Performance Assessment
and the Issue of Bias

Michael Ford

I'm not as much of a David Letterman fan as Ron Berk is. I used to watch Arsenio Hall, who did have an influence on me. You remember Arsenio, he used to talk about things that make you go "hmmm!"

One of the things that made me go "hmmm" was Ron's top 10 awards list. I don't mean this as a criticism. You may recall that number 8 on this list, "Most Politically Correct Procedure," was selecting a broad-based panel of judges. It occurred to me that if the label were changed from "Most Politically Correct" to "Most Equitable," then our perception of the importance of selecting a broad-based panel of judges might change.

The reason I begin here is that it seems to me that there has been or should be a change in perception. I am concerned about bias in testing because as performance assessments become more widespread, we need to look beyond the testing instrument to ensure that the evaluation is fair. We must be concerned with judges, with raters, and with other potential sources of bias.

I come to this as a layperson, with no particular expertise in testing or measurement. I am not involved in the certification of teachers. I am a teacher, a social scientist, and a political economist. I have been involved with National Evaluation Systems as a member of their equity review board for nearly 15 years. It is remarkable to be able to look back over those years and see the positive changes that have occurred.

National Evaluation Systems, and I suspect the testing industry in general, has become very good at identifying and correcting potential sources of bias in test items. As a college administrator, I am familiar with a number of tests that are used in higher education. Most of the tests that I encounter these days are a quantum leap beyond the tests I used to see just 10 or 15 years ago. Recently I was looking at a publication from National Evaluation Systems, and I was amused by some of the examples of bad items. They were typical of the types of items that we used to
encounter, not particularly with NES, but on all kinds of nationally administered tests. Here's an example of the kind of problematic item that I no longer see. "Which of the following is a characteristic of persons with Down syndrome? A. Larger than normal head. B. Obesity. C. Oriental-like folds over the eyes. D. Above average height." This was followed by a discussion of how alternative C, Oriental-like folds over the eyes, is a biased element. It was pointed out that it is inappropriate to liken Asians to people with Down syndrome. Response option C could easily be rewritten to read simply "Downward sloping skin folds over the eyes." Other biased items had content that was offensive to women or African Americans.

In recent years, as the Bias Review Committee has sat down to do its work, it has become a little boring. There is very little bias remaining for us to ferret out. We spend hours reading and reading and reading and not spotting anything. I feel a little guilty because I don't really feel I've earned my money. And that's all to the good.

The old evaluative criteria, defined by NES and used in similar form, I suspect, by many test manufacturers, had us looking for a variety of potential problems. First, we looked for bias in the item content. Then, we focused on language. Was there anything in the language of the item that could potentially disadvantage any culturally distinct group? Was there anything which gave offense in the way the item was presented? Did items perpetuate stereotypes, either negative or positive ones? Did the test as a whole reflect the kind of diversity extant in American culture? And finally, was the test fair overall?

Using these criteria, I think it is accurate to say that efforts to eliminate bias have resulted in the production of model instruments. The problem is that in the last couple of years a curve ball has been thrown to those of us who are interested in and concerned about potential bias. By this I am referring to performance assessments. Performance assessments are remarkably different from the kind of testing we are used to. They are different because what is being tested is the tool itself. The product of a performance assessment might be an interview, a videotape, an in-basket exercise, or a writing sample, to name just a few. Obviously, these kinds of responses are not machine scorable. They are scored by trained raters or judges. The question then becomes how to ensure that the judges are fair. What do you do?
Well, as Ron Berk noted, you select your judges carefully. You try to select people who will score the responses in such a way that the score accurately reflects the candidate's relative proficiency on the ability or skill the test is intended to measure. If you are interested in teacher performance, the videotape ought to reflect the skills of the teacher who presents it. But what do you do to ensure that the raters are evaluating each performance fairly and accurately? How can you be sure that the raters produce reliable and valid results? How can you be certain that an outstanding performance is always scored as such, regardless of individual differences in the teachers giving that level of performance?

You want good teaching to be consistently judged as good teaching, which means that judges should be reliable. There have already been, and I am sure there will continue to be, cogent and provocative discussions about reliability. But it surprises and distresses me to see relatively little discussion of the issue of bias in the context of performance assessments. The modest amount of attention to such bias that I have seen has been concerned with students, rather than teachers. I believe that there are very different issues that must be taken into account in order to ensure that performance assessments of teachers are done in a fair and unbiased way.

The old criteria which I mentioned earlier are not sufficient to ensure fairness in performance assessments. Performance assessments depend on the decisions of judges who are presumed to have the qualifications, the temperament, and the knowledge to assess the candidate's performance. When one is looking for potential bias in performance assessments, the focus must be on the judges or raters, rather than on the instrument. The judges must not only derive a set of standards that are not inherently contaminated by bias, but they must also bring to their work the attitudes, outlooks, and knowledge that allow them to assess the product or performance without bias, or with as little bias as humanly possible. What kinds of procedures would ensure that judges are free of bias to the same extent that the old machine-scorable tests were after they had been continually reviewed, amended, and corrected? There has been very little work done in this area to date and very little thinking about how to ensure that the judges who score performance assessments are unbiased. I have some suggestions, but I make these with a great deal of humility because I don't have much to guide me.
Raters and judges should first be free of biased attitudes that might affect their evaluation of an individual's response or product. How do you ensure that the raters are free of bias? A couple of ideas come to mind. I suspect that if they were implemented they might be criticized for being nothing more than politically correct procedures. However, I think they deserve some consideration, since the goal is to ensure fairness.

