

EDUCATIVE PERFORMANCE ASSESSMENTS

Jon Snyder
Bank Street College

In my comments I wish to address measurement and assessment. I hope to do something unheard of in the annals of academics talking about assessment: I will be succinct and simple. I apologize in advance to any psychometricians reading this because when one makes things “simple” one leaves out nuance.

I will start with three “basic facts” about assessment. From those basic facts, I will propose two principles that are actually an alternative tradition in the world of assessment. I will conclude by demonstrating the use of those two principles in the assessment of pre-service school-based leaders (principals). The principles, however, have applicability for the assessment of children, teachers, principals, and even chancellors.

BASIC FACT 1: All measurement has error.

That is neither good nor bad, right nor wrong, just the way it is. There is error when using a ruler to measure length or when using a thermometer to take one’s temperature. No matter the measurement, there is some degree (from relatively large to relatively miniscule) of error.

The error factor increases when working with “moving parts” (like human beings). Emmanuel Kant put it this way: “Nothing straight can be built from human timber.” This is why you get what is sometimes called the “Aunt Emma” syndrome. Some study is conducted and it says X. But everybody has an Aunt Emma where X didn’t happen.

For instance, there is considerable measurement/assessment evidence that smoking causes cancer. My father, however, who never smoked, exercised regularly, and ate healthily, died of lung cancer. Our neighbor, who smoked, rarely exercised, had doughnuts for breakfast, fried chicken/burgers and fries for lunch, and Jack Daniels for dinner lived for 10 years longer before dying in an automobile accident. No matter what the statistic there is almost universally a counter individual example.

The error factor increases further when assessing humans doing complex tasks. There is less error assessing whether someone knows that $2 + 2 = 4$ than assessing whether someone can adjust a recipe to feed twice as many people as outlined in the cookbook (let alone actually cook the meal).

The error factor increases even further for assessments that are supposed to predict something that will happen in the future. Sticking with the cooking example, there is even more error trying to predict whether a person will adjust a recipe accurately several years from now when trying to prepare for a large dinner party. Or, using a

football metaphor, there is a scouting combine for college players entering the NFL that uses all sorts of very precise assessments. Some high draft picks are stars and others never make it in the league.

This means that when you make a decision about an individual based upon a measurement of some sort, you won't be right all the time. And if you are trying to assess a complex task you will be right less of the time. And if you are trying to predict the future, you will be right even less of the time.

So, the question isn't whether the numbers are accurate, they are not. The question is the level of error/uncertainty with which one can "live" before making decisions based on that assessment. That is always related to the "stakes" attached to those decisions. In general, one would hope that the greater the consequences, the greater the certainty. But it doesn't always work out that way. Eisenhower launched D-Day (thankfully successfully) on some very error prone weather forecasts.

One way to hedge one's bets because of this basic fact of measurement is to use multiple sources of evidence before making a decision. The more sources of evidence one uses, the lower the level of uncertainty. It is what Leigh Shulman calls a union of insufficiencies. Returning to Eisenhower and D-Day, in addition to the Farmer's Almanac, he also used some error prone intelligence reports about German troop strengths, and error prone maps of Normandy beaches.

Certainty would be nice, but measurement and the numbers generated by measurement cannot provide certainty. One of the biggest problems is that people think assessments can provide what they cannot. On the other hand, waiting for certainty means never being able to act until after the fact. It is a little bit like the rock skipping contests I used to have with my younger brother. We would both skip rocks, and then after the fact, I would define the criteria. Not surprisingly, I always won!

BASIC FACT TWO

Assessments/Measurements of complex tasks assess a proxy for the whole goal, they sample parts of the universe, they do not measure the whole universe.

The cooking example used above provides an easily understandable case. Knowing how to read a recipe and adjust the recipe as needed is an essential component of cooking a meal. It is not the entirety of cooking dinner ... and cooking is only one of many essential components of the broader goal of a good meal. There is locating and buying the ingredients, knowing how to use the utensils involved, preparing the table (or where-ever the eating is to be done), serving the food, and even facilitating the social interactions that are an essential component of a good meal (unless one is dining alone).

This is important because of the "what gets tested gets taught" syndrome (especially when the stakes are high for an assessment). This is not what is often called an

“unanticipated outcome.” This is an absolutely anticipated/expected outcome. Back to the poor kid in the kitchen, the outcome is that if all that is focused upon is reading and adjusting the recipe, he’d make a pretty lousy bananas foster, if the banana was rotten, or his guests or family would have a pretty difficult time eating a wonderfully prepared beef wellington out of a wine glass without a knife and fork. It would be like in golf, if we only assessed driving distance in one of those video golf courses ... it wouldn’t provide a particularly accurate prediction of actual golf on an actual golf course. The educational examples are obvious. The capacity to read short passages and respond to questions about those passages is a sample, a proxy, of our many goals for public education. The proxy or the sample is not the whole and the whole can be harmed by not understanding the difference.

BASIC FACT THREE

Measurements/Assessments NEVER tell you what to do.

Numbers never explain what they mean ... the interpretation of numbers and what the appropriate actions in response to them are not measurement/assessment issues. They are human judgment issues.

