DOCUMENT RESUME

ED 355 257 TM 019 598

AUTHOR	Glowacki, Margaret L.; Steele, D. Joyce
TITLE	A Synthesis of the Research on Alternative Assessment Methods in Teacher Education.
PUB DATE	Nov 92
NOTE	26p.; Paper presented at the Annual Meeting of the Mid-South Educational Rese^rch Association (21st, Knoxville, TN, November 11-13, 1992).
PUB TYPE	Information Analyses (070) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Check Lists; *Classroom Techniques; Cost Effectiveness; *Educational Assessment; Higher Education; Literature Reviews; Multiple Choice Tests; Observation; *Rating Scales; Self Evaluation (Individuals); Simulation; Synthesis; Teacher Education; *Testing Problems; *Test Use; Time Management
IDENTIFIERS	*Alternative Assessment; Oral Examinations; *Performance Based Evaluation

ABSTRACT

This synthesis of research on alternative assessment methods in teacher education reviews literature on performance assessment, describing the various types of assessment discussed by M. Priestley (1982), including: (1) actual performance assessments (work-sample tests, identification tests, supervisor ratings, peer ratings, and self-assessments); (2) simulations (simulated performance tests, simulated identification tests, written simulations, and management exercises); (3) observational assessments (checklists, rating scales, and anecdotal records); and (4) oral assessment (oral examinations, interviews, and prepared presentations). The major problems identified in the use of performance-based assessments are those of time and cost. Some concern has been expressed that the domain covered by performance tests may be even more narrow than that assessed by multiple-choice tests. Many educators and researchers see the best assessment approach as the use of alternative assessments in conjunction with standardized multiple-choice tests. (Contains 38 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

□ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARGARET L. GLOWACKI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED355257

A Synthesis of the Research on Alternative Assessment Methods in Teacher Education

Margaret L. Glowacki and D. Joyce Steele
Evaluation and Assessment Laboratory
The University of Alabama

Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, Tennessee, November 11-13, 1992.

019578

BEST COPY AVAILABLE

ntroduction

Controversy regarding the use of standardized multiple-choice tests in education is a continuing topic of research and discussion, part of which is directed at teacher education and licensing programs. Anrig (1992) perceived the debate surrounding testing as becoming polarized.

> Many business people and elected officials love tests; they see them as the end-all and be-all that will solve all problems in education. On the opposite side, many educators, and particularly teachers, see the growth in accountability testing as an unqualified evil and call for the abolishment of standardized testing. (p. 3)

Many reasons have been given for the controversy surrounding multiple-choice tests. Murnane (1991) perceived the low correlation between scores on multiple-choice tests and scores on measures of teaching effectiveness as part of the controversy. The Southern Regional Education Board (1982) believed that the perception by the public that a large number of teachers are unqualified also has led to discussion and research regarding the use of multiple-choice tests in education.

Haney and Madaus (1989) estimated that standardized testing has increased between 10% and 20% annually over the last 40 years. As testing has increased so have complaints about the tests. Haney and Madaus listed what they perceived to be criticisms of standardized tests. They stated that the tests:

> • give false information about the status of learning in the nation's schools;
> • are unfair to (or biased against) some kinds of students (e.g., minority students, those with limited proficiency in English, females, and students from low-income families);
> • tend to corrupt the processes of teaching and learning often reducing teaching to mere preparation for testing; and
> • focus time, energy, and attention on the simpler skills that are easily tested and away from higher-order thinking skills and creative endeavors – the Achilles' heel of the nation's education system today, in the view of many observers. (p.684)

Neill and Medina (1989) discussed many of the same criticisms. They stated that standardized multiple-choice tests have come to dominate education in the United States in the last twenty years. They maintained that although standardized tests are sold as objective and reliable scientifically developed instruments. the basic psychological assumptions underlying construction of these tests are often erroneous, and that reliability and validity studies of these tests are often inadequate. Other criticisms include bias, narrowing the curriculum, and controlling placement of students in various educational programs.

Mehrens (1991) disagreed with many of the criticisms. He stated that there is a great deal of evidence to indicate that most objective tests contain very little bias and that the tests can measure higher-order thinking skills. Forsyth (1990) as cited by Mehrens has illustrated that multiple-choice test items can tap higher-order thinking skills. Neill and Medina (1989) stated that multiple-choice tests narrow the curriculum, but Mehrens maintained that multiple-choice test domains are determined from very thorough reviews of existing curricula guides and textbooks.

Anrig (1992) perceived testing as going through some dramatic changes throughout the 1990s. The changes will take place in three areas: more performance-based assessment, assessment will become more closely linked to instruction, and assessment will become more closely tied to technology. For example, the 1992 National Assessment of Educational Progress (NAEP) assessment will include performance-based assessment for the reading skills test (40%), the writing assessment (100%), and the math test (60%), but will still include multiple-choice questions (Anrig, 1992).

