DOCUMENT RESUME

ED 333 008 TM 016 429

AUTHOR Mehrens, William A.
TITLE Using Performance Assessment for Accountability Purposes: Some Problems.
PUB DATE 11 Apr 91
NOTE 33p.; Paper abridged from a paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; Comparative Analysis; *Educational Assessment; Elementary Secondary Education; *Evaluation Problems; Licensing Examinations (Professions); Literature Reviews; *Multiple Choice Tests; Student Evaluation; Teacher Certification; Teacher Evaluation; *Testing Problems; Test Use
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

Problems with performance assessment (PA) and multiple-choice tests (MCTs) are outlined, with reference to the literature on accountability. PA for individual teachers who should integrate their assessments with their instruction; PA as a supplement to more traditional examinations for licensure decisions; and some limited, experimental tryouts of PA for other accountability purposes are supported. The anti-MCT demagogues, and making PA the latest fad are not supported. Reasons for PA's popularity include: old (but inaccurate) criticisms of MCTs in terms of bias, irrelevant content, and measurement of only recognition; cognitive psychologists' belief that many parameters that they want to study require formats other than MCT questions; increased concern that MCTs delimit the domains that should be assessed; wide publicity of the Lake Wobegon effect of teaching too closely to MCTs; and claims that teaching to MCT formats has deleterious instructional/learning effects. PA problems vary depending on several dimensions, such as secure versus non-secure assessments, matrix versus every student assessment, and accountability versus instruction. PAs have difficulty meeting the five "apple" criteria required of high-stakes tests used for accountability purposes: administrative feasibility, professional credibility, public acceptability, legal defensibility, and economical affordability. It is concluded that MCTs measure some things very well and efficiently; however, they do not measure everything and their use can be overemphasized. PAs can measure important objectives that cannot easily be measured by MCTs. A 52-item list of references is included. (RLC)

4-11-91
Draft

USING PERFORMANCE ASSESSMENT FOR ACCOUNTABILITY PURPOSES: SOME PROBLEMS

William A. Mehrens
462 Erickson Hall
Michigan State University
East Lansing, MI 48824
(517) 355-9567

Abridged from a paper presented as a part of a symposium: Frechtling, J. Performance assessment and accountability programs: Match or mismatch?, given at the 1991 AERA Annual Meeting, Chicago.

DRAFT. Please do not quote without permission.

USING PERFORMANCE ASSESSMENT FOR ACCOUNTABILITY PURPOSES: SOME PROBLEMS

William A. Mehrens

DEFINITION OF PERFORMANCE ASSESSMENT

As Fitzpatrick and Morrison pointed out twenty years ago, "there is no absolute distinction between performance tests and other classes of tests" (1971, p. 238). The distinction is the degree to which the criterion situation is simulated. Typically what users of the term mean is that the assessment will require the examinee to construct an original response. Some people seem to call short answer questions or fill in the blank questions performance assessments. However, it is more common in performance assessment for the examiner to observe the process of the construction so there is heavy reliance on observation and professional judgment in the evaluation of the response. One of my favorite examples of a performance assessment question has been graciously provided by Steve Koffler (1990).

> MEDICINE: You have been provided with a razor blade, a piece of gauze, and a bottle of Scotch. Remove your appendix. Do not suture until your work has been inspected. You have fifteen minutes.

FAD VERSUS ADVANCEMENT?

It is easy to be impressed with the enthusiasm, energy, and optimism displayed by those doing research on performance assessment. However, it is impossible to be impressed by the lack of objectivity or scientific rigor of many of those advocating the current use of performance assessment. Unfortunately, some have put on their advocacy hats before the data support it.

A simple statement of my position is that I am in favor of performance assessment for individual teachers who should integrate their assessments with their instruction; I am in favor of performance assessment as a supplement to more traditional examinations for licensure decisions;[1] and I am in favor of some limited, experimental tryouts of performance assessment for other accountability purposes. Many questions must be answered and problems must be overcome before it should be used on a wide-scale basis. Further, I am "anti" the anti-multiple-choice demagogues; and I am against turning performance assessment into the latest fad.

> One of the most important reasons for the continuing existence of the educational pendulum is that educators rarely wait for or demand hard evidence before adopting new practices on a wide scale (Slavin, 1989, p. 753).

WHY "NEW" PERFORMANCE ASSESSMENTS?

The first point that should be stressed is that performance assessment really is not new. It was employed when the Gilead Guards challenged the fugitives from Ephraim who tried to cross the Jordan river.

> 'Are you a member of the tribe of Ephraim?' they asked. If the man replied that he was not, then they demanded. 'Say Shibboleth.' But if he could not pronounce the 'sh' and said Sibboleth instead of Shibboleth he was dragged away and killed. As a result 42 thousand people of Ephraim died there at that time (Judges 12: 5-6, The Living Bible).

That obviously was a performance examination. I point it out because I heard a speaker at a recent professional meeting say that "performance tests have only been around a couple of years." That person obviously had some gaps in his historical knowledge. Even a reading of the twenty year old Fitzpatrick and Morrison chapter in the second edition of Educational

---

[1] This is due to the high costs of false positives in licensure.

