The Role of Performance Assessment in Studies of Educational Achievement

Richard M. Wolf

Columbia University New York, USA
Teachers College

Abstract

Performance assessment, despite popular belief, has had a long history of use in education. However, that use has, by and large, been confined to the classroom. Teachers routinely use extended assignments, essays, projects, and reports not only as a way ofassessing student performance but also as learning activities. Today, a number of people are advocating the use of educational achievement. There are a number of theoretical, methodological, and practical problems involved in the use performance assessment in such programs and studies. These are identified and discussed. Suggestions for how these problems may be addressed are presented. It seems that it will take some time to solve these problems.

The measurement of educational achievement in large scale studies of student performance at the primary and secondary school levels is in the midst of a revolution. Ten, twenty and even thirty years ago, large scale studies of student achievement relied almost exclusively on the use of multiple choice or short answer questions. This is not to say that many nations did not make use of other forms of questions, notably essay questions, but their use was largely confined to examination systems that had a long history of use. Research studies relied primarily on the use of multiple choice questions because they could more efficiently cover subject matter domains and could be scored quickly and easily. The first mathematics study of the International Association for the Evaluation of Educational Achievement (IEA) used mathematics tests that consisted of about 90% multiple choice questions and 10% short answer fill-in items (Husen 1967).

Today, the situation is rather different. Much more extensive use is made of performance assessments. These are problem situations that require a considerable amount of time for students to work through a problem situation

and produce a response that may be quite extensive, such as a write-up of the results of an experiment or an extended essay. A number of people are actively promoting the use of performance assessment as a major and, in some cases, an exclusive way of appraising student proficiency. There appear to be two streams of thought contributing to this view. First, a number of people feel that multiple-choice and short answer questions are not able to measure the important proficiencies that students should be acquiring in their education. For these people, performance assessment is the sole way to measure higher order thinking skills and problem solving abilities. Second, there is a belief that performance assessment can serve as a lever for educational change. There even is a term for such a view. It is called "measurement-driven instruction" and reflects a belief that teachers and students will focus their time and energies on acquiring the proficiencies on which students will be tested. Accordingly, by changing the nature of the questions and problems that students are asked to answer, one can change an entire educational system. These two streams of thought have resulted in a considerable amount of enthusiasm for the use of performance assessments in examination systems and even in research studies. Advocates of the use of performance assessments, regardless of their motivation, are highly enthusiastic about its possibilities.

While many advocates of performance assessment see this as a new development, the history of appraisal in education shows that this is not the case at all. Multiple choice and short answer questions did not come into widespread use until the 1920's. Up until then, student performance was routinely appraised through the use of essay questions, oral questioning, and student products. The introduction of multiple choice and short answer questions was instituted mainly to overcome problems of extreme subjectivity in the appraisal of student performance and to achieve broader coverage of subject matter domains. Also, the much larger number of students entering the school systems of various nations required that efficiencies in appraisals be made. Performance assessment has been used continuously in a number of areas. This is especially true in vocational subjects such as welding where the production of welds meeting certain specifications has always been the major form of appraisal. In art education, the production of art objects or paintings has always been a dominant form of appraisal. Even in language learning, the testing of writing has continuously relied on the production of essays.

Clearly, advocates of performance assessment have a long history of use on which to base their enthusiasm. However, this base rests primarily on experience at the classroom level. Teachers in all school subjects have long been used to having students do projects, write reports, and engage in other

activities that are designed not only to serve as learning activities but also for purposes of appraisal. As long as teachers are able to closely monitor such activities and observe students at work on such projects, the situation has worked well. Problems arise, however, when an attempt is made to standardize performance assessment across classrooms, schools, regions of a nation, and even between countries. Standardized performance assessments require considerable compromises in order to work. Oftentimes the effort to standardize performance assessments has resulted in widespread dissatisfaction since the kinds of tasks that would be appropriate in one setting may not be appropriate in another. Furthermore, when performance assessment is incorporated into a large scale research study or examination system, the time demand on both students and teachers increases dramatically. In England, it has been reported that a new national examination scheme required over thirty hours of student time plus incalculable hours of teacher time to prepare and grade.

The problem of deciding what to assess and how to assess it may appear straightforward but is not. Where there is a standard curriculum, decisions about exactly what to assess become extremely problematic since, with performance assessment, one is quite limited in what can be assessed due to time constraints. Uneasy compromises about what to assess are commonplace with the result that few, if any, are quite satisfied with the scope of an assessment. Decisions about how to structure performance assessments are also difficult since there are typically a number of different ways in which particular competencies can be assessed. Thus, even educators who are committed to undertaking performance assessment on a large scale face formidable problems.