According to Mehrens (1991), performance-based assessments are not new types of assessments, although discussing them in the context of being the "latest solution to our educational problems is a new phenomena" (p. 3). Reasons cited by Mehrens for the support of performance-based assessment include

> (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

ERIC Full Text Provided by ERIC

> multiple-choice tests; and finally, (5) claims that there are deleterious instructional/learning effects of teaching to multiple-choice test formats. (p.3)

The search for alternatives to standardized testing is not new. When the National Education Association (NEA) and other education organizations criticized standardized testing in the 1970s, there was widespread interest in alternative assessments (Haney & Madaus, 1989). A resolution passed by the NEA encouraged "the elimination of group standardized intelligence, aptitude, and achievement tests" (Quinto & McKenna, 1977, p. 7). At that time, suggested alternatives included teacher-made tests, student work samples, teachers' professional judgement, contracts with students, interviews, and criterion-referenced tests (Haney & Madaus, 1989).

Performance assessments have been used in business and management for many years. Several categories of assessment techniques were discussed by Priestley (1982) including actual performance assessments, simulations, observational assessments, and oral assessments. According to Priestley, although types of assessments can vary significantly, all have three basic elements: a task to be performed, the conditions governing the performance of that task, and the method of scoring the result. The variations on these three basic elements are what make assessment techniques different from one another. Priestley also discussed factors that need to be weighed before choosing an assessment technique:

1) Clarify whether the a.sessment involves a product or a process to be measured. A product is a tangible object resulting from an examinee's performance and a process involves procedures or methods used to reach a particular point. A product may or may not result when a process is assessed.

2) Clarify how realistic the assessment should be. Federal guidelines state that any instrument used for selection of licensing must be a representative measure of the domain being measured, and must measure the skills in conditions approximating the actual job setting (U.S. Equal Employment Opportunity Commission, 1978).

3) Clarify how precisely to measure the assessment. This is a function of the skill being measured and the techniques used to measure the skill. Skills measured by an assessment technique can be categorized into four domains: cognitive, affective, psychomotor, and perceptual. The cognitive domain refers to the intellectual or mental skills that can be measured objectively (e.g., analyzing, identifying, calculating, and comparing). The affective domain refers to attitudes and behaviors that cannot be measured objectively. The psychomotor domain refers to skills that involve physical movement or manipulation, and the perceptual domain refers to the use of the body's sense organs.

4) Clarify the use of the information obtained from the assessment. Determine what type of information is desired and whether it will be formative or summative evaluation.

Following are descriptions of some of the various types of assessments discussed by Priestley, including advantages and disadvantages. Many of these assessment techniques were developed for use in business and management and are being researched for use in education.

Actual Performance Assessments

Actual performance assessments are administered in actual work or classroom settings and are used to evaluate a product, a process, or both. These types of assessments are appropriate for skills in any of the four domains.

Work-Sample Tests

Work-sample tests involve assigning a task to an examinee and evaluating the result, and can involve a product, a process, or both. If the assessment is of a process, it is generally conducted by judges using checklists or rating scales. If the assessment is of a product, checklists, ratings, or quantitative analyses can be used. Job selection and training are the areas where work-sample tests are most often used. Advantages of work-sample tests are that they assess actual performance in realistic settings; they can measure skills that are not readily measurable by other methods; direct observations of performance are provided; and specific, constructive feedback is provided by an observer. Disadvantages include the requirement of a one-to-one administration, cost and time limitations, observations may be subjective and not very reliable, a possible lack of standardized testing conditions exists, and the complete domain of required skills cannot be measured.

Identification Tests

This type of assessment does not generally require actual performance. It measures an examinee's ability to identify something in a job context. It is usually the first step in an actual performance and involves the identification of actual objects in a realistic setting. In some cases, the examinee is required to explain uses or characteristics of an object and how to repair, replace, or improve it. This type of assessment is most useful as a screening test, but can be used to measure skills required for completion of a training program. Many times this type of assessment is used in combination with work-sample tests or simulations.

Advantages of identification tests are that they can measure perceptual competencies, may be administered to small groups, involve real objects or substances, provide efficient measures of entry-level familiarity, and provide opportunities to assess diagnostic skills. Disadvantages include that the identification tests usually measure simple skills or basic knowledge, they don't measure actual performance, and they may require expensive objects.

Supervisor Ratings

Supervisor ratings are unobtrusive observations that occur under normal circumstances while an individual is working and can be used to assess a process and/or product. Supervisor ratings are formalized ratings, recorded in relation to predetermined criteria. Checklists, rating scales, or anecdotal records are used to provide guidelines for the observation and to record results. Supervisor ratings generally are used as on-going checks, but also may be used to assess trainees such as student teachers.

