DOCUMENT RESUME

ED 276 721 SP 028 400

AUTHOR Murray, Stephen L.
TITLE Considering Policy Options for Testing Teachers.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 86
CONTRACT 400-86-0006
NOTE 40p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; *Educational Policy; Elementary Secondary Education; *Policy Formation; State Standards; *Teacher Effectiveness; Teacher Improvement; Test Format; *Testing Programs
IDENTIFIERS *Teacher Competency Testing

ABSTRACT

This paper focuses specifically on paper and pencil testing as a tool contributing to a quality teaching force. Information is presented for those who have responsibility or interest in state level policies for using such tests to promote educational quality. Using institutional stages of a teacher's career as an organizing scheme for test use, the paper provides a framework for examining a range of policy options, discusses requirements for tests to support different decisions, and identifies issues important to implementing these options. Focus is upon paper and pencil testing as a policy tool because of the tremendous amount of interest it has received in the past few years. It is advocated that testing be considered as only one of many means to control teacher quality. As a backdrop for examining teacher testing policy, three fundamental questions are posed: (1) Why test teachers? (2) What decisions will teacher testing support? and (3) What are the requirements for tests? A framework addressing these dimensions of teacher testing policy is offered to help policy makers analyze the appropriateness of specific policy options. It offers information for: (1) examining whether a testing option under review will be consistent with the underlying purposes of the policies; (2) identifying decisions supported by the testing; and (3) revealing technical and legal requirements of the tests to be used. Guidelines are offered for avoiding the primary pitfalls of teacher testing programs. (JD)

CONSIDERING POLICY OPTIONS FOR TESTING TEACHERS

Stephen L. Murray
Northwest Regional Educational Laboratory
300 S.W. Sixth Avenue
Portland, Oregon 97204

November, 1986

This publication is based on work sponsored wholly or in part by the Office of Educational Research and Improvement, U. S. Department of Education under contract number 400-86-0006. The content of this publication does not necessarily reflect the views of the department or any other agency of the U. S. Government.

CONSIDERING POLICY OPTIONS FOR TESTING TEACHERS

Stephen L. Murray

Northwest Regional Educational Laboratory

INTRODUCTION

How should we improve and maintain quality in the teaching force? We all have an interest in this question, whether as parents, concerned citizens, educators, researchers or legislators. A Gallup poll conducted in 1984 revealed that 89 percent of the general public favored using state controlled tests to certify prospective teachers for those subjects in which they planned to give instruction (Gallup, 1984). The 1986 Gallup poll of attitudes toward public education revealed that 85 percent of those polled endorsed periodically requiring experienced teachers to pass a statewide basic competency test in their subject area or areas (Gallup, 1986).

Over the past several years, coinciding with the general movement to improve educational quality, we have witnessed a significant increase in testing prospective and, in some cases, already certified teachers. A recent report by the Educational Testing Service (ETS) shows that all but five states mandate, or have plans to mandate the testing of prospective teachers (Anrig, 1986). As of the summer of 1986, almost one-half of the states were actively considering new developments in their policies governing the use of tests to certify teachers. Two national teacher organizations, the National Education Association (NEA) and the American Federation of Teachers (AFT), have endorsed some applications of teacher testing. Clearly, the interest and support for teacher testing is widespread.

In the late Spring of 1986, two national reports prompted further public interest in improving the quality of the teaching force as a key to achieving long term educational reform. The first of these two landmark reports, Tomorrow's Teachers: A--Report of The Holmes Group (1986), was developed by a consortium of education deans. The second report, A Nation Prepared: Teachers in the 21st Century, was prepared by the Carnegie Forum on Education and the Economy's Task Force on Teaching as a Profession (1986). Each report calls for major reforms of policies governing the quality of the teaching force. Among other measures, the Carnegie Task Force report recommends the creation of a National Board for Professional Teaching Standards to oversee a rigorous national teacher certification system. The Carnegie Foundation has already funded efforts to establish such a Board and begin the technical planning for an approach to certification testing.

Any agency examining methods for controlling the quality of the teaching force needs to keep pace with developments proposed to improve the quality of the certified teacher pool. This paper focuses specifically on paper and pencil testing as a tool contributing to a quality teaching force. It presents information for those who have responsibility or interest in state level policies for using such tests to promote a quality teaching force. Using institutional stages of a teacher's career as an organizing scheme for test use, the paper provides a framework for examining a range of policy options, discusses requirements for tests to support different decisions, and identifies issues important to implementing these options. We focus on paper and pencil testing as a policy tool because of the tremendous amount of interest it has received in the past few years. Like many others, we advocate that testing be considered as only one of many means to control teacher quality.

As a backdrop for examining teacher testing policy options, we pose three fundamental questions:

1. Why test teachers?
2. What decisions will teacher testing support?
3. What are the requirements for tests?

A framework addressing these dimensions of teacher testing policy will help to analyze the appropriateness of specific policy options. It will allow one to: (1) examine whether a testing option under review will be consistent with the underlying purposes of the policies, (2) identify decisions supported by the testing and (3) reveal technical and legal requirements of the tests to be used.

An issue that we do not address in this paper is how the supply and demand of teacher candidates and certified teachers will affect the long range attainment of policy goals for teacher testing programs. Teacher supply and demand and incentives (e.g., pay) for teachers will interact with new testing programs and influence the success of the policy. Raising standards without increasing pay or improving working conditions would most likely reduce the supply of teachers.

POLICY DIMENSIONS OF TEACHER TESTING

Why Test Teachers?