Measurement (1971) could have prevented such an inaccurate statement. However, it is true that the popularity of talking about performance assessment as the latest solution to our educational problems is a new phenomena.

Like all "new" (or recycled) developments (fads) performance assessment is backed by a very large number of people for a variety of reasons. Several of the major reasons are as follows: (1) the old (but inaccurate) criticisms of multiple choice tests; (2) the belief of cognitive psychologists that many of the things they are interested in assessing require formats other than multiple-choice questions; (3) the increased concern that multiple choice tests delimit the domains we should be assessing; (4) the wide publicity of the Lake Wobegon effect of teaching too closely to multiple-choice tests; and finally, (5) claims that there are deleterious instructional/learning effects of teaching to multiple-choice test formats. Certainly these five points are related and overlapping, but they will be discussed separately.

TRADITIONAL (BUT INCORRECT) CRITICISMS OF MULTIPLE-CHOICE TESTS

There have been three main criticisms of objective paper/pencil tests: They are biased, they measure irrelevant content, and the format demands only the ability to recognize an answer--not to actually work problems.

Bias

This paper is not the place to refute the bias charge, but much has been written about that issue and there is a great deal of evidence that most objective tests have very little bias.

Irrelevant Content

The issue of content relevance is related in part to the issue of whether the multiple-choice format can only be used for a limited number of educational objectives/goals. But the issues are separable. To give you a flavor of the criticism, consider the following quote:

> We're spending hundreds of millions of dollars on tests that don't tell us anything about what kids know or know how to do (Shanker, cited in Putka, 1989).

While the above quote was directed more at existing commercial standardized tests than the objective format per-se, the rhetoric stems at least in part from incorrect beliefs about what multiple-choice tests can measure. In addition to the concern about irrelevant content, there is the concern about the narrowness of the content and its mismatch with the curriculum (see Baker, Freeman, and Clayton, 1991).

There will never be universal agreement about the goals/objectives of education. However, one must keep in mind how standardized multiple-choice achievement test domains are determined. They are determined based upon very thorough reviews of existing curricula guides and textbooks. These, one would assume, have been developed and/or adopted because they have some match to the goals of the local schools. Most parents do want their children to learn the content domains sampled by multiple-choice standardized achievement tests.

M-C tests measure only recognition

Consider the following quotes:

> Standardized multiple-choice tests have drawn increasing fire as too simplistic, measuring the ability to recognize knowledge rather than the ability to think and solve problems, an important skill in today's jobs (Fiske, 1990, p.1).

> It's testing for the TV generation--superficial and passive. We don't ask if students can synthesize information, solve problems or think independently. We measure what they can recognize" (Darling-Hammond as quoted in Fiske, 1990, p.B8).

The notion that multiple-choice items can not measure higher-order thinking skills is unfortunate and incorrect. Forsyth has over the years given any number of talks illustrating that multiple-choice achievement test items can tap higher-order thinking skills (see, for example, Forsyth, 1990a). If his examples have not convinced the doubtful, they simply are not open-minded about it--or perhaps they don't think at a high enough level. Look at the sample multiple-choice questions sent to students who register for the SAT. You could not possibly answer those questions without engaging in some problem solving and/or higher order thinking.

COGNITIVE PSYCHOLOGISTS' INFLUENCE

Over the past decade or so, many individuals have been hypothesizing on "what cognitive psychology seems to offer to improve educational measurement" (Snow and Lohman, 1989, p. 263). As Snow and Lohman state, "measurement experts now need to know much more of cognitive psychology than they were taught or are likely to learn without a precis" (p. 263). It is impossible to argue with that point. However, it is possible to argue about just what the measurement implications are from the current writings of cognitive psychologists. As Snow and Lohman inform us, there are many controversies among cognitive psychologists (p. 264), cognitive psychology has its critics (viewed as just by Snow and Lohman), and the field has been fragmented and noncumulative (p. 270). Snow and Lohman suggest that the implications of cognitive psychology are largely for measurement <u>research</u> (p.312), and that "cognitive psychology has no ready answers for the educational measurement

problems of yesterday, today, or tomorrow" (p. 320). Other researchers generally seem to agree with this assessment (see Ohlsson, 1990; Lesgold et al., 1990; and Linn, 1990). None of the researchers referenced are suggesting wide adoption of their exploratory research.

Based on his research, Siegler warns us

> that even seemingly well-documented cognitive psychological models may be drastically incorrect, and that diagnoses of individuals based on these models could only be equally incorrect....the time does not seem ripe to advocate their use in the classrooms (1989, p. 15).

All this brings to mind the question of how to determine what is new and what is true in current cognitive psychology. Bader, in discussing the "new" reading objectives quotes Roe, Stoodt, and Burns who stated that "activating schemata involves recalling existing schemata that are related to a specific subject and relating these schemata to the content being read. Students must activate appropriate schemata." Bader asks us to contrast this statement with the following one by Huey published more than 80 years ago in 1908: "When reading, the learner forms meaning by reviewing past experiences that given images and sounds evoke." (Both quotes taken from Bader, 1989, p. 627).

I suspect current theorists would argue that the schemata theories are different from what Huey said in 1908, but I am drawn to a statement Bracey made recently:

> No current construct is trendier, squishier, and murkier than that of 'schema' ...(1991, p. 416).