Advantages of supervisor ratings include the direct measurement of skills and behaviors, recommendations or personal comments may be provided, less anxiety is produced than with other types of observations, and feedback can be provided. Disadvantages include the possibility of bias, criteria may be difficult to establish, unreliable or inconsistent scoring may occur, there may be difficulty in obtaining standardized ratings, and supervisor ratings that are collected in an unsystematic manner have not been upheld in court as valid and reliable.

Peer Ratings

Peer ratings are evaluations made by colleagues or peers and criteria for peer ratings are nearly identical to those used in supervisor ratings. This type of evaluation is most helpful in training programs. Advantages are generally the same as those for supervisor ratings, except evaluations from peers may be less threatening than those from supervisors. In addition to the disadvantages described for supervisor ratings, peers may not know what skills and behaviors are essential on the job, reliability and validity have been found to be low for peer ratings, peers are untrained in evaluation, and the peers are being evaluated by the same standards as the examinee.

Self-Assessment

Self-assessments use the examinee's own personal ratings. Informal methods of self-assessment include discussions with a supervisor or teacher; formal techniques include self-assessment tests, structured interviews, checklists, and rating scales. The main purpose of self-assessment is to aid the examinee in learning. Self-assessments allow examinees to rate their own performance and determine strengths and weaknesses and are not generally used in licensing programs. Advantages include assistance in helping examinees identify their own strengths and weaknesses; assistance in training examinees in uses of assessment criteria and objective judgement; and aiding examinees in perceiving the value of defined guidelines, quality-control criteria, and step-by-step procedures. Disadvantages include that examinees are inexperienced in conducting evaluations, and inflated self-assessments may result. Self-assessments are not appropriate for use in assessment programs having legal consequences such as licensing because of their lack of objectivity and reliability.

Simulations

Simulations are used in situations in which actual performance tests are impractical due to danger, cost, consequences of mistakes, or the inability to arrange for actual performance situations. They are used more often in training than evaluation, but in either situation are designed to be as realistic as possible. Simulations offer the ability to predict how well an examinee will perform in a real situation on the basis of the simulated performance. The examiner is able to control most of the variables in a simulated situation and can standardize the assessment across administrations.

Simulated Performance

Simulated performance assessments consist of assessment techniques requiring physical performance in a simulated setting. Generally, simulated performance assessments focus on practical and technical procedures, behavior, and management skills.

Advantages of simulated performance assessments are that they provide realistic, direct assessment of on-the-job skills and behaviors; they allow for prediction of responses in a real situation; and they can be standardized. Disadvantages are the time and money required and the complex and lengthy scoring procedure.

Simulated Identification Test

A simulated identification test consists of a controlled situation in which the examinee identifies parts or problems, or manipulates a model. This type of assessment is most useful in situations in which an examinee's mistakes could cause serious consequences (e.g., surgery) or when the reproduction of a malfunction would be destructive (e.g., no oil in an automobile).

Advantages of simulated identification tests include that they are more direct assessments than paper-and-pencil identification tests, they allow for control of the situation, and the consequences of making mistakes are less than in a real situation. Disadvantages include that they are less direct assessments than actual performance assessments, they can be expensive, it may be difficult to obtain materials, and they measure cognitive understanding but not actual ability.

Written Simulation

A written simulation is a paper-and-pencil exercise simulating a decision-making process. The examinee is presented with a situation (written, filmed, recorded, role-played, spoken, or graphically presented) and is required to make inquiries, make decisions based on results of the inquiries, and reach a solution. The examinee solves the problem or faces unacceptable consequences resulting from his or her decision. The written simulation can be constructed to have several acceptable solutions, allowing the examinee to pursue several reasonable approaches.

Advantages of using written simulations are that they provide relatively realistic settings for decision-making or problem-solving skills, feedback is immediately available, and there is freedom to choose an approach. Disadvantages include the time and expense involved in developing complex problems; written simulations are better suited to training than testing; and if they are used for testing, examinees should have prior experience in taking this type of test.

Management Exercises

Management exercises include role-playing, simulatec terviews, fact-finding exercises, and case studies. Role-playing requires the examinee to assume a particular role, and presented with a problem situation, the examinee interacts with one or more individuals. The situation provides a realistic but controlled setting in which a sample of the examinee's behavior, interpersonal and communication skills, and problem-solving skills can be observed. The advantage to role-playing is that the examinee's social interactive skills (otherwise not easily measurable) can be observed in a realistic setting. The disadvantages include the time and number of personnel required for development, administration, and scoring; and the

ERIC Full Text Provided by ERIC

subjectivity intrinsic to any rating or scoring system used to assess complex performances. Because of the disadvantages, role-playing is generally better suited to training than assessment.

