DOCUMENT RESUME

ED 447 190 TM 032 078

AUTHOR	Crehan, Kevin D.; Hess, Robert K.; D'Agostino, Jerome V.
TITLE	The Technology of Teacher Licensing Testing: A Discussion of Issues and Recommendations for Practice.
PUB DATE	2000-04-00
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cognitive Processes; *Job Analysis; *Licensing Examinations (Professions); Reliability; *Teacher Certification; Teacher Evaluation; Teachers; *Test Construction; *Validity
IDENTIFIERS	Test Specifications

ABSTRACT

This paper focuses on teacher testing issues related to job analysis, test specification development, reliability, and validity. It emphasizes the conceptualization and operational definition of appropriate validity evidence to assess the quality of licensure testing decisions. It is suggested that the process of job, or practice, analysis would be improved by adopting strategies that are more analytic of the settings, activities, and competencies involved in teaching and which focus on cognitive processes that go beyond the recall of facts. Reliability is not sufficient to establish the credibility of a test, but without adequate and appropriate reliability, a test cannot be judged as valid. There is some debate over the necessity of documenting evidence of predictive validity during the validation process, but many educators believe that predictive validity evidence is essential in teacher testing. Others note that micro-level validity studies are also important, and that information must be gathered at the item quality level to insure that the validity decision is correct. Test specifications, the technical qualifications of the people who write the test items, and the documentation of each item as it is developed are all important. It is also important to examine the statistical characteristics of the items. The responsibility for gathering all validity evidence lies in the hands of the contractor for the tests, who must compile the validity evidence, review the evidence provided by the developer or other authorities, summarize the evidence, and establish the validity of the use of the results of the test. (Contains 39 references.) (SLD)

Running head: TEACHER TESTING TECHNOLOGY

The Technology of Teacher Licensing Testing:

A Discussion of Issues and Recommendations for Practice

Kevin D. Crehan

University of Nevada, Las Vegas

Robert K. Hess

Arizona State University West

Jerome V. D'Agostino

University of Arizona

Paper presented at the annual meeting of the American Educational

Research Association, New Orleans, LA, April 2000

Introduction

In response to concerns for the quality of instruction in their schools, a number of states have passed legislation mandating teacher licensure testing. Presently, 43 states require examinations for elementary licensing and 44 states have testing requirements for secondary licensure (Mitchell & Barth, 1999). Although there has been a recent surge, the use of tests to certify teachers is not new in American schooling. Teacher certification tests have been in existence for over 60 years. At the request of a group of superintendents in the spring of 1939, the National Committee on Teacher Examinations was appointed by the American Council on Education to oversee the construction of a teacher certification test. The Carnegie Foundation for the Advancement of Teaching provided funds for the development of the project, and the National Committee assigned to the Cooperative Test Service of the American Council on Education the task of preparing the annual forms of the battery examinations to be used. The National Teacher Exams (NTE) were administered for the first time in 1940. Full responsibility for preparing, administering, and scoring the examinations was transferred to the Educational Testing Service (ETS) in 1950. Until 1983, when the test was revised, scores were compared to the original norm group, which took the test in 1940. ETS replaced the NTE with a new set of teacher examinations labeled the PRAXIS series. At this time, ETS and National Evaluation Systems (NES) are the only test companies that provide teacher certification tests.

The case for certification is built on certain fundamental principles (American Association of Colleges of Teacher Education [AACTE], 1984). First, each state has the right and obligation to determine whether or not an individual is competent to practice

before an individual is allowed to do so in those situations where incompetent practice in an occupation may lead to harm or injury. Because teaching involves the imparting of knowledge, and the development of skills, attitudes, and values that are vital for citizenship, the potential for harm is present. Also, inappropriate teaching procedures can lead to short and long term bodily and psychological harm to students. Second, certification standards that stipulate completion of certain education and training experiences are believed to be necessary to ensure that candidates have prepared themselves adequately for the competencies
they will be expected to demonstrate and for the tasks they will be expected to perform.

The fundamental purpose for testing is to remove from the teacher pool those with inadequate preparation and, by inference, to improve classroom instruction. According to Pugach and Raths (1983) and Hyman (1984), this purpose is predicated on the following set of major assumptions:

- a person must have a minimal level of content and professional knowledge to teach,
- valid tests can be designed that measure teacher knowledge,
- there is agreement among educators regarding what to test,
- a meaningful cut-score can be set,
- those that pass the test will be more effective teachers than those that do not,
- hiring teachers who pass certification tests will improve student learning, and
- public opinion of teachers and the teaching profession will improve.