I am one of the few remaining living smokers (and if I continue, not for long). I know I should not. I know all the facts and could pass any assessment of my knowledge of the facts that smoking is bad for me and for those around me. I can be very clear on my goal – to quit smoking. I can assess that goal with a remarkable amount of accuracy each evening. Did I smoke that day? I know my goal/standard. I have a highly accurate assessment. But what does it mean when my yes/no assessment is no every night? It just tells me I smoked that day. It tells me neither what to do any better the next day to quit smoking nor why I am not quitting.

Alternatively, take the example of a young pitcher returning from surgery, let us name him Strasburg and say he pitches in our nation’s capital. Let’s collect all sorts of historical assessments about pitch counts and injuries and career expectancy. We have a great deal of assessment data, but the accumulation of data does not really predict what will happen if he pitches a lot more or a lot fewer innings ... and is even less precise in predictive ability when it is a few more innings here and there. Nor does that wealth of data take into account the balance between long term and short term issues for him and for the team. Could the team win the World Series if he pitched more innings this year? Would they have beaten the Cardinals if he was pitching? Will they ever have this good a chance to win the World Series in the future? Ultimately, those are all human decisions requiring juggling multiple factors, multiple variables, and all with imperfect precision.

The research in teacher education provides another example. A very thorough study was conducted of different pathways into the New York City teaching force. One of its goals was to compare the student achievement of alternative (meaning start teaching

before completing a teacher education program) and traditional (meaning start teaching after completing a teacher education program) routes into teaching. In the first two years the students of “traditional” entrants into teaching did better than the students of “alternative” entrants. In the third year, after the alternative entrants had completed an approved teacher education program (a state requirement in New York State where the studied took place); the students of the alternate entrants did better than those of traditional entrants. One interpretation of that is that teacher education programs improve outcomes for kids; another is that the alternate entrants did better in the third year because they were no longer wasting their time in teacher education programs. Interestingly enough, the media headline was that TFA was better than anybody, which wasn’t even what the study was assessing!

It is not just that the all measurement has error; it is that measurement does not, cannot, tell you what to do. Humans interpret those measurements and humans determine actions based (or not) upon those measurements.

These three basic facts have implications for the development and use of assessments. Those implications can be summarized in two principles.

PRINCIPLE 1:

For complex tasks, assess actual performance, not the knowledge that might be involved in performance.

Our “individual” goals (not to mention, but not to forget, our social goals) for public education are complex – college and career readiness, productive contributors to communities and citizens of a democracy, healthy life style, and fruitful family members – certainly all involve a complex constellation of tasks. It is not enough to know that smoking causes cancer (even if my neighbor smoked and didn’t die of cancer) but rather that I NOT smoke. It is the performance that matters, not the knowledge.

I will use an existing governmental policy as an example. To get a driver’s license involves exhibiting some knowledge of the rules of the road via a multiple choice test of such. This is a relatively inexpensive assessment to administer, its reliability from one administration to another is pretty high, and since a machine scores it, the inter-rater reliability is pretty high as well. But we wouldn’t think of giving someone a driver’s license for the first time without seeing if they could actually drive. Unlike with teachers, we would also not think of letting someone drive for two years and then if they did not get in a wreck give them a license.

This is an example of a performance assessment --- as well as multiple measures. It also, however, is still an example of assessing only a proxy of the entire goal. When I took the exam and when my son took the exam, we parallel parked, backed around a corner, turned left, etc. We did not drive in traffic, in a heavy rain, on a free-way, or in snow on a country road (or all the other things that drivers will face). Thus it provides an

example of the basic fact that assessments do not measure the entirety of the desired universe. It is also an example that assessments cannot provide absolute certainty. There are still bad drivers and there are still accidents. Multiple measures that include a performance assessment, however, improve the odds in our favor.

PRINCIPLE 2: ASSESSMENTS SHOULD BE EDUCATIVE

Assessments first received large scale governmental investments and became a “science” as a by-product of World War One. There were all these people going into the army and the military needed a way to sort them into an appropriate role in the war effort. Two sets of assumptions drove its development:

1. Intelligence (for lack of a better word) was assumed to be relatively fixed , that “intelligence” as measured was what mattered for success, that there was a continuum of intelligence and people were relatively fixed on that continuum.
2. There were roles/tasks available for people all across that continuum.

Today neither of those sets of assumptions holds:

1. Cognitive ability is NOT fixed, there are other elements of success (other things matter, such as grit, creativity, etc.) and it (whatever “it” is) can and does change based upon the influences of the environment.
2. There are not enough roles/tasks in the flat world available for people all across that continuum. Jim Clifton, the chairman of Gallup, in his book “The Coming Jobs War,” points out that of the world’s five billion people over 15 years old, three billion said they worked or wanted to work, but there are only 1.2 billion full-time, formal jobs.

Thus, assessments have to be educative. Their goal can no longer be to sort people, but rather to help educate people. It may be difficult and expensive, but it is possible and it is essential. THE GOAL OF EDUCATION IS EDUCATION. If an assessment only tells me that I am better than you, and does not help educate both of us, then it is a waste of time and money and has fully anticipated harmful outcomes.