In spite of the somewhat cautionary tone of the above paragraphs, I am convinced that cognitive psychologists do have something to offer those of us in measurement. However, I, like Snow and Lohman, think that it is primarily in terms of helping measurement specialists to develop new, and hopefully

better, theories. We should not jump on any "performance-assessment-for-accountability" band wagon.

DELIMITED DOMAIN

Partly as a result of the cognitive psychologists' influence there has been increased concern that multiple-choice tests can not assess all the important domains of educational goals/objectives. This fact has been known almost forever, and the concern is not totally new. Across the decades measurement specialists have agreed that objective tests can not adequately cover all objectives. For example, no one believes they are a good way to measure perceptual motor skills. However, as measurement driven instruction has increased, the concern about the delimitation of the measured domains has increased.

Cognitive psychologists distinguish between declarative and procedural knowledge (or content knowledge and process knowledge). As Snow and Lohman point out, all cognitive tasks require both types of knowledge, but different tasks differ in the relative demands they place on the two. It is generally accepted that some types of procedural knowledge are not amenable to multiple-choice types of assessment. The increased (and in my view correct) push for procedural knowledge goals has led to an increase in the attempts to engage in performance assessment. However, this should not result in a replacement of objective tests.

As Weinstein and Meyer (1991) make clear in their chapter on the implications of cognitive psychology for testing, many different educational tasks require simple recall--particularly in the lower grades and in introductory courses. Further, experts differ from novices in their knowledge

base, and research suggests "that domain knowledge is a necessary but insufficient condition for acquiring strategies and expertise" (1991, p. 42). One example of the research on the importance of a knowledge base is the effects of prior knowledge on reading -- and as the quote earlier by Huey suggests, that is not a new idea. [2]

Collis and Romberg, advocates of performance assessment in mathematics, admit that multiple-choice items provide

> an efficient and economical means of assessing knowledge of and ability in routine calculations, procedures, and algorithms. All seem to agree that these skills are still an important part of mathematics education...(1991, p. 102, italics added).

In spite of my belief in the importance of procedural knowledge and the importance of doing some assessing by other than multiple-choice testing, I remain puzzled by some of the writings regarding this "new" performance testing. Some suggest that multiple-choice tests are indirect and what we need are more direct measures of achievement. But cognitive psychologists focus on processes (such as metacognitions) which are not amenable to direct measurement. They demand indirect measurement (Weinstein & Meyer, 1991, p. 49). Baker, Freeman and Clayton were concerned with content-curriculum mismatch but found current textbooks did not "allow the development of deep understanding" (1991, p. 138), so for their research, they used new material--certainly creating more mismatch. Others, seemingly not too fond of the concept of measurement-driven instruction, wish to use performance tests to reform the curriculum, which seems a lot like being in favor of measurement-driven instruction to me. Baker et al. were also concerned with pressures to

---

2 See Hirsch (1988) for a supportive view of the importance of knowledge to read with comprehension or to be culturally literate.

test in "a relatively limited number of subject matters" (1991, p. 133), and Carlson suggests that there has been a narrowing of the curriculum as a result of not using performance assessment (cited in Rothman, 1990). But performance assessment certainly is less efficient at covering broad domains of subjects than are multiple-choice tests. As Finn correctly pointed out, the limited number of items on performance tests may narrow the curriculum even more (cited in Rothman, 1990).

Thus, there seems to be confusion regarding the domain issue. Some think the problem is that multiple-choice tests do not cover a broad enough domain. But performance tests will access narrower domains--perhaps in more depth.[3] Some are concerned with the curriculum-test mismatch and the efforts of educators to change the curriculum to increase the match -- these people generally see measurement-driven instruction as a bad thing. Others are interested in using new assessment procedures to reform the curriculum and hope there is a teaching to the assessment. All of this confusion gets compounded by those who refuse to separate the issues of content vs. form of an exam (which are related, but not identical issues).

LAKE WOBEGON EFFECTS

High stakes tests can lead to teachers teaching too closely to the test, thus raising scores without raising the inferred achievement. Some advocates of performance assessment suggest that it is appropriate to teach directly to

---

[3] Actually the evidence regarding whether multiple-choice tests and other assessments cover the same domains is quite mixed. Some research suggests the same domains/constructs are being measured--other research suggests that there are some differences (Ackerman & Smith, 1988; Bennet, et al., 1991; Birenbaum & Tatsuoka, 1987; Farr et al., 1990; Martinez, 1990; Traub & Fisher, 1977; Traub & MacRury, 1990; Ward, 1982; Ward et al., 1980).

that type of assessment because the instructors will be teaching appropriate material in ways they ought to be teaching it. Consider the following quotes.

> teaching to these [California Assessment Program] tests is what we want, because the tests are 100% connected with real-world on the job performance (Honig, cited in Pipho, 1989, p.263).

> if schools spend three or four weeks a year teaching to a performance based test, at least they'll be teaching things they ought to be teaching in ways they ought to be teaching it (Shavelson, cited in Rothman, 1989, pp. 12-13).

However, those who feel that performance assessment is the solution to teaching to the test are sadly mistaken. Their reasoning misses the point about inappropriate test preparation. They basically ignore the domain/sample problem that is exacerbated when one delimits the sample as one must in a performance assessment.