A review of the teacher testing area suggests four underlying reasons to test teachers:

1. Limiting the number of incompetent teachers
2. Encouraging teacher professionalism
3. Promoting public confidence in teachers as a group
4. Promoting excellence

First, testing is a means to limit the number of incompetent teachers. Vorwerk and Gorth (1986), for instance, state that the primary outcome of every teacher certification system is to protect the public from incompetent teachers. A teacher who lacks teaching skills or content knowledge is likely to do more harm than good and should not to be allowed to teach, a position to which nearly all would subscribe. If agreement on the skills and knowledge that are essential can be reached, and tests can validly and reliably identify those who lack these essential skills and knowledge, we have a method to implement the policy goal of limiting incompetence through testing. Teacher candidates who fail to demonstrate minimum competence will be prevented from entering the classroom as teachers. Achieving this goal is consistent with the purpose of licensing programs in general, and with teacher certification programs as a form of licensing (Shimberg, 1981).

Paper and pencil tests typically used for certification are appropriate for identifying only one form of incompetence--lack of knowledge. Tests of subject-matter knowledge or basic communication skills are not appropriate for assessing other areas in which teacher competence may be a concern. Bridges (1986), for example, reports that the leading cause for teacher dismissal, in over seventy years of research, is weakness in maintaining student discipline. Problems in maintaining rapport with other teachers and parents, and failure to produce intended classroom outcomes are other frequent causes for teacher dismissal. Clearly, tests of knowledge are not designed to predict the ability to maintain discipline and rapport, or to produce intended classroom outcomes. Other forms of assessment and evaluation are needed to validly measure these areas of competence.

A second reason for teacher testing is to encourage teacher professionalism. Shanker (1986) and Schulman (1986) have argued that assessment and testing systems should be modeled after professional certification systems such as nongovernmental medical specialty boards. Such testing would contribute to greater professional legitimacy for teachers. It would emphasize the upgrading of teaching rather than draw attention to those candidates who do not measure up to minimum knowledge expectations. The content of tests for teachers would reflect the expert knowledge required to match the complex job requirements. The tests would also contribute to a greater public valuing of teaching as a profession (Schulman, 1986).

The distinction between limiting incompetence and increasing professionalism parallels the distinction between licensing and certification.

Licensing is defined "as a process by which an agency of government grants permission (emphasis added) to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected." (U.S. Department of Health, Education, and Welfare, 1977, p. 4).

Certification, on the other hand, "is the process by which a governmental or nongovernmental agency grants recognition (emphasis added) to an individual who has met certain predetermined qualifications set by a credentialing agency...Unlike licensure, a certification law does not prohibit uncertified individuals from practicing their occupations." (Shimberg, 1981, p. 1138).

Where state laws prohibit teaching by one who is not certified in that state, teacher certification is serving the more restrictive licensing function even though it goes by the name of certification.

Proponents of teacher testing reform, such as Shanker (1986), advocate that teacher testing focus on higher standards with greater fidelity to the complex job of teaching. They argue for rigorous standards that go beyond what is minimally required to maintain the public welfare. They emphasize certification rather than licensure because they feel that teacher tests overly simplify what it takes to be a good teacher.

A third reason for implementing teacher testing policies is to promote public confidence in teachers as a group. Gallup polls taken in 1979, 1981 and 1984 have shown strong public support for testing prospective teachers in the subject areas they will teach as a condition for their employment as teachers (Gallup, 1984). The percentage favoring teacher testing has gone from 85 percent in the 1979 poll to 89 percent in the 1984 poll. Gaining or maintaining public confidence in teachers depends on a multiplicity of

approaches. Publishing summary testing results, so one argument goes, will show that some of those who aspire to teach are screened out. Although one cannot predict what an acceptable failure rate would be, a system that screened no one out would lack credibility to the public.

Yet another point of view, one consistent with the purpose of increasing professionalism, sees public confidence being more influenced by data reflecting higher levels of teaching competence. Such tests, along with other information, would be used to support rewards (e.g., merit pay, promotion and granting special status) based on demonstrated expertise.

A fourth reason for testing teachers is to promote excellence in education. Teacher testing programs that are part of more general educational reforms promote excellence indirectly by symbolizing that higher standards of performance are expected. Excellence prevails when the best teachers are hired, when superior teachers are recognized and when good teachers are encouraged to stay in the profession.

As one formulates or examines policy options for testing teachers, each alternative should be reviewed in light of these four underlying purposes. Not all policy options will serve all purposes equally well, and some options may be contrary to the more fundamental ends of some policy makers. For example, tests used to eliminate teachers who lack minimally necessary communication skills, computational skills and subject matter knowledge for a beginning teacher will do little to promote teaching professionalism. A test that focuses on pedagogical skills, however, may fail to identify some candidates lacking basic skills required to teach. One must be clear on the policy goals to be achieved.

Used in an institutional setting, tests are designed to contribute to decision making. The following section outlines five types of institutional decisions that may be supported by teacher testing.

WHAT DECISIONS DOES TEACHER TESTING SUPPORT?

Institutional Decisions

Broadly conceived, teacher tests can be used to support institutional or individual decisions (Cronbach and Gleser, 1965). Most teacher testing policies, however, are intended to support institutional decisions or recurrent choices made about individuals by an agent or agents acting on behalf of an institution. The individuals about whom choices are made include applicants for teacher training, those who are trained as teachers and are seeking certification, others who are seeking certification, and currently certified teachers. State agencies, colleges or universities involved in teacher training, and local school districts are the primary agencies involved in using test results to support decisions about these individuals.

Consequences of these decisions affect the general quality of the teaching corps and the opportunity of individuals to pursue careers of their choosing. Given these consequences, there should be little wonder why testing is, itself, the object of vigilance. This scrutiny and the debate around testing often highlight the common and conflicting interests and values of those with a vested interest in who is chosen to teach.

Common institutional decisions involving teacher testing include:

1. Admitting candidates to teacher preparation programs
2. Certifying or licensing teachers as sufficiently competent to teach
3. Selecting certified teachers for specific positions
4. Recertifying practicing teachers
5. Granting promotion, rewards or special status

Newer policies have called for expanding the institutional use of tests to include recertifying practicing teachers and promoting teachers in career ladder programs. In addition to these uses of data for decisions about individuals, certification test results for teachers are often used in the evaluation of teacher training programs. Each of these decision types, as well as the underlying policy rationale for teacher testing initiatives, places specific demands on the type of testing necessary.

