On Performance Assessment: A Conversation with Grant Wiggins

An advocate of "tests worth taking," Grant Wiggins warns: when it comes to any kind of testing, one size doesn't fit all.

RON BRANDT

I've heard you tell a true story about a group of high school students who were asked to design a boat. They knew a lot of theory, but had never applied it. When they built their boats and tried them in the school swimming pool, most sank.

Are educators in the same situation? We've been talking theory about performance assessment and we've convinced ourselves. We've even sold it to the governors and the legislators, some of whom have turned to us and said, "Okay, you're right. Standardized tests aren't very good, so use performance assessment instead." Are we like the students? Is our boat going to sink?

I think some assessment boats are already sinking — or at any rate, they're taking on water. I see parallels here between kids and adults: if students need models, criteria, and feedback, so do adults. Teachers now face the same situation as students who are asked to do a nonroutine task: how do they design new forms of assessment, the likes of which they've never seen? They know what they don't like about conventional testing — and they're absolutely right — but they don't know what this new vision looks like.

When you consider that performance assessment is complicated logistically and technically, you've got a serious problem, particularly if it's a high-stakes situation.

A lot of people in the technical community seem to think that assessment reform is bound to fail.

Yes, because high stakes will be attached to them too soon, judgments are going to be unreliable, and there may well be lawsuits. But do those same technical people complain about a 60 point margin of error on the SAT, when boards in small districts browbeat faculties over a meaningless five point drop from one year to the next? The problem is high-stakes, one-shot accountability tests of *any* kind. And nobody's saying that this reform has to be "either-or"; we need to redress an imbalance in our testing methods, that's all. Just because there are technical problems to be worked out in direct assessment doesn't mean there's no need for basic reforms in what gets measured and how.

What are some of the technical issues that educators need to be aware of?

A basic issue is reliability: making sure that the score is justifiable, precise enough, accurate enough. As a profession, we know quite a bit about how to deal with that, but a lot of faculties left to their own devices don't.

Can you say briefly how to achieve reliability?

You have to know the behavior you're looking for and have enough evidence to feel confident that the score given is

apt and representative. Make sure you have enough information collected over time on similar tasks. Practice and refine the scoring process. Unpiloted, one-event testing in the performance area is even more dangerous than one-shot multiple-choice testing, because multiple-choice tests have many different but related items, which makes reliability easier to get and measure.

Second, use multiple judges where possible and require high inter-rater reliability. With most teachers now testing and grading in isolation, we often end up with wildly varying grades for what is really the same quality of work. One reason the public has so little faith in transcripts is that they know they can't count on grades meaning the same thing. But you can set standards and get reliability through well-established techniques such as those used in scoring the Advanced Placement exams: fixed anchor papers, good scoring rubrics, and proper training.

How about validity? Is that an issue, too?

Of course. Just because a task is authentic doesn't mean it's valid for inferring mastery of a complex capacity. Technical people talk about "generalizability": does a particular task "generalize" to other similar kinds of tasks? For example, if the student writes an essay, is the score on that particular paper likely to be representative of his or her scores on other similar types of writing tasks? Writing prompts and performance situations in general are quite particular. What happens when we slightly vary the prompt or the context? One of the unnerving findings is: the student's score changes. Those who've studied the problem suggest that students may need to do at least six different tasks of a similar kind to make sure our inferences about overall mastery are valid.

But is it realistic to expect so many samples?

It is if assessment and accountability systems are built around local work, collected over time. That permits the teacher to select from a portfolio of student writing that has many samples of the same kind of work. The Australians now use local work in all state assessment. So does Vermont, and California and Kentucky will soon.

The need for multiple performances shouldn't surprise us. We can look at it as a new kind of reliability problem: the reliability of the novice. Who believes that novice performers are consistent? One performance, even by professionals, is often a risky basis for inference about general ability. Think of a football team. You watch them play a game, and they score 37 points. Is that typical of what the team can do? You don't know; you need more games played against different opponents under differing conditions.

This doesn't mean that performance testing is too messy to bother with, by the way. It just shows that a number of traditional views of validity and reliability need rethinking.

An example of the generalizability problem you mentioned earlier is the work of Rich Shavelson, who's been trying to assess science problem solving in California.[1] If people think they can give a student one task that assesses a domain as broad as scientific problem solving, they're in trouble.

Big trouble. We're not even sure how to operationalize some of these concepts. Hands-on is not necessarily mind-on. People may say one of their valued outcomes is critical thinking, but I can show you a hundred different ways people operationalize that term, many of which are inconsistent with my conception of what critical thinking is all about.

What you're saying is really scary, because we have a history in American education of trying to use a good idea, but messing it up. We have lots of examples, such as open classroom. The latest may be whole language: what it ought to look like vs. what it actually looks like. How can we prevent this happening with performance assessment?