DELETERIOUS INSTRUCTION

Tied to all the above issues is the belief that if one tests via a multiple-choice test, and if one instructs so that the students will do well on the multiple-choice test, the instruction must be deleterious; however, if one assesses via performance measures, the instruction will be beneficial.

It is true that the format of the assessment will have <u>some</u> effect on instructional practices, that this effect will be greater if the assessment is for high stakes accountability decisions, that answering multiple choice questions is not a task that is done a lot outside of school, and that excessive instruction tied too closely to an unrealistic <u>form</u> of assessment is a poor instructional strategy. Nevertheless, it is <u>not true</u> that performance assessment will necessarily lead to high quality instruction. The Honig and Shavelson quotes above are just not true. The California Assessment Program's (California State Department of Education, 1989) five performance items in

math are certainly not "100% connected with real-world on the job performance." Further, teachers could spend time teaching correct answers to these questions without "teaching things they ought to be teaching in ways they ought to be teaching it."

Certainly many teachers would not say that the performance assessment of teachers has resulted in increased learning about how to teach. Further, I submit that if student performance measures become the criteria for teacher or school accountability, teachers will complain about those measures also. It is important to keep in mind Linn's admonition that we need to do more than just assume that the alternatives to multiple-choice items will have no bad side effects of their own (see Moses, 1990).

Again, I have perhaps sounded cautionary -- that is the role of a person trying to contain a fad. However, writing assessment has probably increased the instruction of writing and that is a good thing.[4] I suspect performance assessment of safety procedures in the science laboratories might increase the efforts of teachers to teach safety procedures, and that would be a good thing. But we must be somewhat prudent in our charges regarding the ills of multiple-choice tests and our claims about the wonders of performance assessment for instruction.

---

[4] Evidence appears mixed on this. Seventy-eight percent of California junior high school teachers said state writing assessment increased the number of writing assignments given to students (Moses, 1990). However, 1988 NAEP data allow the authors to conclude that "the recent interest in encouraging writing across the curriculum does not appear to have been carried out in practice" (Applebee, et al., 1990, p. 7).

PROBLEMS WITH PERFORMANCE ASSESSMENT FOR ACCOUNTABILITY

IMPORTANT DIMENSIONS

Like other forms of assessment, the particular problems that are likely to be faced with performance assessment vary somewhat depending on a variety of dimensions such as (1) secure vs. non-secure assessments, (2) matrix versus every pupil assessment, and (3) accountability vs. instruction.

Secure vs. non-secure instruments

One extreme disadvantage of performance assessment is that, with only a few questions, there is no way to keep the exact content of the exam secure. Once performance assessments have been used, they cannot be reused to test the same higher-order thinking process. One can memorize the answer to a higher-order question just as well as one can memorize an answer to a basic-skills question. Thus, performance assessments will have to be new each year -- adding to the developmental costs and making across-year-comparisons of growth very difficult.

Baker, Freeman, and Clayton took a different approach. They have suggested that

> only if the tasks and scoring criteria are made public ...can teachers guide students to meet such standards, and then only if the same tasks are used (1991, p. 137).

While I grant that this may be done without corrupting the inference for some physical performance tasks (e.g. diving), performance assessment tasks that have a metacognitive component do not allow for such release and reuse of the tasks.

Matrix sampling vs. every pupil testing

Different cost issues arise with these two methods. Assessments that would be cost prohibitive for every pupil testing may be reasonable in a matrix sampling approach. However, this makes the assessments much less useful to individual teachers. Further, some high stakes tasks such as those used for licensure and high school graduation requirements demand every pupil testing.

Accountability vs. instruction

The title and thrust of this paper is on the use of performance assessment in accountability programs. Yet most of the research and rhetoric regarding the advantages of performance assessment has been in the realm of individual pupil diagnosis. When one switches from local classroom assessment for individual diagnostic purposes to mandated assessment for accountability purposes, different issues arise. Most measurement experts I know believe that if you use performance assessment for high-stakes accountability purposes, the same kinds of problems as have occurred with multiple-choice tests will exist.

High-stakes tests used for accountability purposes need to meet what Baratz-Snowden (1990) has referred to as the five "apple" criteria: Administratively feasible, Professionally credible, Publicly acceptable, Legally defensible, and Economically affordable.[5] I maintain that performance assessment is likely to have difficulty meeting all of those standards. Currently it appears to meet the professionally credible and

---

[5] Admittedly, her writing pertained to licensure tests, but I believe the generalization of the criteria to accountability assessment is reasonable.

publicly acceptable criteria -- but that is because it is in the fad stage. More careful scrutiny may change that.

ADMINISTRATIVELY FEASIBLE/ECONOMICALLY AFFORDABLE

Because resources are always limited, the costs of performance assessment must be of great concern. ETS has reported that

> one state with a strong commitment to educational assessment found that redesigning its state program around performance tasks would increase by tenfold the cost of the existing state assessment program (1990, p. 6).

Given my belief that most performance exercises are not reusable without distorting the inference, there are some very real questions about the developmental costs in performance assessment for accountability.