Admission to Teacher Preparation Programs. The first institutional decision point in a teacher's career sequence is for a teacher training institution to decide whether to admit a college student into an undergraduate teacher preparation program. The admissions decision, which is typically made after the candidate has completed two years of college, is intended to select those who will successfully complete the preparation program and who will subsequently become certified to teach. In other words, to select those who show promise as a teacher. Although admissions decisions are made by the teacher training institution to which the candidate has applied, states may, in some cases, impose common standards to be used by all state approved teacher training institutions.

Usually, the number of teacher candidates admitted to an institution's preparation program depends upon the number of persons the training program can accommodate. The admissions decision, therefore, operates with a selection quota and is norm-referenced. Depending upon the quota, the size of the applicant pool, and the number of applicants who meet minimum standards for entry, the proportion of the candidates admitted will vary from year to year. A school may not fill its quota when an insufficient number of

applicants meets the minimum admissions requirements set by the school. The teacher training institution should take into account the effectiveness of its teacher training program in relation to the students it serves. The more effective program would, in theory, yield better prepared graduates with the same admission standards as a less effective program.

As with most other forms of decision making, decision makers should consider tests as but one of many sources of information used to make admissions decisions. Generally, the weight given to test-derived information is not specified and may be difficult to determine in practice. The conditions under which nontest data are allowed to compensate for poor test performance is a policy issue of some importance. Will even the poorest test performance be allowed to outweigh nontest data, or must candidates achieve at least a minimum score on the test to be admitted? A very low minimum acceptable score, one that a high percentage of applicants can be expected to pass, may actually give the test less weight than a decision allowing poor test performance to be compensated for with other information. In other words, using an absolute cutoff score, by itself, does not determine the importance (weight) of the test in making the decision.

Tests used to support admissions decisions typically measure basic literacy or academic skills and include such tests as:

1. The Pre-Professional Skills Test (PPST)
2. The California Basic Educational Skills Test (CBEST)
3. The Alabama English Language Proficiency Test (ELP)
4. The Connecticut Competency Examination for Prospective Teachers (CONNCEPT)
5. The California Achievement Test (CAT)
6. The Scholastic Aptitude Test (SAT)

Teacher training applicants generally are not expected to have pedagogical knowledge or a high level of subject matter expertise, because they are tested prior to teacher specific training. The PPST, CBEST, CONNCEPT, and ELP are designed specifically for teaching candidates, the CAT is a general measure of academic achievement, and the SAT is a general measure of academic aptitude often used for general college admissions decisions. The PPST, which measures basic proficiency in reading, writing, and mathematics, was developed by the Educational Testing Service (ETS) to test content similar to that of the National Teacher Examination (NTE) but appropriate for teacher training applicants not yet exposed to specific teacher training. The CBEST is similar to the PPST, having been developed to the specifications of the California State Department of Education by the ETS. The major difference between the PPST and the CBEST is in the writing section. The CBEST includes two written essays while the PPST has only one essay and an objective, multiple-choice section.

The ELP and the CONNCEPT are custom designed criterion-referenced tests. The ELP measures competencies needed for successful completion of course work in the teacher education program and for effective classroom teaching (Baker and Fennel, 1986). The CONNCEPT was designed with the same general goal in mind (Pecheone, Tomala, and Forgione, 1986).

When admissions decisions are based on a quota and the goal is to select the best of many applicants, the admissions test must reliably discriminate between applicants across a broader range of talent than will a typical certification test, which needs only to discriminate between those who possess the minimum required level of knowledge and those who do not.

On the other hand, an admissions test is often used as a preliminary hurdle (minimum cutoff) in which case the test needs only to identify candidates with minimally acceptable performance.

In considering the use of tests to support admissions decisions, the primary concerns will depend on the institutional purposes which the test will serve, which in turn will depend on the "position" of the teacher training institution. For example, a highly selective teacher training institution, one that attempts to train a small number of highly qualified teaching candidates, needs a selection process that identifies those candidates they predict will become knowledgeable and effective teachers. These schools will be concerned primarily with the predictive validity of the tests they use. These same institutions will need to guard against any bias that the tests may insert into their selection decisions.

Teacher training institutions that are not in a position to be as selective will be more concerned that their admission tests measure what is minimally required to teach, but which the institutions do not expect to teach as part of their preparation programs. These training institutions will be concerned that the tests they use for admissions validly measure that knowledge or those skills that must be possessed by those who will be licensed or certified to teach.

Initial Certification. The initial decision to certify one to teach follows teacher preparation (which may include practice teaching or some form of internship) and is the responsibility of a state governmental agency such as a state department of education or another certifying agency. The typical goal of certification testing is to validly, fairly and efficiently identify candidates minimally competent to teach in the state. Therefore, tests used in certification, generally paper and pencil tests, usually measure knowledge that the state has demonstrated is essential for beginning teachers in that state. As such, certification is strictly a state level licensing decision,

the purpose of which is to limit the number of incompetent teachers. Each state has its own certification requirements, although many states have reciprocal agreements to deal with teachers prepared out of the state in which they apply for certification. Most recent teacher testing policy developments have been in the area of teacher certification.

The certification decision is criterion referenced in that there is not a fixed quota of positions to fill at a given time. What a certification test determines is whether the teaching candidate is qualified to teach in the state, not whether they will be more or less successful. Theoretically, the percentage of candidates taking the test and passing it could vary from 0 percent to 100 percent as long as the test validly and fairly discriminates between those who possess the minimally required knowledge and those who do not.

The logic of certification testing requires that specific cutoff scores be set for each test being used. Candidates scoring above the cutoff are certified to teach while those scoring below the cutoff must either retake and pass the test, or fail to be certified. The test, therefore, should be effective at discriminating between acceptable and unacceptable candidates. Because of this requirement and the fact that any test is only a limited sample of behavior, tests for certification will not do a particularly good job of discriminating across a wide range of knowledge. An implication for policy is that a test that is both valid and efficiently designed to certify beginning teachers will probably not be a good test for identifying teachers with superior knowledge in the area tested.

