Some Cautions Regarding the Use of Performance Assessments

H.D. Hoover

The role I've been asked to fulfill at this conference is that of critic. I want you all to know from the beginning that the comments I'm going to make regarding performance assessment and the use of performance assessment are purposely one-sided. I definitely think that the kinds of tests I've worked on over the last 25 years, such as the "Iowa Test of Basic Skills," have been overused in schools. They've been used for purposes they weren't intended for. I also believe that there are many things that we don't measure particularly well with multiple-choice items. We truly need an expanded set of testing options, not only in elementary and secondary education, but also in certification testing, the kind of testing that is primarily being addressed here. I'm very much a believer in the expanded role of different kinds of tests, and I don't want the comments I make to lead people to believe that that's not the case. I'm going to make some fairly strong technical arguments that must be addressed if you're thinking about different kinds of tests, different kinds of assessment, and, especially, an expanded role for performance assessment. I'm sure having a statistician get up here and say I'm going to talk about technical stuff thrills you.

As we all know, teachers are not very knowledgeable about measurement and measurement issues, even though a large proportion of their job is associated with assessment. People talk about tests and performance assessment as if the field were new, as if little were known about it, and as if we didn't know a lot about the reliability and validity of such measurement. In fact, there's a long history of the use of performance assessment in the United States and around the world. Performance assessment preceded what we do today. There were sound reasons why we changed to the kind of tests that we see primarily today, and we have continually incorporated improvements based on what we have learned along the way.

Dr. H.D. Hoover is Professor at the University of Iowa and Director of the Iowa Test of Basic Skills Program.

I think it's very important that anybody who is considering the use of tests with performance assessment components attached, especially those being used in high-stakes situations, recognizes that scathing attacks can be made on such assessments. In addition, there are tests that are used for professional certification with many associated legal implications.

When testing is the topic, there are four issues you must deal with. I know that people assume this is material only psychometricians are interested in, but frankly, anytime you perform testing, you have to deal with these four issues. This is true even if you change the name to measurement or assessment.

The first is the fundamental issue of reliability—what you always think we measurement people talk about. You sometimes hear people say that they gave a performance assessment, had someone score it, had a second person score it, and the two scorers agreed; therefore, the score was reliable. But this is not test reliability in any sense. What is being talked about is simply *rater* reliability. The reliability of a test is a characteristic of the behavior of those people who take the test—students. What is really meant by test reliability is what happens if I give a test today and a similar test tomorrow. Do students rank order in roughly the same way from one day to the next? That's what reliability means in terms of a test score. In most cases, if people are relatively careful, rater reliability is a solvable problem. But let's talk about reliability coefficients of tests. The real question is, when a teacher administers a given task and observes the performance on that task at a given time, does that single performance generalize to a different day, a different subject, with a student who is behaving differently?

For example, in the standardization of the writing supplement to the "Iowa Test of Basic Skills," we took a national sample of students and had them write narrative essays on one topic one day and on a different topic the next. This is rather like having a student take a reading comprehension test one day and a different, but parallel, form the next day. When we did this, the average correlation, or reliability, was .48. When we had students write a narrative essay one day and write a persuasive essay the next day, we got an average reliability coefficient of .36. This is considered extremely low for test reliability. What researchers tend to find is that reliability in the sense we've been talking about—how well a test score correlates with another score at another time in the same general domain of content—shows up with values of about

.35 everywhere. It happens in mathematics, in science, in writing, on all kinds of tests. For example, what happens if you give a science experiment one day and a different science experiment the next day? To quote Shovelson and Baxter (1992), "To get an accurate picture of an individual student's science achievement, the student must perform a substantial number of investigations, perhaps between 10 and 20." Such findings are central to the use of any kind of performance assessment and relate primarily to the reliability question. Now we may hear a person use the word *generalizability*—how well does this performance generalize—which may be a better word, but it's still essentially the reliability question.

Now I want to move to the validity issue. When people push for more performance assessment because they think validity will increase, they're usually talking about only one aspect of validity. This type of validity has been given the name *consequential validity* recently by people like Samuel Messick and Bob Linn. What are the consequences associated with the use of this test? The big argument for performance assessment is that it improves consequential validity. One view held by many is that tests have caused all kinds of bad things to happen, but if we change the nature of the tests, the consequences will be better. I happen to like this idea, and I agree that the use of tests is the big issue. Measurement people argue for different kinds of tests primarily because of the consequences associated with administering those tests.