Even after performance assessments have been developed, the costs of administering and scoring them are high. Frequently special equipment is needed for administration and it is not feasible to have enough copies for simultaneous administration. Consider, for example, the four components being planned for an assessment of teachers' laboratory skills (Wheeler, 1990). There will be a pre-observation questionnaire, a pre-observation conference, an observation and a post-observation conference. The observation is to last 30 to 45 minutes. Observers in the pilot study were trained for three days. All this will certainly be expensive.

PUBLICLY ACCEPTABLE

So far the performance assessment advocates have done a good job with public relations. But, as with multiple-choice tests, once they have been used awhile for accountability purposes and the teachers complain (correctly) about their lack of validity for accountability inferences, there may be a reduction in public acceptability. Once the public understands that the costs

will be substantially higher, one might expect some loss of acceptance of the process.

LEGALLY DEFENSIBLE

> Legally, performance assessment is considered a test (Nathan & Cascio, 1986, p.1).

Whether that is how all courts would decide the issue, prudent individuals developing performance assessments for high-stakes decisions would be wise to act as if this were the case.[6] Experts for plaintiffs generally psychometrically attack tests based on whether the Standards (AERA,APA,NCME, 1985) have been followed. One should expect them to do the same for performance assessment. Whether performance assessments can meet the various psychometric standards of reliability, validity, etc. is doubtful. But other legal concerns also need to be considered. For example, if there is any disparate impact on protected groups, how might one deal with the fact that graders may be aware of the group status of the students? If there is debate about the scoring process will there be documentation of the performance so rescoring can occur?

PROFESSIONALLY CREDIBLE

Professional credibility pertains at least to three overlapping groups: teachers, those involved in teacher education, and psychometricians. Because of effective P.R. and face validity, performance assessment probably has more credibility than multiple-choice testing for the first two groups. It is impossible to know if that will continue if performance assessment becomes

---

[6] See Watson v. Fort Worth Bank and Trust, 1988, for a discussion of this issue in employment testing.

widely used for accountability. Certainly wide use would result in more scrutiny than such assessments have currently been given, and the whole movement could implode following such scrutiny. Psychometricians will hopefully place or withhold their stamps of approval based on evidence regarding the psychometric properties of the assessments. This may place them at a different place on the credibility continuum from those individuals who claim that psychometric properties such as reliability do not matter (who cares about random error anyway -- if we are measuring the right thing).

Let us turn to a discussion of some specific psychometric issues: validity, reliability, scoring/scaling/equating, and bias.

Validity

Generally, psychometricians believe it is important to validate new approaches to testing before any wide implementation (see Nickerson, 1989). Unfortunately, others say validity is a "red herring" (Carlson cited in Rothman, 1990, p. 12).

Performance assessments have face validity -- or what Popham (1990), a veritable virtuoso of verbosity says can be more pedantically described as verisimilitude. Face validity helps in the acceptance of an assessment procedure. Some level of face validity is essential for public credibility. But it does not take the place of real validity and is simply not sufficient. Yet, many of the advocates of performance assessment act as if it is.

In studying the validity of performance assessments, one should think carefully about whether the right domains are being assessed, whether they are well defined, whether they are well sampled, whether--even if well sampled--one can infer to the domain, and what diagnostically one can infer if the performance is not acceptably high.

Correct Domains?

A wish to assess the correct domains was a major reason for implementing performance assessment, and I am, in a general sense, in favor of what cognitive psychologists and reform educators are stressing. Nevertheless, the appropriateness of performance domains are as subject to debate as are those domains assessed via paper/pencil tests. As mentioned earlier, multiple-choice tests do not measure everything. But neither do performance assessments. And some domains being proposed for performance assessment can much more efficiently be measured by multiple-choice tests. In general performance assessment measures a narrower domain that m-c testing, but assesses it in more depth. Is this good? What narrow domains need to be assessed in depth?

Well Defined Domains?

If one is satisfied that the right domains are being assessed, one should still consider whether they are defined tightly enough. Critics of standardized tests have suggested that the domains are not well-enough defined in those tests. My feeling is that the domains of multiple-choice achievement tests that have been used for accountability purposes have been more tightly defined than many performance assessment domains.

Adequate Sampling?

The major problems for valid performance assessment relate to the limited sampling and the lack of generalizability from the limited sample to any identifiable domain. One of the generally accepted advantages of multiple-choice testing is that one can sample a domain much more thoroughly than by performance assessments. Because performance assessment takes more time, fewer tasks (questions) can be presented. Thus, the sampling of the domain is

less dense. For example in California, there were only five mathematics items on their performance assessment. One would be hard pressed to generalize to any curricular domain from such a limited sample.

Generalizability?

Even if sampling is adequate, there is the question of whether one can generalize from the sample to a larger domain. This is dependent upon the intercorrelations between the portions of the domain in the sample and those portions not in the sample. Certainly research has indicated that higher order thinking skills and problem solving are specific to relatively narrow areas of expertise and there appears to be little transfer from one subject matter to another on these constructs.[7]

But even within a subject matter area, generalizability is "iffy." As Herman has pointed out

> research in performance testing demonstrates how fragile is the generalizability of performance (1991, p. 157).

She gives as one example the research that indicates writing skill does not generalize across genres.

