



PODCAST TRANSCRIPT

Key Considerations in Measuring Educator Effectiveness

Talk by Stanley Rabinowitz, December 2, 2011 at Learning Innovations, Woburn, MA.

Basically what we are talking about is teacher evaluation for high stakes. That's really what the law is saying and what states are looking to do. I've been doing this for almost 30 years now and I've never worked in an area where policy is so far ahead of practice as teacher effectiveness, educator effectiveness. It is almost impossible what states are saying they are going to do. There was a big crush, last minute crush, round one Race to the Top, to pass these laws in order to be eligible to get those extra 20 or so points for their applications. And outside of the northeast basically most states, they passed these laws, they didn't get the money, and now they are stuck implementing them. And if you thought you could then back off on it the waiver process for NCLB added in and included in there is the same expectations about educator effectiveness, high stakes assessment-driven systems to measure teacher effectiveness and more generally principal and school effectiveness. So states that thought, "well we didn't need it, we didn't get the money," they certainly want waivers. Massachusetts was a first round waiver state. Pretty much every state is going to be a second or third round waiver state. And so these laws are there and they are going to put it in their workbooks for NCLB accountability or I guess we should say ESEA accountability.

So we are stuck with these laws that on the surface make sense. There are basically two principles behind them. One is that educators should be evaluated like other people, like we do in all our jobs. Second, assessment should be a major driver of that system. Again, assessment, typically is the most objective available data. So it's hard to argue that teachers should be evaluated and that assessment shouldn't be a part of that system. But that is a 30,000 feet view. Everything looks good at 30,000 feet or isn't really distinguishable at 30,000 feet. As you start landing, that's when you say, "that's not a runway" or "you know we've never really landed this thing before. Getting up is easy, getting down is hard."

So what I would like to talk about is what states should do, given that nobody knows how to do this. We know sort of how to do it and there models of doing it, but to get it right in the next year or two, which is what most states need to do, is impossible. So I'm going to be talking about mostly today a research plan of how you can get it right in five years from now, which I think is possible, or at least get it "righter." And we can do that.

So basically, what are the characteristics of reliable and valid indicators, what's the role of assessment, and to me more importantly, how can additional indicators improve measurement. Because assessment for several reasons is not the silver bullet. The value-added. Colorado growth model, growth vs. status: You are going to hear a lot of terms thrown out there. The system that I'm going to talk about is model neutral. I don't think there is a best model. The model that you are going to select should be based on some technical requirements and some history: What's the values

within the state?

So one big difference in growth models is that some growth models are actually normative or it's not to grow to meet a particular standard but to grow more than the typical school or the typical teacher. Those are two very different models; they both are defensible and they both have problems. Clearly you want kids to grow to the standard, but a school where most of the kids are so far below the standard is at a real disadvantage. And so a normative growth model is one where we reward growth better than typical. That sounds fair, right? Except I can keep growing better than anybody else in my situation, but my kids may never reach the standard. So I'm going to keep rewarding you for growing in what I call the Disneyland Rule. Where you get up to that ride and there was that line there, that red line, and your nephew, or son, or daughter could've said "Well I grew 6 inches last week. You should let me in the line because no other kid in my class grew 6 inches last week. I've got better than average growth." Then the person at the line will say "that's really cool. Come back next week when you've grown another 6 inches and you're actually above that line." Because that line means something and we really don't want you falling off of this ride.

It's the same thing with certain growth models. If you don't grow to the standard, if you don't grow to that line, then it's nice, you are better off if you didn't grow but it's still not good enough, because that standard, hopefully, means something external. It means you're college career ready. It means you're ready for fourth grade. It means you are an effective teacher. That line at the amusement park actually is externally validated. Somebody has done a study that if you are smaller than this you could fall out of that ride. Most of our education red lines are internally validated. Meaning we bring a bunch of educators together and they guess where the standard should be and we have elaborate methodology to help them guess. Very few of our red lines in education are externally validated, meaning that you are college ready if you pass the high school graduation test. And most state graduation tests, and I've run one, are designed not just if you are ready but so that too many kids don't fail. And it's a balance between the external, credible, line and the internal, politically livable, line.