First, we need to provide teachers with models and criteria for what good performance assessment looks like. That way, they can compare their own efforts with models and standards. Educators have to understand that it's just not enough to say, "I designed this task and I say it's valid." Having good models should also help teachers do a better job of evaluating commercial tests. We must be able to test any test against criteria, just as we test each student's performance against criteria. Then, just as with kids, provide incentives to improve performance.

We also need to beg, borrow, and steal real tasks from the professions, just as all good vocational education and arts programs do. Why reinvent the wheel? Work with college and business professionals.

Some might say a better answer would be to work at the national level, get some excellent examples, and spread them around the country. Why not national standards and well-designed performance assessment tasks keyed to those standards?

Sure, that's part of it. I've been a willing participant in some of the work that Lauren and Dan Resnick and Marc Tucker are doing on the New Standards Project. I've worked with them because I agree that it's very

difficult for local educators without the financial resources, the technical means, and the intellectual milieu to invent performance assessment with the necessary power and credibility. So let's bring together some of the best teachers and test people and curriculum people and design good stuff, but let's not use state and national efforts to further damage the capacity of local people to design their own assessments. We don't need more high-stakes generic testing that isn't linked to local curriculum.

But the New Standards Project is a national program, isn't it?

It's a national effort, but I like the procedure they propose to follow: regional "moderation," similar to what is done in other countries, where teachers get together to agree on standards — with an effort to calibrate student work in the South with student work in the East.

What's cockamamie is the search by other agencies for tests that are solely national, and the failure to keep "standards" distinct from uniform tests. We have a worthy 150-year tradition of local control and local curriculum, so a broad-based generic test that is also authentic is a contradiction in terms. You can't have a generic test if you want to do justice to the idea of local curriculum.

Well, some people are beginning to say, "Then let's not have a local curriculum either."

There's something to be said for getting outside of parochial norms and expectations. But there's a difference between that and mandating bland standardization. What many policymakers are trying to do is reinvent an education system that's a parallel of the very economic system Eastern Europe is walking away from — namely, a centrally run, centrally designed, centrally mandated "command" form of government. It's going to fail in education just as it failed in economics, because it doesn't empower and energize the entrepreneurship of local people.

Ted Sizer asks the right two questions: *Whose* standards? And by what right? Put that way, it's irresponsible to turn education over to unidentified — and unaccountable — "experts" from afar.

You mentioned Advanced Placement earlier. Isn't that an example of a national program that works pretty well?

I'd say that, given its limited purpose, the AP program works fine for the population involved. But let's turn that around. Why should we invent a new national assessment system when we already have the APs?

Yes, why?

Well, I wonder myself given the current debate. I suppose it's because many people think, "Oh, the APs are just for a select group of students going on to select colleges." But that just shows how badly confused we are about standards versus expectations. Rather than saying, "If it's good enough for the best, it's good enough for everybody who wants to go to college," we're now trying to search for a new test with a new standard. Why? Why *not* the APs? Why not the International Baccalaureate? Why not borrow what other countries already do if it's so great?

What's the answer?

I think we're still not sure that one size fits all. We're not really ready to mandate that sort of thing for everybody, because not everyone needs a traditional college education. Oregon recognizes that students have different needs in its new law requiring high school students to choose between college preparation and vocational preparation.

Now personally, I think American schools would be infinitely better if the AP program set the standards for academic programs. I happen to be a fan of the program, precisely because it puts teachers in touch with the best work that students around the country do in traditional disciplines. It's a model we can use, and I know that it has influenced the Resnicks. The question is, is that kind of program sufficient to encompass our objectives for all our kids? I don't think so.

The slogan is, "All kids can learn."

Of course — but it's not at all obvious that there is only one appropriate college-prep syllabus nationally for all students. I'm also a fan of vocational education done well; why not require some of their best tasks? It's still not clear anyway why we want a *national* curriculum in a country of this size and diversity.

I think what's going on is something more radical than rethinking testing. What we're really doing is rethinking our purposes. We're not sure what we want for our diverse population of students. We haven't figured that out. And until we do, casting this as primarily a technical debate about testing will obscure the real questions: What is worth testing? And are there only uniform national answers to that question? My answer to the second question is: no, there are not. □

[1] R. J. Shavelson and G. P. Baxter, (May 1992), "What We've Learned About Assessing Hands-On Science," *Educational Leadership* 49, 8: 20-25.

Grant Wiggins is Director of Research and Programs, Center on Learning, Assessment, and School Structure, 39 Main St., Geneseo, NY 14454. **Ron Brandt** is ASCD's Executive Editor.