First, as part of the selection and training process, judges should be screened for cultural bias. Judges are employed to make critical decisions, and it is necessary to ensure not only that they have subject matter mastery and content mastery, but also that they bring to the task an attitude and approach that guarantees equality of opportunity for all people.

Second, judges or raters should themselves constitute a cross-section of the cultural groups in American society. It is desirable to include people who have had a wide range of experiences and are familiar with a wide range of approaches. This is likely to occur only if the panel is fully representative. While such diversity does not guarantee that a panel will be unbiased, it does make it more likely. That is why I started with a little poke at Ron Berk's awards list.

How do you achieve such a panel of judges? It's fairly simple. We have a great deal of experience in creating groupings of people who are representative. The bias review panel at NES is one such group. Creating similarly representative groups of judges and scorers is an essential element in their recruitment and training to minimize or eliminate bias.

Third, judges and raters ideally should have had experiences or training which has familiarized them with a broad range of permissible performances. People who have had greater exposure to different styles and approaches are more likely to recognize as valid those performances or responses which depart from the typical but which are, in fact, effective.

My experience at Hampshire College has influenced my thinking in this area. Hampshire College is an unusual institution in American higher education. It has been referred to as a pot where all the educational reforms of the 1960s were thrown in together, creating an educational environment that is almost a stereotypical product of the 1960s. We don't have grades. We don't have credits. The list of things we don't have is equaled in length only by the list of the unusual things we do have.
Students at Hampshire College are required to pass a series of exams. We call them exams, but they are not really exams at all. They are stages of educational development. During the first stage, students take courses or work in all four of the academic schools at Hampshire and present evidence that they have done so. The evidence need not be grades. In fact, students are not even required to take courses, although most do. What is required is that students present work demonstrating mastery of the mode of inquiry in each of the schools. In the second stage, the concentration stage, students work with faculty committees to organize and complete independent study programs. This program usually takes about two years to complete. As the final step of the program the faculty committee conducts an extensive oral examination based on their reading and evaluation of the work the student has completed. In some cases, particularly in the Social Sciences, this evaluation consists only of looking at all the papers that have been written. It is, in fact, an unusual case in which the portfolio contains only papers. Many other performances and products are typically included, such as internships or work products. In the Natural Sciences and the Humanities and Arts, there is great variation. Finally, at the last stage, students spend a year producing a senior-level product. It might be artwork or social research, but it must be at an advanced level. It must demonstrate the student's mastery of the content, methods of inquiry, and modes of analysis and presentation of the chosen area. Again, this product is evaluated by a committee.

Thus, the system we have in place at Hampshire College is essentially a performance assessment system. We have been using this system for some time and have had to face many of the issues that are now facing those who are involved with performance assessments of teachers.

One issue which is very important to us at Hampshire College, and is of course important in teacher certification as well, is that the decisions we make on the basis of the performance assessment be accurate. For all of us at Hampshire College, our worst nightmare is that the school develop or maintain a reputation for allowing young people to graduate who have not mastered the things that we say they have mastered. Related to this fear is the spectre of variation—that students who aren't very good might obtain the same kind of evaluation results as the students who are outstanding. This would not benefit the students, especially those who go on to graduate school, and it would certainly damage the
college's reputation. So how have we built a predictive quality into our evaluations, such that we have confidence that the evaluations truly reflect what the student has mastered?

There are a variety of ways, and often, through trial and error, we have learned some things from our student performance assessment which may be useful to people involved in teacher performance assessment. First, we use something which is referred to in the social science literature as the triangulation approach. In the triangulation approach a variety of faculty members work as a committee to review and evaluate a variety of aspects of a student's work, viewing the work from different vantage points, coming as they do from a variety of academic fields and from different schools within the college. As part of the process there is almost continuous feedback between the faculty members and the student. The faculty committee uses reasonable, common-sense standards to judge whether the student has mastered his or her concentration and whether he or she has done outstanding independent work.

The work of the committee is reviewed both at the school level and at the college level. This helps to ensure that within a school, and within the institution as a whole, we are all using the same standards. It prevents students who have not met these standards from graduating. This review is an important part of the system.