One way in which certification testing policies differ is in how they provide for candidates to retake the test and what assistance, if any, is provided to help candidates pass the test. Policies differ in the amount of

time allowed between testings and the number of times a candidate can retake the test. A second important difference in the testing policies of different states is in the level of difficulty reflected in the cutoff score. While stringent cutoff scores imply more rigorous standards for who will be certified to teach, states with more rigorous standards may be those who have a larger supply of potential teachers or who allow for more easily granted provisional certification.

Information used to support certification decisions, which is not limited to test information, may include:

1. General knowledge
2. Knowledge of teaching methods
3. Knowledge in subject(s) the candidate plans to teach
4. Communication skills
5. Successful completion of an approved teacher training program
6. Commitment to teaching
7. Acceptable trial performance of teaching functions

Assuming that the candidate has completed training, it is reasonable to expect information more specific to functioning as a teacher and not to require information about more basic academic skills. Basic academic skills measures used to screen students for admission to teacher training programs have already been used implicitly in certification. But, of course, one of the reasons behind testing basic academic skills in the certification step is the concern that teacher candidates from institutions outside those under the control of the state in question may not have been subject to comparable

quality screens. One of the specific reasons for standardizing teacher certification testing, therefore, is to control for less than standard information on the quality of teacher candidates.

Published tests used to certify teachers include:

1. The National Teacher's Examination (NTE)
2. The Pre-Professional Skills Test (PPST)
3. The California Basic Educational Skills Test (CBEST)
4. The Georgia Teacher Certification Tests

A number of states have developed their own certification tests, contracting to such agencies as National Evaluation Systems (NES) and the Instructional Objectives Exchange (IOX). Oklahoma's program includes criterion-referenced tests for more than 75 different certificate areas (Folks, 1986).

Of the testing alternatives for certification, the NTE, which is published and managed by ETS, is used most widely. The NTE testing program, which began in 1940, comprises objective, standardized measures of academic preparation for teaching. The primary purpose of the NTE battery was to allow school systems "to evaluate the achievement of individuals from different colleges and universities which may have dissimilar standards and grading practices." (Rosenfeld, Thornton, & Skurnik, 1986, p. I-1). Recently revised, the NTE Core Battery now includes tests of professional knowledge, general knowledge and communications skills. The NTE Specialty Area Tests measure 27 content areas.

As of the summer of 1986, 17 states used the NTE Core Battery for teacher certification. Thirteen of these states also used the NTE Specialty Area

Tests for certification. To support the legal use of the NTE in a state, the test must be validated and cutoff scores established for that state. Validity studies must establish the content validity of the NTE and specific cutoff scores in relation to the minimum knowledge required to function as a beginning teacher in that state.

Modified tests and custom made tests are used in states where resources were available for their development. Whether a state decides to use a test off of the shelf or develop its own test, however, they are responsible for validating the test for use in their state and setting state standards of minimally acceptable performance. Unlike tests used for admissions to teacher training programs, tests used for certification must be validated in terms of job relevance, and cutoff scores must be based on what is minimally required to perform as a beginning teacher in a state.

Selection for Teaching Positions. A third use of teacher testing is in the process of selecting applicants for a teaching position. The norm is for local school policies, rather than the state, to prescribe how teachers will be selected for local positions. Only one state, Hawaii, has a state level policy specifying the use of tests for selecting teachers. They require use of both the NTE Core Battery and the Specialty Area tests as part of the information considered in hiring teachers.

Use of a test in support of hiring decisions is also subject to the Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission et al., 1978).

Recertification of Practicing Teachers. A fourth type of decision for which test data may be considered is to recertify currently certified teachers. Three states, Arkansas, Georgia and Texas, have implemented

programs to test teachers who are already certified. The Texas program, which uses the Texas Examination of Current Administrators and Teachers (TECAT), also tests administrators. These programs are subject to considerable controversy and are opposed by both the AFT and the NEA. States which require testing for recertification are forced to acknowledge that either job requirements have changed or that previous certification standards were unacceptably low. Notably, ETS has forbidden use of the NTE for recertification decisions.

Career Advancement. A final type of decision in which test data may be considered is to support career ladder programs in which teachers are given opportunities to increase their level of professional responsibility by taking on special assignments. Florida and Tennessee use tests in career ladder programs. As with using tests for recertification, there is little experience with using tests for career ladder programs. Controversy over these testing applications is based on their use to distribute financial advantage rather than the right to teach.

Individual Decisions

Individual decisions relate to setting personal goals and direction rather than making routine institutional decisions such as admissions, certification and selection. A teaching candidate may decide, for example, to concentrate on teaching in a specific content area to take advantage of an assessed strength, or he/she may decide to remedy a weakness revealed by a test. Shimberg (1981), discusses the use of self-assessment testing as a tool to ensure the continued competence of practicing professionals. In self-assessment testing, professionals voluntarily take tests with the assurance that they alone will know the results. The concept that underlies self-assessment testing is that some practitioners may be unaware of their own weakness and that with self-assessment much of the anxiety and opposition that relates to testing is eliminated or, at least, reduced. Individuals use their results to plan their own refresher training. Group data, although not representative of the population, may be used to plan educational programs for the professional group. Self-assessment testing can be combined with recertification testing to give practitioners a way to assess weaknesses prior to the "official testing."

Testing programs differ in the types of decisions they serve. Policy makers should consider the extent to which they wish their testing programs to serve institutional and individual decisions. Generally, this will mean that policy makers should decide whether tests designed to support institutional decisions should also have diagnostic utility.

What are the Requirements for Tests?