All this was a long way to say that no model is perfect. So you pick the model that makes the most sense for your state, your system, and your values. Understanding that somebody won't be happy and they properly should be because everything is a compromise. I love assessment. I've lived my life making assessments but I also know the warts. When we're talking about educator effectiveness there is that magic 69% rule. That's on the average the number of teachers who don't teach 3 through 8 math and reading, high school math and reading, and sometimes science. That's what we call the non-tested grades and subject areas—your science teachers in others grades, your social studies teachers, the school nurse.

So states are going in two directions and I hold my head on one of them. They are taking the model of their reading and math tests and trying to extend it to everybody. Here's my typical response to that approach: One is that you can't afford it, because it is actually expensive to do that right now and second is, who in this room believes that if didn't have reading and math tests now, that we would design them the way most reading and math tests are? So let's see, let's have a kid read a page of a passage and then answer some questions about it and we that reading. We wouldn't do that. We accept it because we have always done it that way and so what I tell my states to do is use the 69% to both get better models for the 69% but get better models for our current reading and math tests. Let's not bring a bad model to everybody, let's improve the incomplete model we have for everybody. So some states are forming what they are calling content collaboratives to come up with

better assessment models for those other subject areas. I'm fighting a district—there's a district in Arizona that wants to give me three million dollars to develop reading and math tests-like for their other content areas. I'm stupid enough to say you really don't want to do that. I want your three million dollars though don't get me wrong but I really don't want you just developing multiple choice tests of art. Or music. There is a role for that; there's some knowledge you have to have but wouldn't you also rather have these other components, a performance piece. Wouldn't you really like to have a performance piece for our math tests and certainly our science tests? So we're going to compromise. We'll have multiple choices but we will have at least some part of those other pieces, and I hope they will generalize that to the reading and math tests.

So we need multiple measures, what we call triangulation in assessment. What is basically means is I don't have exact measurement, a single best measurement, so what I'm going to do is I'm going to surround the indicator with a bunch of other measures that don't duplicate it. I want to measure effectiveness broadly, where assessment on demand is just one of those pieces. I also want school climate. I want teacher control of that classroom, which may or may not show up in assessment. The bottom liners say it all should show up in higher test scores. I buy that 51%, which is what a lot of states have in their laws. It has to be majority assessment driven, educator effectiveness, but I can't buy any more than that 51% because there are some things that don't translate directly into test scores that are equally important: perseverance, attitude. Again, it's a very simplistic argument that says we can just do it and it will show up in higher scores.

So when we triangulate we look for other kinds of non-duplicative measures. The problem with them is that they are less reliable, they are more subjective. But we can build—the classic, I don't know how many of you read about six months ago, Gates came out with a report on using videotape to measure principal observations of instructors. That as many of you know, many of you have been principals or teachers, know that to be one of the most subjective endeavors. But what Gates does and what others that work in this area do, is you videotape it, so you have an artifact. You have a strong rubric of what you're looking for and you train external people to apply the rubric. What we now have taken the basically subjective compliance type approach, added objectivity, and added a little bit more science. And that is the kind of way we can take as subjective, suspect measure and make it more objective, and make it defensible in this kind of system. And here you are clearly getting at different information than the assessment does. You are seeing if the teacher or the principal can apply certain instructional practices that you value as a state or as a district. So you can see where we are going with this.

Two big lessons; don't make the mistakes we've made for reading and math and apply them elsewhere. And use a broader set of indicators which can be made objective through training and other kinds of—if not more objective, less subjective as you build up. I believe that the 50% rule is necessary. Assessments are the currency of the realm and so we need to use them as a valuable, we need to expand them, then we need to supplement them. Indicators should be added based on availability. Are they there? That is really important. They got to be there. We have to be able to afford them. We have got to develop them. They need to be technically adequate. The public and educators need to believe in them. Teachers don't love assessments. They don't believe that's getting at what they do. But if you add those other pieces, if you say “You give me assessment, I'll give you observation. I'll give you student service, parent service, any number of measures that can be made more reliable over time.”

I love the value of versus. burden comparison. It means that if people can't do it or won't do it it's

not worth it no matter how wonderful it is. Again, you make these decisions, the same thing here. If you make a better test, then the value of that test goes up and the lost instructional time, which is a burden, goes down, because the test provides information that is worthwhile.

