scores were signaling far more than the relative differences regarding a human trait. Which is exactly what happened. A useful but imperfect invention appeared to the naked eye to say what it could not and almost immediately it was if the imperfections did not exist. The methodology has been rife with misinterpretations and misunderstandings since.

Why this methodology has long been preferred by policy makers in America seems to me embarrassingly simple and naive: the schools they perceived to be good schools *consistently* had high predictive test scores and so all schools should have high predictive test scores, and therefore all schools would be good.

The technical issues on that alone are enough to doom it as a rational policy. If students are positioned along a scale from the student with the least of the trait to the student with the most, high test scores for all will be an impossibility. Any policy that says otherwise is being disingenuous, because success will have to look a lot like everyone being above average.

But traits such as numeracy and literacy complicate that further because they occur for a variety of reasons, most notably as a combination of what happens in school and what happens outside school. The time spent learning *in* school is comparable across students and communities, but the time spent learning *outside* school differs dramatically. And while the effects of that time in school will differ, sometimes by a lot, those differences pale in comparison to the differences that occur outside school. Students in wealthier communities are highly likely to have significantly more non-school learning than students in poorer communities.

What policy makers were doing when they picked predictive testing as their preferred tool was ignore the fact that it was presenting to them a big picture of the issues in American society that needed to be corrected. Declaring predictive testing as a signal of effectiveness allowed policy makers not to have to address reality because they could instead treat schools as the cause of the problems they were unwilling or unable to address. That continues to this day.

Also complicating things is the desire to move away from old titles for predictive testing, like standardized tests, or norm-referenced tests. States have come up with any number of new names that cause people to think they've moved to a new methodology more capable of meeting policy goals than the old one, criterion referenced or standards-based being two common ones.

But don't let the names fool you, because at their heart they all rely on Binet's underlying methodology and are thus subject to the limitations and imperfections I've described. Criterion referenced just means someone drew a line in the sand and assigned labels on either side, pass/fail being the simplest and yet also the most inappropriate and inaccurate. Standards-based just means that the relative differences that are being identified are relative to a specific body of content, but the tests still aren't a measure of how much or even what was learned from within that body of content. They are showing the relative differences across a population relative to that content.

No matter what we call the resulting tests (or how often we do them—the latest marketing ploy by testing companies being to do them multiple times a year) all the limitations apply to these results that applied to our hypothetical scale of funniness.

Let me repeat the limitations from earlier but consider them here in an educational context.

1. The point is to organize a trait along a scale based on observations of relative differences between students.
2. It is most useful when the trait cannot be directly measured.
3. We will never know the amount of the trait possessed by anybody. That is not a part of a scale based on relative differences.
4. We'll need to acknowledge the imperfections. We'll need to acknowledge that repeating the scaling exercise on consecutive days will produce different results, and that both of those orderings will differ from an ordering that would be produced if we could directly measure the thing, which we cannot. As a result, we need to acknowledge some amount of error in our work, but we will never be able to say where along the scale the error occurred.
5. The scales produced via a predictive test contain additional imperfections in that the observations are obtained via student responses to carefully constructed questions that further remove the results away from what would be seen in the underlying reality. That means that the estimates will need to be viewed as being not entirely accurate, and any estimate about a student even less so. And no analyses should be done, nor conclusions reached, absent lots of other evidence and information.
6. We can't make a judgment based only on the patterns in the scale. We don't have the information to do that. Judgments are only possible after lots more research.
7. We can't draw a line in the sand and say, there it is, there's the passing line. Everyone above this line demonstrates effectiveness, and everyone below demonstrates failure. Any line will be invalid and meaningless because of how often we'd be wrong.
8. The scale is about the trait, not the people on it. Each point along the scale contains some number of people who appear to have a similar amount of the trait, but the point is to understand the trait.
9. New people can be analyzed using the scale to determine the relative amount of the trait they possess. But what that means must be determined elsewhere.

Despite these limitations, the analyses that can be done from scaling a trait based on relative differences via predictive testing can still be useful. For example, we could pull in demographics and other information and observe the patterns that emerged during COVID and then make additional observations over time as we attempted to remedy any negative patterns we observed. Of course, we couldn't make a judgment just from the resulting test data, because first researchers would need to seek out causes independent of the scores. But the results could help focus their search.