Does testing do what it is supposed to do? Is testing fair and efficient? Answers to these questions depend on clearly conceived, unambiguous testing purposes and means of assessing the consequences and costs of testing. These analyses should also consider the consequences of relying strictly on nontest information to make decisions, a point which many testing critics tend to overlook. Assessing consequences implies the need to examine at least two types of evidence:

1. Evidence of validity
2. Evidence of fairness or lack of bias

Because validity is also fundamental to unbiased decision making, we discuss it at greater length than fairness. A consideration that we do not discuss here, but which should also be taken into account is the cost of testing teachers.

Validity

Evidence for the validity of a test-based inference needed to implement a teacher testing policy is essential. A test that is invalid for an intended use (e.g., measuring knowledge of a defined content domain, predicting success in an academic program) is of no value for that use. Moreover, invalid test data will even be worse than useless if they displace valid nontest information. As discussed, testing teachers can serve different purposes (e.g., admission to teacher training programs, certification, selection for

teaching positions, career advancement and individual career planning). Appropriate validation procedures depend upon the intended test use and the interpretations required to support that use. Teacher testing policy options should be formulated with a clear understanding of the validation evidence required to support the intended test use. It is equally important to determine if it is feasible to conduct the type of validation study needed to support the inference necessary to support the policy.

Standards

Two essential sources of standards for test use, including standards for test validation, are the Standards for Educational and Psychological Testing (American Psychological Association, et al., 1985) and the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission et al., 1978). The Standards provides general guidelines for developing and using educational and psychological tests; the Uniform Guidelines specify the legal requirements for using tests to make unbiased employee selection decisions.

The Standards specify three strategies to produce validity evidence: construct-related, content-related and criterion-related. As defined in the Standards, "The evidence classed in the construct-related category focuses primarily on the test score as a measure of the psychological characteristic of interest. Reasoning ability, spatial visualization and reading comprehension are constructs as are personality characteristics such as sociability and introversion." (p. 9). Construct-related validity is important when the inference to be made from the test, references an attribute of the individual.

In defining the second type of evidence for validity, the Standards state that ". . . content-related evidence demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content." (p. 10). Note that as it is defined in the Standards, content-related evidence deals with what is covered in a test, rather than an attribute of an individual taking the test.

According to the Standards, evidence of criterion-related validity is used to demonstrate "that test scores are systematically related to one or more outcome criteria" (p. 11). Outcome criteria, although they are the variables of primary interest, are not used for routine assessment or decision making because they are too expensive to measure, or they are not available until some time after a decision is to be made. Predictive validity is a special case of criterion-related validity that applies when the outcome criteria are collected after the test results are known.

Admissions and Certification Tests

Tests used to support admissions decisions for teacher training programs should predict success in those programs; at a minimum they should screen out those who lack the basic skills to succeed in training. For that reason, tests validated as measures of academic aptitude are a frequent choice for admissions decisions. There has been a recent trend, however, to develop admissions tests that measure basic skills in reading, writing and mathematics to help screen out applicants who are deficient in skills that the teacher training institution does not include in its curriculum. These tests are developed to demonstrate content-related validity.

Recently, a good deal of attention has been given to validating certification tests, the purpose of which is to determine whether the teaching candidate meets the minimum knowledge requirements to be licensed in a state (Cross, 1985). In certification decisions, the test is expected to assess the minimum level of knowledge in a prescribed content area such as reading literacy, writing competence, general knowledge and subject specific knowledge. Validating a test to be used for certification decisions requires evidence that the test is content valid. Thus, the question guiding the validation of a test to be used for certification, is:

> Do the test items provide a representative sample of the content domain of essential knowledge for a beginning teacher? If they do, then performance on the test assesses the candidates' knowledge in relation to that which is required of a beginning teacher.

One of the problems in defining the appropriate content domain for a teacher certification test is that the full range of assignments a teacher may be given is not usually known when the teacher is certified. Consequently, the teacher must be prepared to do more than they will eventually be called upon to do. Therefore, they should be certified as minimally competent for a broader content domain than would seem necessary based on more specific assignments. In terms of the test validation process then, the content domain to be sampled is the sum of all the tasks and knowledge that the prospective teacher could be called upon to use by virtue of their certification.

Identifying the appropriate content domain for teacher certification testing has been the subject of some strategy differences in the last 10 years. These differences point out that tests are often used and validated

for uses other than those originally intended by the test developer and that the political consequences of legal and technical standards has had great impact on interagency relations.

Two content domains that have been of interest for certification tests are:

1. The curricula of teacher training institutions (curriculum relevance)
2. The knowledge needed to perform the job tasks required of teachers in the state (job relevance)

In the former case, content validity requires evidence that test items reflect the content of what is taught in the teacher training institutions; in the latter case, content validity requires evidence that the test items reflect the content knowledge necessary to be able to perform as a teacher in the state. Initial controversy over which of these two content domains is most appropriate had been resolved in favor of job relevance for certification. Some states, however, continue to examine both curricular relevance and job relevance. This strategy has the side benefit of involving both public schools and teacher training institutions in defining what is essential knowledge for a beginning teacher.

A popular method for establishing the content validity of an existing test is to have a panel, or panels, review and rate individual test items on a scale assessing the relevance of the content measured by an item to the knowledge expected of an entry level teacher. The item judging method is required for validating the NTE. A limitation of this form of validation, which begins with a given set of test items rather than a defined content domain of essential skills or knowledge, is that it ignores content knowledge essential to performing teaching tasks but not covered in the test. Some validity studies have included steps to assess the comprehensiveness of

existing tests by identifying knowledge important to acceptable performance as a beginning teacher but not covered in the test being reviewed (Poggio, et al., 1986). This validation tactic adds important evidence in validating existing tests for new purposes.

A more rigorous and costly alternative to the judgmental procedures that comprise content validation strategies is to conduct a job analysis of teaching. Job analyses frequently take the form of surveys of practicing teachers to determine the job demands for skills and knowledge and the frequency with which those skills and knowledge are required. Job analyses may also involve direct observation, interviews and analysis of records. ETS has recently reported results of an analysis of the important job tasks that cut across specific grade levels and subject matter specialties (Rosenfeld, Thornton, and Skurnik, 1986). Their methods of defining the content domain of knowledge essential for a beginning teacher included literature searches, use of advisory committees, and interviews with teachers and administrators. Their research, because it focuses on more generalizable content validity than state specific validity studies, will be useful to those interested in the feasibility of multistate or national validation strategies.