Simulated interviews are specialized uses of role-playing specifically to measure interview skills. The most common use of simulated interviews is to assess examinees whose jobs require interviewing on a regular basis. The advantages and disadvantages are the same as those of role-playing.

Fact-finding exercises are most often used in business and corporate settings. A specific problem that exists within a company is presented with a minimum amount of background information about the problem and company. The examinee must ask questions and examine documents to assemble and analyze information, and recommend solutions. The disadvantages to these exercises include their use in small-group settings, which allows the domination of one person and/or the benefit of others' answers, and little room for creativity or use of personal problem-solving styles unless a number of optional sources of information are available.

When using case studies, one or more examinees are given informational case descriptions to be analyzed and diagnoses made or solutions proposed. Scoring can be based on final solutions or on the basis of how the solution was derived. Case studies have not been widely used in education and Merseth (1991) discussed some of the reasons for this. Additional information regarding the use of case studies in education can be found in Barnett (1991), Kagan and Tippins (1991), Shulman (1991), and Wineburg (1991).

Observational Assessments

Observational assessment techniques are methods for rating or scoring performance and generally are used in combination with other assessment techniques. The observational assessment techniques can be used to observe a process and/or a product, and all techniques share some characteristics. These characteristics include definitions of critical components of a product or process, use of a recording form, observation by an assessor of the behavior or product to be rated, and indication on the recording form by the observer of whether the critical components were observed. Observational assessment may be obtrusive (i.e., a specific testing situation is devised and specific behaviors performed) or unobtrusive (i.e., behavior occurring normally without an artificial testing situation). Observation instruments have three major functions: to structure and govern the observation, to serve as a recording form, and to explain to the examinee the basis for evaluation. The instruments generally have directions, criteria for observation and recording, and instructions for scoring.

Issues in observational assessment are based on the subjective nature of judgmental observations. These issues include unfair judgements based on the time periods that observations occur (e.g. ten minutes of an eight-hour day, misbehavior of a student), scoring reliability, and the use of observations over long periods of time. Time sampling is one method available to aid in solving the problem of unfair judgements based on the time periods that observations occur. Time sampling requires that observations be planned to occur at specific times in advance of the observations. The selection of the times should allow for a variety of activities and should be as regular as scheduling allows. In order to ensure scoring reliability, observers should be carefully and thoroughly trained in the subject area being observed and in observational assessment. Even when carefully trained observers are used, though, personal bias, the halo effect (i.e., rating the examinee the same on all dimensions), and logical errors (i.e., if an examinee is low in achievement, that person must be low in intelligence as well) must be guarded against.

Checklists

Checklists are one technique for observational assessments and can be quantitative or qualitative. Both types of checklists list the dimensions to be observed and use two opposite choices for each dimension (i.e., yes/no, correct/incorrect). Quantitative checklists assess only the presence of a behavior or attribute, while qualitative checklists assess whether the behavior or attribute is of desired quality. Because qualitative checklists require judgement, qualitative terminology must be clearly defined to minimize rater bias and ensure consistent assessment.

Quantitative checklists are useful when it is sufficient to know whether a behavior or attribute is present (e.g., the teacher asked questions), to observe a procedure with definite steps, or to evaluate a product with definite characteristics. Qualitative checklists are useful when the presence of a behavior or attribute does not provide adequate information (e.g., questions asked by the teacher were relevant to the subject matter). The possibility of bias is greater with the qualitative checklists, but both checklists have been used with success in education. The checklists are relatively easy to use but can be expensive and time-consuming to develop, administer, and score.

One of the advantages of both types of checklists is their use in guiding and scoring observations when the behaviors or attributes to be observed are known in advance. In situations where the behaviors or attributes are unpredictable, checklists are ineffective. Quantitative checklists have the advantage of rater subjectivity being less of a problem than with other observational techniques because of the simplicity of checking yes or no, but the disadvantage is that they do not provide measures of quality, speed, and accuracy. The advantage of qualitative checklists is that they do provide a measure of quality, but the dimensions of quality must be clearly defined, and at least two observers should score the product or process independently.

Rating Scales

Rating scales are another observational technique. They are similar to checklists, but rather than having two opposite choices, the degree to which a behavior or attribute is present is indicated by a point along a continuum. The four basic types of rating scales are numerical, graphic, descriptive-graphic, and ranking scales. Rating scales can be used for the same types of assessments as checklists, but usually provide more information than checklists.

Rating scales generally are used to measure affective traits such as attitudes and opinions, because attitudes and opinions are not considered right or wrong but are measured on a continuum.