Then you always have to remember that purpose drives the definition of adequate. The subtitle for this is “lawsuit.” What’s good enough for informal feedback between the principal and the teacher is not necessarily good enough if you are going to fire the teacher based on the results. That’s the difference between the principal doing the observation and the videotape with the rubric with an external reviewer. One is for feedback; the other is for promotion, accountability, and a whole bunch of other—we always like to say “rewards and sanctions.” So as the purpose gets more higher stakes for individuals or the system, then you really need to have much more reliable, technically sound indicators and evaluation accountability systems.

I talk a lot about different level data. The most defensible is what I call Level 1, technically sound state-wide data. Like your third grade math test is Level 1 data. Uniform across the state, there is enough adequacy and it can probably be used to some extent to evaluate students, schools, and teachers. Not totally. The problem is, as I said we have a 69% rule, and we really should be thrilled about just using assessment data even when we have it. So we have Level 2 data or other measures that have some technical adequacy but they may not be fully aligned or uniform across the state.

The easiest example I can give is in your state schools may use different interim assessment systems. Those have some degree of alignment to the state standards. They are just not uniform and they are not fully aligned. But there may be a way of using them either through some concordance table across them or another way of making them more equitable. Then there are Level 3 data, observation surveys, things like that which right now are too subjective to use but over time, through things like videotaping and training, can rise to Level 2 and potentially to Level 1. Think about what you believe in now, start with that and over time think about a model that can move Level 3 to Level 2, Level 2 data to Level 1.

And the way you do that is through a research agenda. And there is that 1-3-3-5. One means what have we got now and we got a whole lot of nothing in a lot of places but the law says we have got to start, and that is why school nurses get evaluated by reading and math scores. Because that is the best we got now. I can sort of live with this if two things happen, actually three things. One, you hold your nose while you do it, second is you keep the stakes relatively low: it becomes a feedback instead of a firing. And third, most important, you think about how do I get to a better place two years, four years, and ten years.

And by thinking about that now you are in a position to start doing the study, start doing the studies, start doing the research, the investments so by year three, I now have an assessment that I believe in in all my content areas, at least piloted. Because in two years I can build an art test; I can build a dance test; I can build a social studies test, and I can build a better math test. I may not have all my teachers ready to implement yet because it is harder. You have to actually administer it. You have to observe it and you have to artifact it, if I can invent verbs. But in three years, two more years from now, I can have a pretty comprehensive system and I can train my principals to objectively apply a rubric to their teacher observations. I can have some of them videotaped to be externally monitored by some state entity.

The reason I say 1-3-5-10 is you’ll never get past one if you don’t think three and your three will be

too limited if you don't think five. By thinking ten, you can really think about "what do we really want to do? What is effective teaching? What are effective outcomes?" And that can get five years right. Because what you find out is most of what you want to do really won't take ten years. So you've got to think in these stages and you can live with one the way it is if you tell people this is just temporary. We're going to get to a better three and a really cool five.

So that is how I like to push my states, get aggressive in their innovation, be realistic about today. Don't do too much today. The analogy I use is—the biggest mistake individuals make in their life, lower stakes, big mistakes, but a big one, is buying a house they can't afford. You look around at the economy: Something magical is going to happen. I was going to be able to afford this, and so you end up with two really bad things: You own a house but you can't afford to put any furniture in it, so all you have is a house. Or you end up buying stuff that you don't really want to have. Or in the worst case you lose the house. What a lot of states are doing is, they're building a house with their accountability system, they can't afford to run and they don't know how to run. "It's all solar!" But it rains a lot here. Build the house you can afford and then build onto the house or like allot of us do, we bought a really good bedroom set and then two years later we bought a really nice living room set. Build an assessment system you really like now, build the observations over time—that is your new living room. You now actually have a full house as opposed to allot of empty rooms. Having assessment as the only part of your accountability system is like having a big house and you only live in one room. Think about how you are going to add onto it. Think about how you are going to furnish it and how you are going to afford it and how you are going to train people to live in it. Because ultimately, this crashes or burns based on people's ability to believe in your system, implement it, and grow with it over time.