In conclusion, I will provide an example of a principal performance assessment that Bank Street and our partners at Stanford and Vanderbilt are beginning to develop that meets these principles.

Our goal is to create a performance assessment system that meets multiple functions:

1. Assure the state of its capacity to meet its legal and ethical responsibility to only grant a principal’s credential to someone who has demonstrated the necessary performance to safeguard the educational rights of our children. (Like the state’s responsibility to only give a driver’s license to someone who has demonstrated at least some capacity to drive safely.)
2. Enhance the learning of the individual candidates for the credential.
3. Improve the preparation pathways capacity to educate the candidates in their care.

It is noteworthy that while our goals are educative, they are also “responsibility” oriented as well. Responsibility and Education are NOT mutually exclusive. In fact, they are mutually interdependent. You cannot have one without the other. This assessment system will allow the state to offer a credential for a principal to begin (meeting its responsibility to our children) and help principal candidates learn, and help pathways to improve.

The first issue to deal with is to define the attributes of the performance tasks.

Because leadership is complex and requires the integration of multiple sets of knowledge and skills, we envision an assessment system that includes three to five tasks that:

- Each possess multiple components assessing two or more standards (not a task for each standard);
- Assess two or more challenges/dimensions of leadership, requiring candidates evidence to integrate the constructs;
- Reflect the complexity of leadership;
- Require authentic work that is part of the lived world of educational leaders;
- Include sufficient documentation of contextual information to enable sound assessor judgments (the answer to any educational question is always “it depends” and knowing what “it depends on” in the context is therefore essential to generating appropriate responses);
- Are sufficiently robust to limit capacity to “game the system.” (Cheating is NOT OK, and it shouldn’t be an anticipated outcome of an assessment system. To be on the safe side, however, it is always better to try to decrease the temptation by making it difficult to do so.);
- Include addressing the process (planning), the end result (product/outcomes), and evidence of “learning from the experience” (capacity for continuous improvement as an essential “standard” for principals);

Finally, to be used for assessment purposes, the tasks must, as accurately as possible, assess and differentiate levels of performance.

Since our goal for principals is not to recreate what is, but to create better schools and improved outcomes for all our children, one task, for instance, could focus on Continuous Improvement. One of the greatest challenges 21st century leaders face is implementing educational change that results in improved student learning. New leaders must possess the knowledge and skills to create improvement plans that provide evidence to support the need for change, underscore developing capacities and relationships to sustain change over time, target resources and systems realignment strategically, and incorporate inquiry practices for ongoing monitoring and feedback.

In order to do so, effective leaders must:

- understand the principles of high performing schools (knowledge, but also);
- be able to execute the steps to plan, guide and manage continuous improvement (e.g., knowing where to begin, how to coordinate and pace the improvement initiatives);
- build her or his capacity to develop a widespread understanding of what school change intends to accomplish among constituents (e.g., which constituents to engage and how);
- promote the strategic engagement of relevant stakeholders in identifying, creating and achieving common school improvement purposes, as well as to continue to support and enrich a sustainable process of continuous improvement (e.g., how to analyze data and present information to guide the improvement process);
- engage staff and their community in an inclusive process that forges solutions to problems that produce better outcomes by promoting trust among stakeholders, addressing the interests of diverse constituencies, and developing a common understanding of the reform efforts needed (e.g., what resources need to be mobilized); and
- Understand the systems implications of implementing, managing and sustaining change (e.g., how to engage central office support).

One possible task to assess this desired performance would be to ask the candidate, using his or her school, to develop a comprehensive school improvement plan. The plan should focus on improving the performance of all students including those who are most at risk of not meeting challenging performance standards. It should:

- be based on current school data (student, teacher, and program),
- reflect the research-based practices of highly effective schools,
- acknowledge contextual barriers that need to be addressed for change to take place,
- identify mechanisms and strategies for engaging stake holders in the process,
- indicate the allocation and reallocation of resources,
- specify a timeline and the tools that will be used to monitor progress and make necessary in course adjustments.

In order to “pass” this task, the candidate would need to:

1. Describe the nature of the task;
2. Describe the issue or problem being addressed in the task;
3. Briefly outline the critical contextual factors;
4. Present data about the problem or issue, with attention to equity considerations;
5. Provide evidence of planning;

6. Explain the research and evidence-based rationale for options considered and choices made (e.g. allocation of resources or selection of programs);
7. Describe the product;
8. Outline the anticipated challenges to adoption or implementation;
9. Reflect on the leadership implications and leadership skills developed through this task.

This one task, of a proposed system of three to five assessment tasks, provides an actionable example of the assessment principles as they could be applied to the assessment of pre-service principals. The development of, and more importantly, the enactment of this proposed Principal Performance Assessment system will require the “good will” and the conceptual contributions of the professional community (in higher education, from alternative providers, and from the K-12 world). We think, however, that it is possible (as does the Commonwealth of Massachusetts that is funding this work). More importantly, our children deserve and require nothing less.