Numerical rating scales use numbers, and to be used reliably, the numbers must have the same meaning for every behavior or attribute listed. Graphic rating scales use words (e.g., poor, fair, good) instead of numbers. Ratings can be different for each behavior or attribute. Graphic scales are more widely used than numerical scales due to flexibility, but reliability may be a problem because the distinction between descriptions (e.g., fair, good) is subjective and may vary between observers. In an attempt to reduce subjectivity, descriptive-graphic rating scales provide detailed descriptions for each point on the scale. This type of rating scale is useful when standard criteria exist. Ranking scales are used to rate a product or process in comparison with similar products or processes. Normative-referenced ranking consists of comparisons within a group. Criterion-referenced ranking is used to compare a product to a predetermined minimum standard.

Anecdotal Records

Anecdotal records are an objective method of recording events that might be forgotten or incorrectly remembered, especially atypical events. They are factual, written descriptions of an event. Anecdotal records are used as descriptions of events and are not scored or evaluated. Examination of anecdotal records may reveal patterns of behavior that might be important for making diagnoses or recommendations.

Oral Assessment

Oral assessment techniques require some form of oral response and can be used for any non-physical performance test. The three major techniques reviewed by Priestley are oral examination, interview, and prepared presentation. Priestley maintained that oral assessments are useful in situations where examinees need to explain their reasoning or defend their ideas, when written communication is impractical or possibly subject to bias, and when oral communication skills are essential. Generally, a checklist or rating scale is used to record and score an examinee's performance.

Oral Examination

Oral examinations generally are designed to measure either a generalized domain of knowledge or a specialized domain of knowledge, often related to a product (e.g., dissertation defense). Oral exams have been used as part of the requirements for licensing; to assess specific mastery of a limited field of knowledge; to measure oral communication skills; and to assess comprehensive and in-depth knowledge, problem-solving skills, application skills, and interpretive skills.

Advantages of oral examinations are that complex skills difficult to assess with paper-and-pencil tests can be assessed (e.g., oral communication skills), an examinee can explain or defend a response, continued questioning can be conducted if an examinee misinterprets a question, and face validity can be relatively high. The disadvantages include that the examination is given on an individual basis and can be time-consuming, and there may be subjective scoring.

Interview

When interviews are used for assessment, the purpose of the interview is to evaluate the examinee in relation to predetermined standards and characteristics. The interview should be structured and may consist of factual questions, affective questions, or a combination of both. The questions should be constructed to elicit potentially verifiable or factual responses that are specific and sample actual behavior.

The advantages of interviews are that they are often cited as the only way to evaluate certain characteristics regarded as essential to effective job performance, the interviewee can ask the examiner questions, and there is flexibility in an informal interview. Disadvantages are that a number of sources of error exists, consistency of examiners tends to be low, prediction of future success tends to be low, and time and cost are generally high.

Prepared Presentation

In a prepared presentation an examinee or group of examinees provides an organized display of information. The presentation may be entirely oral or use audio and/or visual materials. When a prepared presentation is used as an assessment technique, its purpose may be to assess a product; knowledge; or particular skills such as the ability to communicate orally, to select and organize information, or to persuade. Prepared presentations are used in many situations including the demonstration of teaching skills by preservice teachers.

Advantages are that prepared presentations are the most appropriate method for assessing oral skills that will be necessary in an occupation and a larger number of examiners can be used, which increases reliability of ratings. Disadvantages are subjectivity and that prepared presentations are more costly and time-consuming than group assessments.

Use of Performance Assessments in Education

The Teacher Assessment Project (TAP) at Stanford University spent from 1986 to 1990 exploring alternative methods of performance-based assessment that have the potential for use in teacher licensing and certification. The performance assessment exercises developed by the TAP, sponsored by the Carnegie Corporation of New York, were discussed by Haertel (1990). "The TAP prototype exercises represent a fundamentally new kind of teacher examination, based on structured observations of teachers' performance in situations designed to elicit the same kinds of knowledge and skills used in teaching, lesson planning, textbook selection, or related activities" (p. 15). These exercises are referred to as structured performance assessments and would fall under the category of techniques Priestley (1982) calls simulations. The structured performance assessments are conducted in assessment centers and simulate situations that occur in classrooms.

The prototype exercises have been developed by TAP to assist the National Board for Professional Teaching Standards in developing its teacher-certification tests. Haertel (1990) made a distinction between certification and licensing. He stated that certification "generally refers to a form of recognition controlled by organizations representing practicing professionals, for example, the National Board of Medical Examiners. Certification attests to some level of mature and expert practice" (pp. 17-18). Licensing refers to the issuance of a license by a government agency to practice a profession. Although the structured performance assessments were created for use in teacher certification, they may be useful in the assessment of teacher education students and teacher licensure.