Or consider the generalizability of performance in a science laboratory assessment. Some research has been conducted in California on the development of a science laboratory assessment for new teachers. In their 1990 Final Report, Wheeler and Page wisely state that they do not know if their prototypic exercises will generalize

> across different science laboratory situations--grades K-12;, earth, life, and physical sciences; various types of lab activities; different groups of students; and different lab setting, including field trips. ...conclusions about the generalizability of the assessment should be

---

[7] See Norris (1989), for a discussion of both epistemological and psychological generalizability of critical thinking.

based on a large-scale field testing that includes many more types of situations (Wheeler and Page, 1990, 60-61).

At this point in time we simply do not have enough data indicating the degree to which we can generalize from most of the performance assessments that are being conducted. Much of the evidence we do have would suggest that generalizability is extremely limited.

Correct Inferences About Sample Performance?

Even if the domain is the correct one, it is well defined, the sample is adequate, and generalizability is possible, validity problems remain. One has been alluded to earlier. If the assessment is not secure, students will be taught how to do that particular task. This not only makes the inference to the domain inappropriate, it means one may make an incorrect inference about the sample performance. For anything other than a completely physical skill (e.g. diving), one is typically making an inference about the cognitive processes used. But one can memorize reasons as well as facts. Anytime one wishes to infer something like a metacognition, it is important that the assessment be secure.

Finally, a threat to validity that deserves mention is the lack of ability to make a very precise inference from a poor score on a performance assessment. If, for example, one accepts Anderson's (1983) theory of skill development, there are three stages: the declarative stage, the knowledge compilation stage, and the procedural stage. At which stage is an individual whose skill development is inadequate?

Reliability

There are several threats to reliability in performance assessment. One has to do with the small number of independent observations (the sampling problem discussed above). A second has to do with a lack of internal

consistency (also discussed above). A third has to do with the subjectivity of the scoring process -- to be discussed below.

The evidence for performance assessment reliability is apparently so low in so many instances that, in a "preemptive counterattack," some advocates of performance assessment have told us that reliability is not important. Some have gone so far as to suggest that measurement theory is wrong when it says reliability is a necessary prerequisite to validity. It is the critics that are wrong. Reliability refers to random error in a measurement, and if random error is too great, any <u>perceived</u> relevance of the assessment is illusory because nothing is being measured (Fitzpatrick and Morrison, 1971, p. 268). Thus, one can not possibly make any valid inference from the data.

The only performance assessment area that has reported much evidence on reliability has been writing assessment. There, the major evidence reported is <u>reader</u> reliability. It generally runs in the low .80s. To obtain this level of reliability is costly. It requires' careful selection of and extensive training of the raters, precise scoring guidelines, and periodic rechecking of rater performance. Other types of reliability are less often reported. For other areas of performance, I have heard rumors that preliminary evidence shows internal consistency reliabilities to be as low as .20. While there surely <u>must</u> be data I have not seen, I believe there are serious problems with the reliability of many performance assessments.

<u>Scoring, scaling, equating, and aggregating data</u>

Many issues arise concerning scoring, scaling, equating, and aggregating data. The major issues in these areas will be highlighted in the following sections.

Scoring

It is obvious that there is subjectivity in assigning the scores to a performance. This means that who does the scoring is very important for any test used for accountability. Some telling data regarding scoring by anyone having a vested interest in the results comes from the judgments of teacher performance by principals. State after state has obtained very negatively skewed distributions when principals score teacher performance. When assessing for accountability purposes, it is imperative to have performances scored by those who do not have a vested interest in the outcome. Having teachers score their own students' performances will not work. Further, if the school building or school district is being held accountable for the scores on performance assessments, the scorers must come from outside the district.

The issue of "what" is to be scored is also of considerable importance. Typically, "an examinee response is complex and multifaceted, comprising multiple, interrelated parts" (Millman and Greene, 1989, p. 344). One can either use componential or holistic scoring. As Millman and Greene pointed out, in either case, to develop the scoring criteria requires a clear understanding of what it means to be proficient in the relevant domain (which, in turn, assumes there is a good definition of the domain). Holistic scores are useless for diagnostic/prescriptive purposes so most advocates of performance assessment probably will opt for developing scoring profiles (see Wolf et al., in press). The Standards require that the reliabilities of the sub-scores need to be reported. Further, if the data are going to be used for diagnostic purposes, one should report the reliability of the difference

scores. It is my guess that these will generally be quite low. The profiles for students' performances will likely be so unreliable they are useless.

Scaling

Determining how to scale the data from performance assessments is another challenge. In his paper on the NAEP Proficiency Scales Forsyth (1990 b), convincingly argues that those scales do not yield valid criterion-referenced interpretation. Large scale performance assessments will likely be equally difficult to scale.

Equating

Because performance assessments yield fewer independent pieces of data, and because specific assessments should not be reused, the equating problems seem formidable. I realize that some states have some of the best experts in the nation working on this issue. I am not aware of what the proposed solutions will be. In any case, for longitudinal comparisons and fairness in accountability, the scores on different forms of performance assessments must be equated so that they represent the same level of achievement regardless of when the performance was assessed, which tasks were given, or which raters scored the performance.