A process highly related to validating tests for certification, therefore mentioned here briefly, is to establish a cutoff (e.g., passing) score using the items judged to measure some aspect of the content domain. Methods to set cutoff scores are described in Berk (1986) and Nassif (1986). The importance of the cutoff score in terms of policy is twofold. First, it symbolizes the level of knowledge required of a beginning teacher in a state. Second, it, along with the quality of the applicant pool, determines the percentage of those who will pass the test.

An important feature of paper and pencil certification tests for which only content validity has been established is that the interpretations they logically support are limited. First, a paper and pencil test can only assess some of what may be essential for a beginning teacher to be minimally competent to teach in a state. Paper and pencil tests can assess knowledge and the capacity to manipulate that knowledge within the format of the test. Such tests, however, do not measure the capacity of candidates to apply their knowledge in the classroom, which can only be measured by a direct assessment of actual performance. A test of knowledge, in other words, cannot be expected to be a comprehensive assessment of minimal teaching competence. This is an important point, as certification tests, in general, do not purport to measure teaching competence, but rather measure selected knowledge domains judged to be essential for beginning teachers.

A second limitation of tests validated for certification decisions, which has also been a frequent point of confusion, is that tests of minimal competence used for certification are not necessarily expected to discriminate between more or less knowledgeable or effective teachers. For those who recall reading or hearing the comments of teachers after taking the Texas Examination of Current Administrators and Teachers (TECAT) the confusion can be tied to actual experience. Comments, such as "The test didn't really challenge me. It was too easy," and "The test did not measure teaching competence," reveal a mistaken notion of what the test was supposed to do. While it will not be possible to prevent such mistaken interpretations, it will be important for those using the tests to avoid misinterpretations and to promulgate appropriate information regarding the proper uses and interpretations of the tests.

In the last several years we have seen a reduced emphasis on evidence of criterion-related validity in certification testing. Problems measuring the criteria and other design difficulties, especially the need to assign people to jobs regardless of their performance on the test, are cited as reasons for relying on content-related validity studies.

Some would go so far as to say that we would not expect performance on a test of knowledge to correlate with effective teaching. Passing a test that informed judges say is representative of the content knowledge required of a beginning teacher and, therefore, should be passed by an applicant is seen as sufficient reason for using that test to certify teachers. The Uniform Guidelines for Employee Selection, which have been used to lend support to this argument, include the following provisions:

1. Empirical data should be made available to establish the predictive validity of a test, that is, the correlation of test performance with job-relevant work behaviors; such data should be collected according to generally accepted procedures for establishing criterion-related validity.

2. Where predictive validity is not feasible, evidence of content validity (in the case of job knowledge proficiency tests) may suffice as long as appropriate information relating test content to job content is supplied.

3. Where validity cannot otherwise be established, evidence of a test's validity can be claimed on the basis of validation in other organizations as long as the jobs are shown to be comparable and there are no major differences on context or sample composition.

4. Differential failure rates (with consequent adverse effects on hiring) for members of groups protected by Title VII constitute discrimination unless the test has been proven valid (as defined above) and alternative procedures for selection are not available.

5. Differential failure rates must have a job-relevant basis and, where possible, data on such rates must be reported separately for minority and nonminority groups.

Taken together, these five statements from the Guidelines for Employee Selection highlight the link between test validation requirements and ensuring that tests used are not biased against members of protected minority groups, a point which we turn to next.

Lack of Bias

It is essential that tests used for decisions affecting opportunities granted to individuals be free from bias against protected minority groups. Of major concern is that the content of tests used to make selection decisions be appropriate for minority group members. One method to guard against including content inappropriate for a minority groups is to provide for a review of all items for racial or cultural bias. Such a review would be in addition to a review for job-relevance.

Even with reviews for bias, however, differential passing rates favoring nonminority group members on admissions tests, certification tests and recertification tests are a well-documented fact. In a most recent case, the differential passing rate for Texas teachers and administrators taking the

TECAT revealed that while 96.7 percent of the 202,084 educators passed the test, only 94 percent of the 24,685 Hispanic educators and 81.6 percent of the 15,681 black educators passed the test (Education Week, 1986). Kauchak (1984), commenting on a by-product of teacher certification testing in Louisiana, reports that between 1978 and 1984, "only 15 percent of the 1,394 black students from public institutions who took the NTE achieved a passing score" (p. 627).

Two recent reports by staff from the ETS review data on differential passing rates for teacher certification tests (Goertz, Ekstrom and Coley, 1984 and Goertz and Pitcher, 1985). In California, where the CBEST is used for certification, the passing rates were 76 percent for white test-takers, 39 percent for Hispanic test-takers, and 26 percent for black test-takers. The passing rates for whites and blacks taking the Georgia Teacher Certification Test, developed by NES for Georgia was 87 percent and 34 percent respectively.

In cases such as these, which show the use of a test to adversely affect one or more minority groups, the Uniform Guidelines suggest that the state be prepared to prove that the test is valid in terms of job-relevance and that another less damaging but still valid selection procedure is not available. The legal requirements for teacher testing, however, remain a critical issue requiring careful attention by one qualified to give legal advice.

A more far reaching implication of teacher testing policies, however, is the affect of such selection on the composition of the teaching force. Forecasting the consequences of the move to require testing for certification, Goertz and Pitcher (1985) project that by the year 2,000, the percentage of minorities in the teaching force could be cut nearly in half. They point out that at the same time the proportion of minority students enrolled in school will rapidly increase.

The issue of bias and differential passing rates is certainly not closed. The performance of minority groups on certification tests is confounded with the selectivity and effectiveness of teacher training institutions that serve different proportions of minorities and nonminorities.