Thus at Hampshire College we use a panel of "judges" who view the work from different vantage points, who are representative of a range of different fields, and who have expertise in a variety of content areas and methodologies. Throughout the evaluation process there is continual feedback between the committee and the student. And finally, there is review at the school level and at the college level. What can we learn from this that is relevant to bias review? It seems to me that the judges are critical, absolutely critical. They have to be well trained, and they have to be screened. In addition, it is important that there be a process in place which ensures feedback between the judges, or scorers, and the candidate, so that there is a continuous process to determine if there is any kind of bias. Part of the problem in conducting a bias review is that the people involved sometimes feel as though they are being accused of being racist, sexist, or insensitive. The purpose of a bias review is not to accuse anyone of malfeasance or bad attitudes, but rather to ensure that the test is fair to all candidates. In order to make progress, we must recognize our limitations and work to overcome them. The trick lies in ensuring
that scorers come to the testing situation with attitudes that are most likely to result in unbiased judgments. As I said earlier, this is directly related to the range of experience of the scorers. It is important to select judges who have a wide range of experiences. We have to be sure that there is a breadth of judgment and that the judgment is sound. Raters must have expertise in their area, but it is also important that they be fair and aware of the potential for bias. Additionally, any calibration and recalibration that is done must take the potential for bias into account. Most of the material I have seen so far on training, calibration, and recalibration focuses exclusively on content issues and standards, without attention to bias. The potential for bias must be considered throughout the testing process.

Why the issue of bias is so important seems to be self-evident. Education is a crucial profession. It is also one of the most open avenues of upward mobility for minority group members. Therefore, it is crucial that the judgments which are made about who is to be admitted into the profession be free of any sort of conscious or unconscious bias, to the maximum extent possible.

As noted earlier, triangulation suggests that several judges, coming from different backgrounds, would be more effective than a single judge in making decisions about specific performances, given the complexity of making judgments about human behavior. Similarly, triangulation suggests that the objective of testing, as well as the steps involved in achieving that objective, can best be met through multiple approaches. Performance assessments make a great deal of sense in this light. At the same time, performance evaluations are not the only kind of evaluations we should use. Other evaluations also make sense, such as evaluations which look specifically at outcomes. In addition, there is still a place for traditional tests, which do some jobs exceedingly well and reliably. The triangulation approach suggests that the variety of data sources, research results, and analytical approaches that are inherent in these different approaches are much more likely to provide a sound basis for making judgments than a single source or approach. This is particularly true for judgments about whether a person is competent to do a particular job.

I think that many Americans today feel a great deal of cynicism and distrust toward politics and politicians. You see interviews in the *New York Times* where people say things such as, "I am going to leave this country" or "the real answer is term limits. Throw the rascals out, keep them circulating so they aren't in there long
enough to do continuing damage." Then there are those who say "It's just so distasteful. It's just so disappointing that I won't have anything to do with it." It seems to me that in the 1980s there was as much disillusionment or disappointment with education as there is today with politics. There was an avalanche of studies which described the sorry state of public education in this country and an almost equal number of studies that provided prescriptions for what ailed us. There were those who seemed ready to do away with the entire public education system as we know it and replace it with school choice, a system fueled by the dynamics of the marketplace, a panacea. They argued that this would lead to a better day.

I think that way of thinking is still alive. From my vantage point it seems that focusing on fair, equitable, and sound assessment is much more likely to lead to a democratic process of school reform than school choice would. My experience with school choice suggests that the problem is inequitable resource distribution. What that means in a situation of choice has yet to be resolved. Massachusetts had a school-choice program which I believe has been discontinued but which was in effect for a couple of years. It was a peculiar kind of school-choice program which allowed students to transfer from district to district. If a student transferred out of a school district, the district that received the student received funds equivalent to the sending district's per capita expenditure. Guess which kinds of students left. And guess what kind of communities ended up effectively subsidizing other districts. In almost every instance the students who left were white. They left school systems that were regarded as not as acceptable or excellent in their educational processes to go to ones that were better. The districts that lost students, like Springfield, essentially subsidized the more affluent districts that these students transferred to. All school-choice programs aren't that Draconian, but it does seem to me that school-choice programs, if they are going to work at all, need to confront the inequity issue. If they do so successfully, it seems to me the result is not likely to be a school-choice program, at least not as school choice is typically defined.

I think that a real focus on fairness in performance assessment and a triangulation of procedures is more likely than school choice to lead to equity in education. There are some things that can be done to help us move in this direction. Test professionals ought to continue to review their work in conjunction with policymakers.
Performance assessment, and other kinds of evaluations, can and ought to lead to clarification of the relationship between assessment results and public policy. There ought to be a triangulation of effort that brings together test professionals, policymakers, and front-line educators, so that the resource problems identified through tests can be addressed. This can lead to changes in policies and a redistribution of resources to address differences in performance that are caused by differences in wealth. I believe this approach is more likely to lead to an educational system with a real focus on excellence and fairness.

Performance assessment is a new direction in testing, and a very promising one. It is essentially a holistic approach, an effort to assess complex skills in a realistic way. I support this movement, but I do have some concerns. As I look at the literature in this area, I find very little evident concern about the potential for bias in performance assessments. In traditional tests, we looked for bias in the content and characteristics of the items themselves. In performance assessments, we must look at the judges. We must look as closely at the judges as possible sources of bias as we have looked at the items in the past. Until we do this, performance assessment has the potential to do real harm to minority group members.