Aggregating

Decisions about the unit of reporting will be difficult to make. Certainly for those performance assessments that are based on group activities the unit can not be the individual.[8] However, other types of assessment may lend themselves to individual reporting.

---

[8] See, for example the prototype math exercises for Maryland State Department of Education, 1990.

Ethnic group differences

As mentioned earlier, one of the agendas for moving to performance assessments is that some individuals believe paper-pencil tests are biased. Given the commonly used definitions of bias, the evidence does not support that position. However, some are hopeful that performance assessments will show smaller ethnic group differences. The results are not yet all in with respect to this hope but evidence on writing assessments across the nation do not show smaller differences between black and white performers than are obtained from multiple-choice tests. Further, the data will be more complicated to interpret due to the subjective scoring processes and the potential opportunity for scorers to allow ethnicity to influence their scores.

CONCLUSIONS/IMPLICATIONS

As measurement specialists have known for decades, multiple-choice tests measure some things very well and very efficiently. Nevertheless, they do not measure everything, and their use can be overemphasized. Performance assessments have the potential to measure important objectives that cannot easily be measured by multiple-choice tests.

CONTINUE RESEARCHING BUT DO NOT OVERSELL PERFORMANCE ASSESSMENT

Some research has been conducted regarding performance assessment but much more research is needed. Like Wolf, et al.(in press), I would call for "mindfulness" (p.4) in the performance assessment research, and hope the researchers would "be as tough-minded in designing new options as [they] are in critiquing available testing" (p. 38). Evidence regarding psychometric

characteristics must be gathered. One cannot "pursue these new modes of assessment ... on the mere conviction that they are better" (p. 41). Finally, I agree with Wolf, et al., and wish to emphasize that researchers should be "standing on the shoulders rather than the faces of another generation" (p. 8).

While continuing the research, performance advocates should not be overselling what performance assessment can do. Wiggins has suggested that

> It's wrong to say [performance assessments] were oversold; they were overbought (cited in Rothman, 1990).

I do not see it that way. I think they have been both oversold and overbought, and the sellers have not been truthful about competitive products.

While standing on the shoulders of another generation, performance assessment researchers should <u>not</u> be intentionally or unintentionally misinterpreting what that generation has accomplished and the still current values of paper-pencil assessments.

CONTINUE USING MULTIPLE-CHOICE TESTS

Most large scale assessments have added performance assessments to their existing array of efficient paper-pencil tests, not replaced them. This is good. There is no question but that the multiple-choice format is the format of choice for many assessments -- especially for measuring declarative knowledge.

CLOSING THOUGHTS/QUOTES

From at least one point of view, performance assessment is a good thing for measurement specialists, and education in general. It has resulted in

more money and more resources being devoted to assessment. This has opened up a whole new assessment industry. It should result in more research regarding the effects of testing on teaching and learning. Nevertheless, I agree with Haney and Madaus who suggest that

> the search for alternatives [to multiple-choice tests] is somewhat shortsighted (1989, p.683).

We also need to keep in mind a statement Lennon made more than a decade ago.

> To encourage the innocent to root around in the rubble of discredited modes of study of human behavior, in search of some overlooked assessment "jewels," is to dispatch a new band of Argonauts in quest of a non-existent Golden Fleece (1981, pp. 3-4).

Finally, we should heed the wisdom of Boring:

> The seats on the train of progress all face backwards; you can see the past but only guess about the future (1963, p.5).

REFERENCES

Ackerman, T.A., & Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12(2), 117-128.

AERA, APA, NCME (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Applebee, A.N., Langer, J.A., Jenkins, L.B., Mullis, I.V.S., and Foertsch, M.A. (1990). Learning to write in our Nation's schools: Instruction and achievement in 1988 at grades 4, 8, and 12. Washington, DC: Office of Educational Research and Improvement, US Department of Education.

Bader, L.A. (1989). Communicating with teachers--honestly. Phi Delta Kappan, 70(8), 626-629.

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock and E.L. Baker (Eds.). Testing and Cognition. (pp. 131-153). Englewood Cliffs, NJ: Prentice Hall.

Baratz-Snowden, J. (1990) RFP-National Board for Professional Teaching Standards. Washington DC: National Board for Professional Teaching Standards.

Bennett, R.E., Rock, D.A. & Wang, M.W. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28(1), 77-92.

Birenbaum, M. & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats--it does make a difference for diagnostic purposes. Applied Psychological Measurement, 11(4), 385-396.

Boring, E.G. (1963). History, psychology and science. Ed. R.I. Watson & D.T. Campbell, New York: Wiley.

Bracey, G.W. (1991). Backtalk. Phi Delta Kappan, 72(5), 416.

California State Department of Education. (1989). A question of thinking: A first look at students' performance on open-ended questions in mathematics, Sacramento, CA: Author.

Collis, K., & Romberg, T.A. (1991). Assessment of mathematical performance: An analysis of open-ended test items. In M. C. Wittrock and E.L. Baker (Eds.). Testing and Cognition. (pp. 82-130). Englewood Cliffs, NJ: Prentice-Hall.

Educational Testing Service. (1990). Annual Report, Princeton, NJ: Author.

Farr, R., Pritchard, R., and Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27(3), 209-226.

Fiske, E.R. (1990). But is the child learning: Schools trying new tests. The New York Times, Wed., January 31.