The AACTE principles appear to sanction state legislative mandates for teacher certification. However, implicit in this type of legislative action is the assumption that the technology and expertise necessary to develop defensible licensure measures exists and is accessible to the executive branch charged with implementation of the law. Additionally, the willingness on the part of the state to establish minimum standards for entry level teacher licensure presupposes the willingness on the part of the state to meet minimum standards for the measurement procedures used in the licensure process. Unfortunately, the condition of technology available for the development of licensure testing has not reached a level commensurate to that of other legislative commissioned projects, e.g., road construction. Some general guidance toward this end is provided by the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985; 1999), the Equal Employment Opportunity Commission's (1978) Uniform Guidelines on Employee Selection Procedures, the Society for Industrial and Organizational Psychology's (1987) Principles for the Validation and Use of Personnel Selection Procedures, the Code of Fair Testing practices in Education (Joint Committee on Testing Practices, 1988), Development, Administration, Scoring and Reporting of Credentialing Examinations: Recommendations for Board Members (Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance, 1993a), and the Principles of Fairness: An Examining Guide for Credentialing Boards (Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance, 1993b). These standards, principles, and guidelines are available to the licensing authority and contractor to guide the test development. However, the lack of specificity allows some latitude in practice that permits the development of test products of variable quality. The ultimate responsibility for the quality of the testing program rests with the licensing body and, therefore, it is incumbent on this body to assure the legal defensibility of decisions supported by test results. Additionally, it is not sufficient to relegate this responsibility to the testing contractors. Madaus (1992) raised awareness of the issue in arguing that high stakes test development was too important a public concern to allow the test developers to be the sole determiner of test quality. He proposed the establishment of an independent auditing mechanism for high-stakes test development. Downing and Haladyna (1997) followed Madaus in proposing external evaluation of high-stakes testing programs and suggest a model for external review. Existing internal and external programs of review are discussed (Educational Testing Service, 1984; National Association of State Boards of Accountancy, 1994). The concerns expressed here are in agreement with Madaus and Downing and Haladyna. While a complete detailing of all the issues and options in licensure test development and validation is well beyond the score of this forum, we will attempt to address some to the major areas of concern.

The presentation focuses on issues related to job analysis, test specification development, reliability, and validity. It will emphasize the conceptualization and operational definition of appropriate validity evidence to assess the quality of licensure testing decisions.

Job Analysis (Practice Analysis)

Typical guidelines for the conduct of a job analysis (cf. Council on Licensure, Enforcement and Regulation & National Organization for Competency Assurance, 1993a; Clifford, 1994; Henderson, 1992) suggest that subject matter experts would generate a list of job tasks and related knowledge, skills, and abilities. This listing would be structured into a survey of job incumbents for ratings of task performance frequency and criticality for public protection. The results of the survey are used to construct a table of specifications which guides test construction. Kane (1997) proposed a more systematic strategy for this step in test development preferring the label "practice analysis" to job analysis (see also Knapp & Knapp, 1995). Kane suggests that practice analysis is the preferred label since a job analysis suggests a focus on a particular job. Since licensure provides access to professional practice in a wide variety of settings rather than a particular job, practice analysis is the more appropriate term. Kane describes the goal of a practice analysis as the description of patterns of practice in the profession. Kane assumes that the professionals are operating from the same general knowledge base and applying similar methods. This assumption seems to fit with the practice of teaching. The process Kane describes for structuring a practice analysis provides a model, which results in a more systematic breakdown of the job related tasks than the traditional job analysis. The model provides for specification of settings, activities, and competencies. The estimates derived from the formula driven model provide an assessment of the importance of each competency category. Once developed, the model serves as a theory of practice, and the practice analysis is an empirical test of the theoretical model. The most apparent appeal of the proposed model is that it breaks the often daunting task of job analysis into more manageable parts. It would be most informative to attempt development of a theoretical model of practice for an elementary teacher, say, in a bilingual classroom. (This assumes the identity "elementary teacher" is sufficiently homogeneous.) Once developed, the model could serve as a "straw man" to elicit criticism from interested constituencies and, subsequently be revised and empirically tested.