The development of performance assessment tasks is clearly a difficult undertaking. While individual teachers can and do do this regularly in their own classrooms, standardizing such assessments across classes, schools, and regions of a country (let alone nations) is extremely difficult. At this time, it seems fair to say that experience in doing this is quite limited. Choices have to be made to keep testing within reasonable time limits and this means that what some consider to be important proficiencies to be assessed will be left out. Sampling of proficiencies in performance assessment is a major problem. In addition, research with performance assessment has shown that there is a large component of task specificity in each performance assessment task (Lane, Stone, Ankemann & Liu 1994). Lane and her co-workers have found that it requires about twelve to fifteen tasks in order to obtain an overall score or grade that has sufficient generalizability to provide a meaningful individual score. For this reason, many research and assessment enterprises report results

only at a school level since the time required to obtain a meaningful individual score would require too great.

Another problem that is encountered in the use of performance assessment centers around the scoring of student responses. There is often considerable variability in the scoring of student responses to performance assessment tasks. To some extent this problem is ameliorated through the use of carefully developed scoring guides and a training program for scorers. This has led to some successes in dealing with the problem of scoring, but in many cases it has not. Much more work is needed in this area if one is to obtain dependable scores. In addition to scorer problems, there are other problems that have been recognized only very recently. In many assessment programs, there are interactions between students and tasks and between raters and tasks. These interactions are often considerable. They arise when, for example, a student is given a task that he or she is particularly adept at or one which he or she is not adept at. Also, some scorers rate the responses to one task differently than they rate the responses to another task. These factors can conspire to make scores on performance tasks rather undependable not only at an individual student level but also at a school level.

That problems exist in the use of performance assessment in large scale studies of educational achievement and examination programs can not be ignored. They are indeed formidable. The consequences can also be severe. At the individual student level, it can result in misclassification of the level of achievement which, if tied to a high stakes decision, can lead to incorrect classification as to passing or failing. At the school level, it can result in serious misrepresentation of the performance level of the school. Finally, in international studies of achievement, it can lead to incorrect rankings in league tables of results of achievement testing. In a recent large scale performance assessment in the state of California in the United States, for example, it was found that the percentage of truly superior students in a school could only be said to be between 19 and 41 - a rather imprecise finding (Cronbach, Bradburn & Horvitz 1994). The publication of such results in newspapers in that state created immense problems.

A major reason for these problems lies in the fact that performance assessment is very different from traditional testing and that the models and techniques that have been used to treat traditional test data simply do not work with performance assessment data. Considerable research and experience with this newer form of testing will be needed before we can use it with confidence. Whether this can be achieved in a reasonable time is an open question. Politicians, it seems, are demanding that schools demonstrate that they are

doing the job they are supposed to be doing and producing educated graduates. Many politicians are demanding that complex performance assessments be undertaken despite the fact that educational bureaucrats are ill-prepared to do this. The recent experience in England (Black 1994) indicates the kinds of problems that this can create.

There is no easy solution in sight. Considerable research must be done and experience accumulated before performance assessment can be used in large scale research and examination systems with an acceptable level of confidence. This will undoubtedly require years of work. Until then, it seems that educators will need to rely on forms of testing that they are familiar and comfortable with. This means the continued use of multiple choice and short answer tests. This is not as bad as it sounds. Multiple choice and short answer questions can be used to test a number of higher order thinking and problem solving abilities. Consider, for example, the following mathematics item that was used with thirteen year olds in the first IEA mathematics study:

The floor of a room is covered with wooden rectangular blocks. When blocks measuring a by b inches are used, M blocks are needed. If blocks fit exactly, how many blocks will be needed if each block is measured x by y inches?

A. $\frac{Mab}{xy}$ B. $\frac{ab}{Mxy}$ C. $\frac{(a+b)M}{x+y}$ D. $\frac{ab \cdot xy}{M}$ E. $\frac{Mxy}{ab}$

This is an example of a multiple choice question that is clearly testing problem solving abilities. Other examples can be given from other IEA studies as well as other testing programs that show how multiple choice questions can be used to test more than the acquisition of knowledge. It would seem that such questions should be used, perhaps along with performance assessment tasks, while research and experience with newer forms of assessment proceeds. This would seem to be a reasonable solution to the problem of assessing student performance. If policy makers can be shown how multiple choice and short answer questions can be used to test thinking and problem solving skills while research proceeds with performance assessment, they may be willing to continue the use of such test items until increased competence in the use of performance assessment is achieved. In addition, the use of multiple choice and short answer questions can increase the coverage of the subject matter domains to be tested.

Bibliography

Black, P.J. (1994). Performance Assessment and Accountability: The Experience in England and Wales. *Educational Evaluation and Policy Studies, 16*(2), 191-203.

Cronbach, L.J., Bradburn, N.L. & Horvitz, D.F. (1994).*Sampling and Statistical Procedures Used in the California Learning Assessment System* Report of the Select Committee to the California State Department of Education.

Husén, T. (1967). *International Study of Achievement in Mathematics*. Stockholm: Almqvist & Wiksell.

Lane, S., Stone, C.A., Ankemann, R.D. & Liu, M. (1994). Reliability and Validity ofa Mathematics Performance Assessment. *International Journal of Educational Research, 21* (3), 247-266.