Fitzpatrick, R. and Morrison, E.J. (1971). Performance and product evaluation, in E.L. Thorndike (Ed.). Educational Measurement (2nd. ed.) (pp. 237-270). Washington, DC: American Council on Education.

Forsyth, R.A. (1990a). Measuring higher-order thinking skills. A presentation at the meeting of the Institute for School Executives. Iowa City, IA.

Forsyth, R. A. (1990b). The NAEP proficiency scales: Do they yield valid criterion-referenced interpretations? Iowa Testing Programs Occasional Papers. # 35. Iowa City, IA. Iowa Testing Programs.

Haney, W. & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, 70(9), 683-687.

Herman, J. (1991). Research in cognition and learning: Implications for achievement testing practice. In M.C. Wittrock & E.L. Baker (Eds). Testing and Cognition. (pp. 154-165). Englewood Cliffs, NJ: Prentice Hall.

Hirsch, E.D. Jr. (1988). Cultural literacy: Let's get specific. NEA Today, 6(6), 15-21.

Judges 12:5-6. The Living Bible.

Koffler, S. (1990). Personal communication.

Lennon, R.T. (1981, April). A time for faith. Presidential address at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Lesgold, A., Lajoie, S., Logan, D. & Eggan, G. (1990). Applying cognitive task analysis and research methods to assessment. In N. Frederiksen, R.Glaser, A. Lesgold, & M.G. Shafto (Eds.). Diagnostic monitoring of skill and knowledge acquisition. (pp. 325-350). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R.L. (1990). Diagnostic testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.). Diagnostic monitoring of skill and knowledge acquisition. (pp. 489-498). Hillsdale, NJ: Lawrence Erlbaum Associates.

Martinez, M.E. (1990, April). A comparison of multiple-choice and constructed figural response items. A paper presented at the annual meeting of the American Educational Research Association. Boston, MA.

Maryland State Department of Education. (1990). Maryland school performance assessment program: Protype Mathematics Task. Author.

Millman, J. and Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.) Educational Measurement, 3rd edition. (pp. 335-366). New York, NY: American Council on Education and Macmillan Publishing Company.

Moses, S. (1990). Assessors seek test that teaches. APA Monitor, 21(11), 36-37.

Nathan, B.R. & Cascio, W.F. (1986). Introduction: Technical & legal standards. In R.A. Berk (Ed.), Performance assessment: methods and applications. (pp. 1-50). Baltimore, MD: The Johns Hopkins University Press.

Nickerson, R.S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Norris, S.P. (1989). Can we test validly for critical thinking. Educational Researcher, 18(9), 15-20.

Ohlsson, S. (1990). Trace analysis and spatial reasoning: An example of intensive cognitive diagnosis and its implications for testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds). Diagnostic monitoring of skill and knowledge acquisition. (pp. 251-296). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pipho, C. (1989). Stateline. Phi Delta Kappan, 71(4), 262-263.

Popham, W.J. (1990). Face validity: Siren song for teacher-testers. In J.V. Mitchell, Jr., S.L. Wise, B.S. Plake (Eds). Assessment of teaching: Purposes, practices, and implications for the profession. (pp. 1-14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Putka, G. (1989). New kid in school: Alternate exams. The Wall Street Journal, November 16.

Rothman, R. (1989). States turn to student performance as new measure of school quality. Education Week, 9(10), 1, 12-13.

Rothman, R. (1990). New tests based on performance raise questions. Education Week, 10(2), 1, 10, 12.

Siegler, R.S. (1989). Strategy diversity and cognitive assessment. Educational Researcher, 18(9), 15-20.

Slavin, R.E. (1989). PET and the pendulum: Faddism in education and how to stop it. Phi Delta Kappan, 70(10), 752-758.

Snow, R.E. & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed). Educational Measurement (3rd ed.). (pp. 263-331). New York: NY. American Council on Education and Macmillan Publishing Company.

Traub, R.E. & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1(3), 355-370.

Traub, R.E. & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. In K. Ingenkamp & R.S. Jager(Eds). Tests and trends 8: Jahrbuch der Padagogischen Diagnostik. (pp. 128-159). Weinheim & Basel: Beltz Verlag.

Ward, W.C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6(1), 1-12.

Ward, W.C., Frederiksen, N. & Carlson, S.B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17(1), 11-30.

Watson v. Ft. Worth Bank & Trust. 108 S.Ct. 2777, 1988.

Weinstein, C.E. & Meyer, D.K. (1991). Implications of cognitive psychology for testing: Contributions from work in learning strategies. In M.C. Wittrock & E.L. Baker (Eds). Testing and Cognition. (pp. 40-61). Englewood Cliffs, NJ: Prentice Hall.

Wheeler, P. (1990, April). Assessment of laboratory skills of science teachers via a multi-methods approach. Paper presented in the symposium on Innovative assessment prototypes for the California New Teacher Project at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education. Boston, MA.

Wheeler, P. & Page, J. (1990). Development of a science laboratory assessment for new teachers, grades K-12. Final Report. Mountain View, CA: RMC Research Corporation.

Wolf, D, Bixby, J., Glenn, J., & Gardner, H. (in press). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), Review of Research in Education.