The assessments examined included assessment center exercises, portfolio documentation, and a combination of on-site portfolio documentation with assessment center exercises. The assessment center exercises "required teachers to demonstrate and explain their knowledge and skill in hypothetical situations analogous to actual practice" (Murnane, 1991, p. 17). Examples included analyzing textbooks, evaluating student work, and analyzing video tapes of others' teaching. The exercises consisted of simulated teaching activities using semi-structured interviews (Vavrus & Collins, 1991). Semi-structured interviews are designed to give some structure to an interview without stifling the interviewer's responses. The interview begins with an initial set of designated questions and as the interview progresses, additional questions (called probes) are formulated based on the interviewee's responses. Advantages of this format include contact between interviewer and interviewee which is believed to allow for fuller answers, than if the interview was written and the interviewer can probe for more information (Tyson, 1991).

According to Grover, Zaslavsky, and Leinhardt (1990), the use of semi-structured interviews is due to the belief that semi-structured interviews may be able to capture the complexity of teaching and "that the nature of the assessment upon which licensure or certification is based will ultimately produce changes in the classroom by influencing the nature of teacher preparation programs" (p. 3). The most common interview approach consists of an interviewer asking predetermined questions about a series of tasks designed to simulate a range of significant teaching activities. The interviews are especially helpful in examining teachers' knowledge of their subject area, teaching strategies, student assessment techniques, classroom management, and motivational techniques.

Pecheone and Carey (1990) discussed the Connecticut State Department of Education's work in developing semi-structured interviews to address beginning math teachers' abilities. The interviews use open-ended questions to stimulate responses instead of controlling the responses. The four mathematics exercises used in the interviews consist of: structuring a unit, structuring a lesson, alternative approaches, and evaluating student performance. More complete descriptions of the exercises are given in Pecheone and Carey (1990). Grover, Zaslavsky, and Leinhardt (1990) discussed the development of a scoring system for the Connecticut semi-structured interviews.

Researchers at the RAND Corporation are designing simulation problems for use in assessing skills in instruction, planning, evaluation and assessment, and classroom management. The simulations are written simulations being developed for assessment of teaching skills for the state of California. An example is to ask a candidate for licensure in English to use a set of resource materials and plan a sequence of lessons to meet specific curricular goals for a class of students with certain backgrounds and skills. Candidates may then be asked to evaluate essays written by a different class of students (Murnane, 1991).

The greatest amount of research and interest appears to be directed at portfolios. Portfolios have been used in some areas of education such as art, but are now being examined for use in the areas of teacher certification and teacher licensure. The compiling of portfolios was one of the areas of focus of the Teacher Assessment Project (TAP) at Stanford University. The subject area portfolios were examined for elementary literacy and high school biology. The portfolios consisted of videotapes, lesson plans, samples of student work, and reflective commentaries. The activities documented in the portfolios then were used to help develop simulated exercises for the assessment center (Wolf, 1991).

ERIC
Full Text Provided by ERIC

The Stanford Teacher Educational Program (STEP) also has been examining the use of portfolios. STEP instituted a teacher portfolio program in 1990-91. Lichenstein, Rubin, and Grant (1992) in their discussion of this program defined a portfolio as the "physical document that contained teachers' inquiries' (p. 6). The portfolios consisted of two pieces of practical research called entries. The entries contained three parts: rationale, artifacts, and reflection. The rationale explained why the inquiry was important to the individual developing the portfolio. The artifacts were materials such as videotapes or recordings of classroom instruction, lesson plans, student work, journal entries, or other physical evidence related to teachers' practice. The reflection was a written section about the artifacts collected and pulled together the inquiry, the artifacts, and the teachers' actual practice.

The most common form of actual performance assessments used in teacher training programs is the supervisor rating. This type of rating generally is used in student teaching programs; supervisors observe student teachers in an actual classrooms and rate their performance. Peer ratings and self-assessments also have been utilized. Work-sample tests and identification tests are less commonly used in evaluating teachers or teacher candidates; little research was found regarding these techniques.

Classroom observations are a combination of actual performance assessment and observational assessment. Behavior checklists, category systems, narrative records, summaries, and rating systems are some of the instruments currently in use for classroom observations. Stodolsky (1990) discussed open and closed observation methods for classroom observations. Closed systems focus on specific types of behavior and use a finite number of pre-established categories of behavior. Open systems may use narratives, ad lib notes, films, videotapes, and specimen records as developed by Barber and Wright (1955) and Wright (1967). Stodolsky (1990) also discussed reliability, validity, sampling, and observer training issues in classroom observation.