Validity is something you have to gather evidence about, however, and this is where proponents of performance tests have come up a little short. You've all heard of face validity, I'm sure. In the case of many supporters of expanded use of performance tests, a better term might be *faith* validity. There's a lot of faith validity regarding the use of performance assessment, but I don't see much evidence of other kinds.

Another kind of validity is what might be referred to as the differential validity of performance assessments versus other kinds of assessments, for example, multiple-choice tests versus performance tests. However, exams consisting of both types of items, such as the AP exams, offer little support for the stereotype of multiple-choice and free-response formats measuring substantially different constructs, i.e., trivial factual recognition versus higher-order processes (Bennett, Rock, and Wang, 1991). Many people who gather data on performance assessments say they

measure something different than the multiple-choice task does. That's pretty easy to check, and it doesn't take a highly trained psychometrician to figure it out. By the way, the fact that the correlation between a performance assessment and a multiple-choice test is less than one is not very good evidence that they measure different things. The correlation between two variables will always be lower than one if they are not perfectly reliable. However, using the reliability of the measurements, you can estimate what the relationship between the two variables would be if they both were perfectly reliable. Most evidence I've seen regarding math tests, science tests, AP exams, and so on indicates the two item types do not measure different things. People who think that multiple-choice tests measure trivial facts and performance assessments measure higher-order processes frankly don't know much about measurement. There are literally millions of bad multiple-choice test questions that measure trivial stuff—but there are just as many bad open-ended questions that don't measure anything! Many open-ended tasks don't measure higher-order thinking and, in fact, well-constructed multiple-choice tests are *more* apt to elicit higher-order thinking processes. With well-constructed multiple-choice questions, people must engage in certain kinds of thought processes to get credit. On the other hand, in most cases when people answer open-ended items, you cannot be sure whether they have done any higher-order thinking or not. It's very easy for students responding to open-ended questions to parrot back answers in the form of what they know people want to hear. They get credit for doing higher-order thinking, because people say *that's the way I think, it must be higher order.*

Wainer and Thissen's conclusion (1993) to a different study of the AP exams (by the way, Howard Wainer and David Thissen are heavy-duty theoreticians, which I'll admit is a good reason to mistrust them) is, "A natural conclusion to reach from the weightings-associated constructed-response versus multiple-choice questions is that the former take more examinee time and resources to measure the same thing more poorly than the latter." In most examples of performance assessment, this conclusion holds. Constructed-response items do a worse job of measuring the same thing. It is up to the proponents of alternative assessments who say that this is not true, that the tests measure different things, to gather data and prove it.

Now let's get heavy duty. Let's get into predictive validity. When we talk about teacher evaluation or tests used for selection, the type of validity we are most interested in is predictive validity. Can we accurately predict who is going to be a good teacher and whom we should keep out of teaching? Unless you show me data based on external criteria, the best evidence of the predictive validity of the test is its reliability coefficient. In the AP exams, there is a multiple-choice section and an open-ended performance section. In five of the seven AP exams, if you create a total score including the performance component, the reliability is lower than if you only have the multiple-choice section. Everything else being equal, the predictive validity will be lower, too. It doesn't *have* to follow, and I am not saying it does, but that's the only evidence we have. Imagine you're testifying in court and all you're shown is this new performance assessment through which we are going to select teachers. Of course it appears to have differential impact on who gets into teaching and who doesn't, but we think it's great stuff. What if somebody has data like this and asks you which one is probably the fairest, which type of test do you think will make the most accurate prediction about who's going to be a good teacher? I'd say throw the performance section out. Most of the data we have argues that adding a performance assessment component hurts you from a predictive validity standpoint—the defender of the performance assessment in a court case will need to have some counterevidence.

Fairness, or equity, is an area of interest to many people. Here are a couple of quotes I've always loved, so I have to present them again. "Though time-consuming and costly, these methods are more humane, holistic, and ultimately more fair ways of determining student progress" (NCTE, Council Grams, May 1990). Now, my response to that is, "Where are these people coming from?" I don't have a clue. I read stuff like this and it drives me nuts, because I've spent 25 years of my life trying to make tests fair. When you build a test that students all over the country take, you do everything in your power to make it fair.

