September 14, 2020
The74Million.org

Beyond the Scantron: Harvard Expert Andrew Ho on the 3 Ws of Testing and How to Figure Out What Students Have Lost Academically to COVID-19

(Kent Nishimura / Los Angeles Times via Getty Images)

This interview is part of "Beyond the Scantron: Tests and Equity in Today's Schools," a three-week series produced in partnership with the George W. Bush Institute to examine key elements of high-quality exams and illuminate why using them well and consistently matters to equitably serve all students. Read all the pieces in this series as they are published here. Read our previous accountability series here.

Andrew Ho is the Charles William Eliot Professor of Education at the Harvard Graduate School of Education. A psychometrician, Ho serves on the National Assessment Governing Board, which determines policy for the National Assessment of Educational Progress. His

expertise in constructing and explaining high-quality assessments has led him to serve as an advisor to state education commissioners in states like Texas, Massachusetts, and New York. Ho also has taught creative writing in his native Hawaii and AP physics in California.

The testing expert spoke with us about the standards that guide the construction of standardized tests. He delineated the differences between diagnostic, summative, and formative exams, while emphasizing the purpose of a test. And he offered his suggestions on how schools can best determine what their students might have lost academically because of the COVID-19 pandemic.

Courtesy of Andrew Ho

What are the differences between diagnostic, summative, and formative tests and the role that those play?

There is no more important question to ask when we talk about tests than what is the purpose of your test? And who will use the scores to make what decision?

I tweeted recently about the three W's of testing, which is *who* is making *what* diagnosis to inform *which* decision? That is at the heart of testing in our field.

In fact, we don't say in our field that we validate tests, we say that we validate uses of test scores. That is a subtle but super important distinction. It depends upon what you're using the

test for. The uses of low-stakes tests don't require as much evidence or work as developing a high-stakes state test, where you have public technical manuals that are 500 pages long.

Diagnostic testing implies that you're making a diagnosis potentially to inform a decision. The term is very broad, and we are cautious with the term, because it implies the tests have more power than they do. You would love it if your child could do a worksheet for 20 minutes and you then would know what they need to work on next. But there is usually no way to wring that much information out of such a test unless the test is very long or given many times.

By contrast, a formative test is designed to support learning, usually teaching and learning. Summative tests look back on the learning that has already happened, so they usually happen at the end of a unit or the year. And they are usually a longer test, and more standardized and comparable. The term is sometimes used to refer to tests like the SAT and the ACT, although I prefer to call those selection tests.

I use a three-way division in thinking about the purpose of tests. The first purpose is for selection. Those are high-stakes tests that could diagnose, for example, whether an individual is prepared for college, a particular job, or a profession.

Another purpose of tests is to support learning. Those tend to be formative exams.
The third purpose is for program evaluation and monitoring. How do we understand whether a program, including a school or classroom, is working?

These big-picture categories are more distinct than diagnostic, formative, and summative.

What would you say to parents and others to help them understand what goes into the making of a high-quality test? What are their essentials and how are these tests constructed?

We have a consensus on standards for our field and contractors through the American Educational Research Association, the American Psychological Association, and the National

Council on Measurement in Education. States, districts, and testing organizations of all stripes know about these standards. We are not just making things up when we think about testing. There are standards.

There is particular consensus in these standards about the five sources of validity evidence that good tests must have. I like alliteration, so I call this the 5 Cs of validation: *content, cognition, coherence, correlation, and consequence.*

Content starts with defining what you measure. You look at the questions and ask, is this important? And you have to involve the subject matter experts. If you want to measure mathematics in fourth grade, it helps to know the relevant knowledge and skills, and how these are taught.

I tell education commissioners to carry five note cards with them. When people say that the tests don't measure anything relevant, then pull out a note card, read the question, and ask them if this is important for their child to know. People will usually say "oh, yes," they want their child to know that material. That is because the questions have been vetted exhaustively by teachers and subject matter experts.

Next is *cognition*. After we develop a question, we sit down with kids in cognitive lab sessions to see if these questions engage their minds in the way we anticipate. We actually test questions with kids.

The third C is *coherence*. We want test scores to generally cohere. Scores wouldn't change too much if we gave the test on a different day and if we gave slightly different questions. We collect a lot of evidence to make sure that test scores are stable across different items and idiosyncratic changes. We also call this reliability and precision.