Classroom observations also are being utilized by Educational Testing Service (ETS). ETS is redesigning the NTE tests and is planning to rely primarily on classroom observations to measure performance. The four content areas to be evaluated are: planning for instruction, implementation of instruction, classroom management, and evaluating students' progress. ETS will not administer the assessments, but will provide technical assistance in training observers, developing scoring strategies, and setting minimum performance standards (Murnane, 1991). "The central assumption underlying the ETS strategy is that trained evaluators' observations of a teacher working with students provide the most valid and reliable measures of assessing teaching skills" (Murnane, 1991, p. 140).

Research Regarding Multiple-choice Tests and Alternative Assessments

The National Teacher Examination (NTE), a standardized multiple-choice test measuring three areas of teacher preparation (general education, professional education, and specialty areas), has been used in several southern states for testing prospective teachers (Southern Regional Education Board, 1982). Several studies have looked at the relationship between scores on the NTE and scores for on-the-job performance obtained from ratings by supervisors or principals of student teachers or classroom teachers. The relationship between the Common subtest scores of the NTE and teaching style, measured by the presence or absence of behaviors of the teacher, was examined by Medley and Hill (1970) as cited by Southern Regional Education Board (1982). Correlations with a median of .25 were found. A significant correlation of .43 was found between NTE elementary education test scores and student teaching ratings by university supervisors (Piper & Sullivan, 1981).

A study by the Southern Regional Education Board (1982) examined the relationship between scores on the Georgia Teacher Certification Test (TCT) and ratings on the Georgia Teacher Performance Assessment Instrument (TPAI), and between scores on the NTE and TPAI ratings. The TCT is a teacher certification test developed by the Georgia Department of Education designed to assess teaching field knowledge. The TPAI is a used to assess on-the-job performance in 14 areas including teaching plans and materials, classroom procedures, and interpersonal skills. The correlations between TCT scores and TPAI ratings ranged from -.27 to .45. Correlations between the NTE scores and TPAI ratings ranged from -.12 to .52. The authors of the study stated that the findings support the view of ETS and the Georgia Department of Education "that knowledge is only one part of the complex process called teaching" (Southern Regional Educational Board, 1°82, p. 22). The Southern Regional Education Board found mixed results in the research examining the relationship between NTE scores and ratings for on-the-job performance, but maintained that the results generally support the view of Educational Testing Service (1978) that the NTE does not predict classroom performance.

Murnane suggested a long-term solution to multiple-choice tests as designing tests that require constructed responses, which he maintained will provide more valid measures of subject-matter knowledge than multiple-choice tests. He recommended diversity among training programs because no single strategy works best for all prospective teachers. Sorne of the alternative assessments discussed by Murnane included task simulations, classroom observations, and focusing on teacher/student interactions.

Reliability and validity have yet to be resolved regarding performance-based assessment. Mehrens (1991) maintained that several threats to the validity ana reliability of performance assessments exist: limited sampling, whether correct domains are being assessed, whether the domains are well enough defined, generalizability, scoring, bias, lack of internal consistency, and subjectivity. Trevisan (1991) explored some of the issues surrounding the reliability of performance-based assessment and has recommended caution before giving wholesale acceptance to this type of assessment. He stated that there is little reliability data available on performance-based assessments and research needs to be focused on the psychometric issues of performance-based assessments.

Conclusions

The major problems of using performance-based assessment are time and cost. These have long been two advantages of multiple-choice tests. Estimates have been that performance-based assessments are four to 10 times more expensive than machine-scoreable, multiple-choice tests (Anrig, 1992). It is more expensive to have to train and hire teachers and professors to read examinee responses than to scan a score on a multiple-choice test. The time required to administer a multiple-choice test can be predicted and fairly short, but when using performance-based assessment, examinees must be given enough time to perform. Both time and cost need to be carefully considered.

Some concern has been expressed that although many educators and researchers do not think the domain covered by multiple-choice tests is broad enough, performance tests may access even narrower domains, although possibly in more depth. The question of which domains are being covered by which types of assessments is yet to be answered. The results of research have been mixed. Some research suggests that multiple-choice tests and performance assessments cover the same domains while other research suggests they cover different domains (Ackerman & Smith, 1988; Bennet, Rock, & Wang. 1991; Birenbaum & Tatsuoka, 1987; Farr, Pritchard, & Smitten,1990; Martinez, 1990; Traub & Fisher, 1977; Ward, 1982; Ward, Frederiksen, & Carlson, 1980).