POLICY OPTIONS

It is too common that the underlying purpose of testing and the range of decisions that testing can support are given inadequate consideration in formulating policy. We recommend that deliberations over policy options for testing teachers begin with a consideration of the underlying purposes. We suggested four such purposes in the section on Why Test Teachers. They are:

1. Limiting the number of incompetent teachers
2. Encouraging teacher professionalism
3. Promoting public confidence in the teachers as a group
4. Promoting excellence in education

A second step in the development or review of policy options is to determine the implications for achieving the more fundamental purposes by introducing tests or altering the way tests are used in each of the following five types of institutional decisions:

1. Admitting candidates to teacher preparation programs
2. Certifying teachers as sufficiently competent to teach
3. Selecting certified teachers for specific positions
4. Recertifying practicing teachers
5. Granting promotion, rewards or special status

The third step will be to review the requirements needed of tests to be used for the purposes intended. The primary concerns are for validity and lack of bias. Other lesser but still significant concerns are discussed in the next section on pitfalls of teacher testing policies.

WHAT ARE THE PITFALLS OF TEACHER TESTING PROGRAMS?

Pitfall Number 1: Failure to Establish Clear Purpose

As we have discussed, there are at least four general purposes for testing teachers. They include::

1. Limiting the number of incompetent teachers
2. Encouraging teacher professionalism
3. Promoting public confidence in the teachers as a group
4. Promoting excellence in education

While a negotiated policy will contain compromise, it should, nonetheless, retain a clear enough purpose to maintain support for its implementation and to determine if the policy is accomplishing what it is intended to accomplish. A testing policy that focuses on limiting the number of incompetent teachers will focus attention on passing a test at a low level of difficulty, one that reflects minimal knowledge required to function as a teacher. A test designed to screen out teachers who lack essential knowledge cannot be expected to promote greater professionalism among the teaching force, and it may do little to promote public confidence in the teaching force, as it may diminish further the perception that the public has about what it takes to be a teacher. More importantly, the goal of promoting excellence calls for moving beyond minimal competence and public perceptions.

Guideline Number 1. Policy formation debate should include a discussion of the policy goals for teacher testing proposals. Any policy established should include a clear statement of what educational improvement is expected to result from that policy's implementation.

Pitfall Number 2: Setting Unrealistic Expectations

The second pitfall, setting unrealistic expectations, is closely related to Pitfall Number 1. As we have said earlier, tests are an appropriate tool for identifying teachers who are incompetent only by virtue of less than satisfactory knowledge in those areas judged essential to functioning as a beginning teacher in a state. A test will not screen out teachers whose lack of competence is independent of essential knowledge. For example, failure to maintain classroom discipline, which is the most common problem leading to dismissal of teachers, requires periodic assessment of actual performance rather than a paper and pencil test on classroom management.

Higher level purposes, such as creating greater spirit of teacher professionalism, require much more than a testing program to be successful. Greater rewards, including but not limited to higher pay, will be necessary to accomplish these higher purposes. Granting talented teachers greater responsibility and opportunities for self-determination are also strong incentives.

Guideline Number 2. Establish realistic expectations for what can be expected from implementing any teacher testing policy you consider. Promulgate these expectations among stakeholder groups. Test them against the experience of others who may have implemented similar policies. Obtain expert advice from advocates and opponents of the proposed policy. Establish a plan to evaluate the policy once implemented.

Pitfall Number 3: Inadequate Funding

It should go without saying that to establish a new program requires resources. Even programs that run with "existing resources" force reallocation of staff time which means that other activities are not carried out. Policies that require new test development with little attention to cost, regardless of the source, run the risk of resulting in a poor product, failure to meet timelines and failure to accomplish other important tasks.

Guideline Number 3. Provide for adequate funding to implement policies established. To do this may require that a detailed plan of implementation be developed before making a final judgment about the policy.

Pitfall Number 4: Unrealistic Timelines

Implementing a new program too quickly carries a number of risks. The quality of work may suffer, and political support will be jeopardized. A particular problem with new policies for testing teachers for certification or for recertification is providing adequate time between announcing new requirements and requiring teachers to pass the new test. Even though the test should be based on job requirements in the state, there should be ample time for teacher training institutions to respond with appropriate curriculum revisions and give teacher candidates an opportunity to learn the content to be covered on the test.

Guideline Number 4. Policies should require the development of detailed implementation plans with timelines adequate to support implementing such policies as are established.

Pitfall Number 5: Failure to Resolve Stakeholder Differences

A number of different groups have a vested interest in teacher testing policy. Failure to solicit their support and negotiate differences can lead to such problems as resistance, delays in implementation, lack of financial support, public criticism and less desirable counter proposals depending on the group responding.

Parent groups may support policies simply as a way to keep unqualified teachers out of the classroom; teacher associations may openly criticize policies they have had little opportunity to influence and which, in their eyes, undermine the development of a more professional image for the teaching force; teacher training institutions may see the certification testing policies as an unfair and narrow attempt to hold them accountable for their preparation programs.

Guideline Number 5. Identify and involve stakeholder groups during stages in which policy is being reviewed and formulated. Promulgate information that reveals developments in policy and seek expert advice regarding options being considered.

Pitfall Number 6: Failure to Meet Legal Requirements

Testing policies that impact on teacher selection decisions directly or indirectly (as does certification) are subject to the 1978 Equal Employment Opportunity Commission Guidelines on Employee Selection Procedures, which are intended to guide implementation of Title VII of the Civil Rights Act of 1964 and prevent employment discrimination based on sex, race, color, religion or national origin. Failure to follow necessary validation steps and to report

on differential failure rates for minority and nonminority groups can lead to legal action against the state. Because the Guidelines leave room for interpretation, the threat of legal action is a risk with innovative certification requirements that employ testing. The pending court case in Texas regarding the legality of using the TECAT is an example.

Guideline Number 6. Review legal requirements and key cases. Keep pace with legal opinion and actions related to new policy options and seek legal advice about innovative policies. Be sure to follow accepted procedures for test validation for the purposes they are to fulfill. Document procedures and decisions.