Fourth, we want test scores to *correlate* with things and predict things. For instance, if you have a high SAT or ACT score, generally that should predict something that would happen

in college. It shouldn't predict it perfectly because that would be disturbing, as if the test is some kind of oracle. But, in general, we want things to correlate well. That is the fourth kind of evidence.

Then fifth, and this is an important kind of evidence, we want the *consequences* of testing to generally be positive. If we measure to improve learning and schooling, we should check to see if that improvement is happening. If we go out and do some testing, we would like for the outcomes to be better than had we not tested at all.

Let's talk about cut scores, which can be opaque for people. When done right, how do cut scores look? How are they used?

Cut scores require a judgment. How good is good enough if you take a driver's test? You could determine that arbitrarily and set the answer at the median. Historically, that has been a simple way to do things. Who could argue with average?

But there are two problems with that. If you define a cut-score at the average, you are dooming 50 percent of folks to being below average. Another problem is people often use a historical average. You then start to find the average student is above average. The current average goes up and up compared to the historical average. You end up with the Lake Wobegone effect, where all the kids are above average.

Also, the average sets no goal and has no authority. It is this abstract statistical concept. We have moved over the past 30-40 years to establish criteria for judgments. We start with a North Star when we set standards. We look to terms like "proficiency," "adequate," "exemplary," or "basic."

The most common term is proficiency, which we follow up with a policy definition. I sit on the National Assessment Governing Board, which oversees the National Assessment of Educational Progress (NAEP). The board's definition of proficiency is "competency over challenging subject matter."

That becomes the North Star.

Some states have a definition of proficiency that is different from NAEP's. It might say, "adequate mastery of skills." That is a consequentially different North Star. In that state, proficient means adequate.

As far as deciding where to set the cut score, you will get dozens of teachers and subject matter experts in a room over three-to-five days to debate what phrases like "competency over challenging subject matter" mean. During those debates, they look at evidence about items and responses. This helps them triangulate and reach consensus despite initially different opinions.

That decision goes forth to state policymakers who usually agree with the conclusion. Then, it goes out into the world where we say, "If you score at this point, then this describes you." That is the logical chain. It is sometimes overwrought. But it is quite established, like psychometric "due process."

People who think quantitatively ask, wouldn't it be easier to just set the cut at average. But that is not what parents and policymakers are asking about. It's a good start to know your kid's average but what does that actually mean they know and can do? We want to know if they are on track to college and career readiness. And that's not average right now. People want more ambitious standards for our kids and our schools.

This is important to know. These tests are not ranking kids against each other. They are ranking them against a level of competency set by a panel of educators.

Let's move on to the challenges schools face with the pandemic. What is the right role of tests, and how should we be using the information they give us at the classroom level to the policy level, as we manage a pandemic? What can they give us and what can't they give us?

Testing is not my highest priority in the COVID-19 era, and I do tests for a living. Testing is not as important at the moment as physical, mental, social and emotional health. Your brain's not going to work as well if you're sick.

But we should be very specific about what we mean if we recommend a district-mandated or state-mandated test this fall. Why do we need it? Who would use it, and for what purpose? What if most districts already have high-quality curricula that include assessments?

We need to tap into the silos of information that exist from when teachers in the spring say goodbye to their students, to when the teachers of those students in the fall greet them. Those teachers from the spring know which kids did not show up for online courses. If that kind of knowledge is passed along, it can help the teacher in the fall. Many districts and schools already do this, but there's never been a more important time than this socially-distant era in which to make this informal link.

Having said all that, nothing can answer questions about a student's grasp of a subject the way a state test does. They show where we stand compared to the previous year in a rigorous, standardized way.

I know "standardized" often sounds bad to some. But, to me, "standardized" sounds like fairness. It sounds like comparability, which is our ability to understand where our kids are this year compared to last year. As a citizen, I care about gaps that have increased, about progress that has been lost, and measuring this requires standardization and comparability to the previous year.

If we can have a light touch, through the creative use of tests, we can understand how much we've lost and how yawning those gaps have become. You don't need to sample everybody or give everyone the exact same test. And it might need only an hour instead of the usual longer testing windows.

There is nothing like hard data to lead to a powerful, effective response. I am framing all of this in a hierarchy of needs, but this is one of our needs right now. We shouldn't just speculate, we should measure formally and rigorously how much work we need to do to get our kids back on track.

Anne Wicks is the Ann Kimball Johnson Director of the George W. Bush Institute's Education Reform Initiative.
William McKenzie is senior editorial advisor at the George W. Bush Institute.