Or we could pull in race data and analyze if education policy is having an effect over time, and if not assign researchers to attempt to ferret out what might need to be judged and changed, which would likely include both school and non-schooling issues.

We could, in fact, do a lot.

But that's a pipe dream. We do that rarely at best. Policy makers and the general public mostly presume that a standardized test score on its own offers a signal of effectiveness, for both the student and the school, that it is a perfectly appropriate thing to draw lines in the sand and make judgments about those who fall on one side or the other, and to presume that the cause for success or failure is always and only the school, which in turn can be imputed by just staring at the score.

And teaching to the test, or making detailed instructional decisions, is now, sadly, a central part of a great many curricular efforts.

What we now do as a matter of course with predictive test scores is based upon false understandings, because the tool and its underlying methodology have nothing to offer regarding any of it.

This would all be laughable except the consequences are huge and entirely negative. States spend around $2 billion annually on tests and supporting materials that are used for things that they were never designed to do. That leaves what is arguably one of the most important of our social institutions, and the biggest expense for states, with judgments on those institutions from an instrument that was never capable of rendering judgments, let alone identifying effectiveness. The invalid judgments they produce in turn erode public trust, leave schools vulnerable to accusations of ineffectiveness or declarations of effectiveness that are neither valid nor true, which leaves lots of people wondering out loud if investing in or even funding public schools is worth it.

Not to mention the damning socioeconomic aspect in all this. Trying to use the resulting ordering as a judgment tool by drawing a line in the sand all but guarantees participation trophies to schools in wealthy neighborhoods for opening their doors, and sanctions on schools in poor neighborhoods despite incredible efforts, without evidence that any of the judgments is warranted. That validates and helps keep in place some of the most pernicious biases in American society and precludes many of our citizens that have the most to benefit from an effective education from getting it.

It affects real estate prices and property taxes. It drives decisions on where people will and won't live. It appears to support those who wish to privatize public education when in truth it does not.

It makes the entire enterprise of education worse, not better. All because of a misunderstood research instrument invented more than a century ago.

It is entirely accurate to say that in America we run a $2 billion a year Rube Goldberg machine to figure out where the rich kids and the poor kids live.

Policy makers unintentionally created a massively flawed accountability system for schools that was destined to mislead the public. They passed laws requiring schools to act as if the judgments were true, and now frequently blame schools for having created the mess, all because they picked and continue to believe a reasonable tool for doing what it was designed to do can function a million miles off label. That's a downward spiral we need to escape, and a little intentionality on this will go a long way.

It's time to shift the narrative about what the predictive testing methodology used by states is, and why it is the wrong tool for educational accountability. Not because we should hate it or declare ourselves anti-test or anti accountability, but because predictive testing is a surprisingly limited research tool that cannot answer any of the questions that should be at the heart of a proper accounting of what happens in schools.

This shift could have a profound effect in a surprisingly short period of time, creating space and opportunity for furthering the mission of public education like nothing else we can image. It could be a catalyst for overcoming biases, for enabling and expanding understandings about what happens in a school. It could someday get us to a better place regarding educational policy.

But remember this: the misunderstandings about predictive testing predate and in fact led to our current policy predicament. Policy makers, just like all of us living today, have never known a world outside these misunderstandings. Telling anyone that what they thought they knew all their lives is false, however true, may create a path to something better, but few will choose to take it.

That means that the responsibility to correct these misunderstandings lies entirely in the hands of educators. The choice is to continue to act as if those misunderstandings are true, or act only within the limits of what the methodology can tell us. It is ironic that used within its limits as part of a larger research agenda, predictive testing can aid efforts in improving public schools, while using them as if the misunderstandings are true is one of the surest ways to prevent it.

Even more ironic is that doing the right thing is the only way to satisfy those who have insisted for years that educators to the wrong thing.

Best,
John Tanner
Founder, bravEd

John Tanner is the founder of bravEd, an organization dedicated to getting the accountability function in schools right. He can be reached at john.tanner@brave-ed.com