Pitfall Number 7: Inadequate Test Security

Tests used to make decisions that impact on the lives of individuals are prone to test security problems. This can be particularly troublesome for generally available achievement tests such as the California Achievement Test (CAT) (not to be confused with the CBEST) which some states have used for admissions decisions. It can also be a problem for states that implement policies calling for tailor-made paper and pencil tests unless planning allows for the periodic development of new test items and test equating. Because of the financial rewards involved, tests used in career ladder programs are also prone to problems of test security.

Guideline Number 7. In addition to guarding against unauthorized release of tests, one ought to assume that items will need to be revised periodically.

Pitfall Number 8: Dealing with Low-Incidence Content Areas

A problem for policies that call for testing in a number of different content areas is the low-frequency demand for the test. That is, the number of teachers seeking certification in some content areas may be so few that the economic wisdom of developing and maintaining a test for that subject area is questionable.

Guideline Number 8. Policies should be reviewed to identify cost inefficiencies and, where present, they should be eliminated. In considering cost, however, attention should be given to the benefits of testing as well.

Pitfall Number 9: Changing Job Requirements

Significant restructuring of teaching responsibilities or developments in subject area knowledge domains will affect the content domain of required job knowledge. To the extent that a test is designed to sample very specific content, it will follow that such a test will need to be periodically revalidated and new items added. A more general test, one that relies on more generalizable competencies, will be less subject to changes in specific content domains.

Guideline Number 9. Teacher tests should be periodically reviewed to assure that they retain their content validity against the domains they are intended to sample.

References

American Psychological Association, American Educational Association, & National Council on Measurement in Education (1985). Standards for educational and psychological tests and manuals. Washington DC: American Psychologial Association.

Anrig, G. R. (1986). Teacher education and teacher testing: The rush to mandate. Phi Delta Kappan, 67, 447-451.

Baker, C. C., & Fennell, Barbara (1986). The Alabama english language proficiency test: A criterion for admission to teacher education programs. In W. P. Gorth & M. L. Chernoff (Ed.), Testing for teacher certification (pp. 253-265).

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research. 56(1), 137-172.

Bracey, G. W. (1986). Pandora and pollyanna: Some comments on the rush to mandate. Phi Delta Kappan, 67, 452-455.

Bridges, E. M. (1986). The incompetent teacher. Philadelphia: The Falmer Press.

Carnegie Forum on Education and the Economy's Task Force on Teaching as a Profession. (1986). A nation prepared: Teachers for the 21st Century. New York: the Carnegie Forum on Education and the Economy.

Cronbach L. J. & Gleser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana, IL: University of Illinios Press.

Cross, L. H. (1985). Validation of the NTE tests for certification decisions. Educational Measurement: Issues and Practice, 4(3), 5-7.

Darling-Hammond, L. (1986, July 16). We need schools willing and able to use Carnegie's "teachers for the 21st century." The Chronicle of Higher Education, p. 76.

Equal Employment Opportunity Commission. Civil Service Commission. Department of Labor and Department of Justice. (1978, August) Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. Federal Register, 43(166), 38290-38315.

Folks, John (1986). Stakeholders in teacher certification testing. In W. P. Gorth & M. L. Chernoff (Ed.), Testing for teacher certification (pp. 47-57).

Gallup, A.M. (1986). The 18th annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 68(1), 43-59.

Gallup, G. H. (1984). The 16th annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan. 66(1), 23-38.

Goertz, M. E., Ekstrom, R. B., & Coley, R. J. (1984). The impact of state policy on entrance into the teaching profession (Final Report to the National Institute of Education, NIE Grant No. G83-0073). Princeton, NJ: Educational Testing Service.

Goertz, M. E. & Pitcher, B. (1985). The impact of NTE use by states on teacher selection. Princeton, NJ: Educational Testing Service.

Jaeger, Richard M. (1986). Policy issues in standard setting for professional licensing tests. In W. P. Gorth & M. L. Chernoff (Ed.), Testing for teacher certification (pp. 185-197).

Jaeger, Richard M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 18, 22-27.

Kauchak, D. (1984). Testing teachers in Louisiana: A closer look. Phi Delta Kappan. 65(9), 626-628.

Nassif, P. M. (1986) Teacher certification testing technical challenges: Part I. In W. P. Gorth & M. L. Chernoff (Eds.), Testing for teacher certification (pp. 117-137). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Pecheone, Raymond L., Tomala, Gail & Forigione, Pascal D. Jr. (1986). Building a competency test for prospective teachers. In W. P. Gorth & M. L. Chernoff (Ed.), Testing for teacher certification (pp 99-113).

Poggio, J. P., Glasnapp, D. R., Miller, M. D., Tollefson, N., & Burry, J. A. (1986). Strategies for validating teacher certification tests. Educational Measurement: Issues and Practice, 5-2), 18-25.

Rosenfeld, M., Thornton, R. F. & Skurnik, L. S. (1986). Analysis of the professional functions of teachers: Relationships between job functions and the NTE core battery (Research Report No. 86-8). Princeton, NJ: Educational Testing Service.

Shanker, Albert (1985, Fall). A national teacher examination. Educational Measurement: Issues and Practice, 4(3), 28-31.

Shimberg, B. (1981). Testing for licensure and certification. American Psychologist, 36, 1138-1146.

Schulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 515(2), 4-14.

The Holmes Group, Inc. (1986). Tomorrows teachers: A report of the Holmes group. East Lansing, MI: Author.

U.S. Department of Health, Education, and Welfare, Public Health Services (1977). Credentialing health manpower (DHEW Publication No. (05) 77-50057). Washington, D.C.: Author.

Vorwerk, Katherine E., & Gorth, William P. (1986). Common themes in teacher certification testing program development and implementation. In W. P. Gorth & M. L. Chernoff (Ed.), Testing for teacher certification (pp. 35-43).