Performance-based assessment is being seen as one alternative to multiple-choice tests. There are advantages and disadvantages to both types of assessments and one should not be used to the exclusion of the other. They both have their place depending upon constraints such as purpose of the testing, time, money, and etc. "Although I am a strong supporter of performance-based assessment, I believe we should also resist the temptation to think it will solve all the problems of testing - particularly the problem of differences in test results by race, sex, and ethnicity" (Anrig, 1992, p. 4). Even advocates of performance assessment 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. . ." (Collis & Romberg, 1991, p. 102).

Many educators and researchers see the answer as somewhere in between, the use of alternative assessments in conjunction with standardized multiple-choice tests. There are certain domains of skills and knowledge which lend themselves to performance-based assessment, while others would be much better tested using multiple-choice tests.

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.

Anrig, G. R. (1992). Trends in standardized testing: Nationwide assessment, not a national test. Spectrum, 10(1), 3-9.

Barker, R. G., & Wright, H. F. (1955). Midwest and its children: The psychological ecology of an American town. New York: Harper & Row, Publishers.

Barnett, C. (1991). Building a case-based curriculum to enhance the pedagogical content knowledge of mathematics teachers. Journal of Teacher Education, 42(4), 263-274.

Bennet, 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.

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. (1978, April 14). The national teacher examinations, notes on their reliability & standard error of measurement. Princeton, NJ: ETS.

Farr, R., Pritchard, R., & 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.

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

Grover, B. W., Zaslavsky, O., & Leinhardt, G. (1990). Scoring a semi-structured interview for assessment of beginning secondary mathematics teachers. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Haertel, E. H. (1990). Performance tests, simulations, and other methods. In J. Millman & L. Darling-Hammond (Eds.), The new handbook of teacher evaluation. Newbury Park, CA: Corwin Press, Inc.

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

Kagan, D. M., & Tippins, D. J. (1991). How teachers' classroom cases express their pedagogical beliefs. Journal of Teacher Education, 42(4), 281-291.

Lichenstein, G., Rubin, T. A., & Grant, G. E. (1992, April). Teacher portfolios and professional development. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Medley, D. M., & Hill, R. A. (1970, March). Cognitive factors in teaching style. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis.

Mehrens, W. A. (1991, April). Using performance assessment for accountability purposes: Some problems. Paper abridged from a paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.. (ERIC Document Reproduction Service No. ED 333 008)

Merseth, K. K. (1991). The early history of case-based instruction: Insights for teacher education today. Journal of Teacher Education, 42(4), 243-249.

Murnane, R. J. (1991). The case for performance-based licensing. Phi Delta Kappan, 73(2), 137-142.

Neill, D. M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. Phi Delta Kappan, 70(9), 688-697.

Pecheone, R. L., & Carey, N. B. (1990). The validity of performance assessments for teacher licensure: Connecticut's ongoing research. Journal of Personnel Evaluation in Education, 3, 115-142.

Piper, M. K., & Sullivan, P. S. (1981, January). The national teacher Examination: Can it predict classroom performance? Phi Delta Kappan, 401.

Priestley, M. (1982). Performance assessment in education & training: Alternative techniques. Englewood Cliffs, NJ: Educational Technology Publications, Inc.

Quinto, F., & McKenna, B. (1977). Alternatives to standardized testing. Washington, D.C.: National Education Association.

Shulman, J. H. (1991). Revealing the mysteries of teacher-written cases: Opening the black box. Journal of Teacher Education, 42(4), 250-262.

Southern Regional Education Board (1982). Teacher testing and assessment: An examination of the National Teacher Examinations (NTE), the Georgia Teacher Certification Test (TCT), the Georgia Teacher Performance Assessment Instrument (TPAI) for a selected population. Atlanta, GA: Author. (ERIC Document Reproduction Service No. ED 229 441)

Stodolsky, S. S. (1990). Classroom observation. In J. Millman and L. Darling-Hammond (Eds.), The new handbook of teacher evaluation. Newbury Park, CA: Corwin Press, Inc.

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

ERIC Full Text Provided by ERIC

Trevisan, M. S. (1991, April). Reliability of performance assessments: Let's make sure we account for the errors. Paper presented at the annual meeting of the National Council on Measurement in Education and the National Association of Test Directors, Chicago, IL.

Tyson, P. (1991). Talking about lesson planning: The use of semi-structured interviews in teacher education. Teacher Education Quarterly, 18(3), 87-96.

U.S. Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38309.

Vavrus, L. G., & Collins, A. (1991). Portfolio documentation and assessment center exercises: A marriage made for teacher assessment. Teacher Education Quarterly, 18(3), 13-30.

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.

Wineburg, S. S. (1991). A case of pedagogical failure--my own. Journal of Teacher Education, 42(4), 273-280.

Wright, H. F. (1967). Recording and analyzing child behavior. New York: Harper & Row, Publishers.

Wolf, K. (1991). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. Phi Delta Kappan, 73(2), 129-136.