Then there's this one: "Reforms advocated by assessment experts such as a shift to performance-based testing could benefit minority students whose learning styles are ill-suited to multiple-choice tests" (*Education Week,* 4/18/90). They are? Who says so? Data in general don't agree with this conclusion. We have data going back a long way on the differential performance of blacks and whites, males and females, all kinds of groups, on multiple-choice tests and

performance tests. More often than not, the difference in performance between blacks and whites on performance tests is larger than the difference between blacks and whites on multiple-choice tests.

Here's another quote: "Majority/minority group differences on a California bar exam were not ameliorated with the addition of a performance section. In fact, when the results were adjusted for unreliability in the ratings, mean differences were larger than those that were observed for the multiple-choice portion of the test" (Dunbar, Koretz, and Hoover, 1991). In most cases that's the way things turn out, but special attention should be paid to the following part of the quote, ". . . when the results were adjusted for unreliability in the ratings . . . ." When somebody shows you some data in which the difference between blacks and whites or males and females on a performance task is smaller than it is on a multiple-choice test that is supposedly measuring the same thing, the first thing to ask about is reliability. If you have a test that has zero reliability, there probably won't be group differences. What a heavy technical idea!

The last topic I'll address is feasibility. I often talk to people about implementing performance tasks as an additional part of the teacher evaluation process. The first question people ask when they decide to go into this whole hog is whether they can afford the scoring. They answer, "The teachers! They'll score all this stuff for us. They'll do all this stuff for us." And they did. England did go into performance assessment all the way and required teachers to score the tests. It took them an average of 44 hours. That's 44 hours taken from somewhere else, and they are being paid for it. I would want to have a lot of evidence that all kinds of wonderful things happened on the basis of this shift in instructional time. Of course, what actually happened is the teachers got madder than hell. In fact, the whole thing went belly up in England, primarily because of teacher rebellion. Des Nuthall, who was one of the people in charge of the English program, said, "I suggest the U.S. might learn two other things from the English experience. First, the cost of performance assessment both financially and in terms of time is immense; second, despite all the care and effort, some will still not view it as rigorous enough" (*Educational Leadership,* April 1992).

It also should be noted that teacher rebellion is much more likely in the U.S. than in Europe, where people are more accepting of performance assessment. They have a long history of open-ended

examinations. They don't question something like a reliability of 0.35, it just doesn't come up over there. But it comes up over here, and it will continue to come up over here. They also don't have all the history and issues associated with group differences in performance that we've tried to deal with in the United States for a long time.

The closing quote, from the Wainer and Thissen article mentioned earlier says, *give us some evidence*. They say more strongly than I would, "The data that we have seen are unambiguous. Whatever is being measured by the constructed-response section is measured better by the multiple-choice section. We have never found any test that is composed of an objectively and a subjectively scored section for which this is not true." That's awfully strong. I don't know if I would say it that strongly, although I'll be honest—I can't think of anything off the top of my head that counters this argument. Basically they are talking about caution and evidence. Especially in high-stakes situations where performance tasks are involved, we must face the fact that invariably these assessments are going to be less reliable, they're going to be more expensive, in most cases they are going to show greater differential impact. Especially if it gets into a legal situation—one person didn't get a teaching certificate and another person did—and such evidence can be brought to bear, the people who are responsible for this use of performance assessments in high-stakes situations had better be ready with evidence of the kind I referred to earlier as consequential validity. If you have no counterarguments, you will be in big trouble.

That's my stirring end to this talk. Thank you all very much.

References

Bennett, R.E., Rock, D.A., and Wang, M. (1991). Equivalence of Free-Response and Multiple-Choice Items, *Journal of Education Measurement,* 28, 1, 77–92.

Dunbar, S.B., Koretz, D.M., Hoover, H.D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education,* 4, 4, 289–303.

Shovelson, R.J., and Baxter, G.P. (1992). What we've learned about assessing hands-on science. *Educational Leadership,* 49, 8, 20–25.

Wainer, H., and Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education,* 6, 2, 103–118.