The process of planning and delivering instruction clearly involves complex cognitive processes. However, it appears that much of the job analysis in teacher education results in specification of knowledge level outcomes (Rosenfeld & Kocher, 1998; Tannenbaum & Rosenfeld, 1994; Popham, 1992) and most of the content of initial certification tests is at the simple recall level (Mitchell & Barth, 1999). Given that test content comes from test specifications, which are driven by job analysis, perhaps the strategies used in job analyses need to be modified. A modification of traditional job analysis practice which has promise for improvement in content specifications in teacher licensure testing is cognitive task analysis (Redding, 1992; DuBois, Shalin, Levi, & Borman, 1995; Hanser, 1995). Traditional job and task analysis focuses on the behavioral aspects of job performance. A cognitive task analysis attempts to develop an understanding of the cognitive components of the task and provide a description of these components. Cognitive task analysis determines the mental processes that underlie performance. Very briefly, cognitive task analysis involves a description of the cognitive processes that underlie performance including conceptual and procedural knowledge as well as generative knowledge. Conceptual and procedural knowledge relate to the background knowledge and the "how to" of developing instruction while the generative knowledge addresses the "why" which supports what is done. That is, generative knowledge processes involve analysis of new or transfer problem situations and the framing and classification of relevant dimensions of the situation. Ideally, this analysis leads to an adaptation of conceptual and procedural knowledge that is appropriate to the situation.

It is suggested that the process of job analysis, or more appropriately, practice analysis, would be improved by adopting strategies which are more analytic of the settings, activities, and competencies involved in teaching (Kane, 1997) and have adequate focus on cognitive processes that go beyond recall of facts. Additionally, translation of the practice analysis and survey data to test specifications should use a strategy which assures the more critical aspects of the teaching process, e.g., generative processes, are given appropriate weights. Kane, Kingsbury, Colton, and Estes (1989) suggest a multiplicative model that may serve well in this instance in that it gives greater weight to areas of practice judged as more critical to successful performance.

Reliability Considerations

Reliability is not sufficient to establish the credibility of a test, but without adequate and appropriate reliability a test cannot be judged as valid. Anastasi and Urbina (1997, p. 84) define reliability as "the extent to which individual differences in test scores are attributable to 'true' differences in the characteristic under consideration and the extent to which they are attributable to chance errors." Whether we define reliability in terms of consistency or stability, reliability must ultimately reflect the degree of "agreement between two independently derived scores" (Anastasi & Urbina, p 85). As Feldt and Brennan (1989) point out, the obligation of a test developer is to gather sufficient evidence about a test in order to account for all potential sources of error.

Salvia and Ysseldyke (1998) recommend following Nunnally's hierarchy for gathering test level reliability information:

1. Use alternate forms;
2. Use equivalent (split) halves; and
3. Use Cronbach's alpha.

They, as well as Anastasi and Urbina (1997), recommend that large-scale assessments with potential high stakes need reliability coefficients near or above .90.

Furthermore, Anastasi & Urbina (1997) recommend identifying pertinent characteristics that might distinguish one administration group from another (e.g., institution of training, age, sex, ethnicity, etc.). This permits determining if the reliability for any individual administration is influenced in any fashion by the range of individual differences in an administration group.

A third issue of reliability pertains to the issue of item sampling or content sampling, which is the extent to which scores on these tests reflect idiosyncrasies peculiar to the selection of items on any form of the test. Another way of asking this question might be, "Would someone working from just the test blueprints be able to develop a test that would result in essentially the same scores for examinees?" The best way to determine the quality of content sampling is through a balanced split-half reliability procedure (Anastasi & Urbina, 1997; Salvia & Ysseldyke, 1998).

Decision-Consistency

Anastasi and Urbina (1997) make note of the need for any test whose results will be used to classify students to undergo some type of decision-consistency analysis. As Berk (1984, p. 235) points out, "The initial choice of reliability category [for decision making] is contingent upon the intended interpretation of the test scores, type of decision to be made with these scores, and the consequences or loses associated with false mastery and false nonmastery decision errors." The determination of the quality of the decision can be established by computing the appropriate indexes of agreement and significance values (Anastasi & Urbina). Berk and Subkoviak (1984, 1988) both recommend the use of a threshold loss function to determine the extent of false classifications (master and non-master) regardless of their size. A reliability estimate and the z-score of the passing value are needed to estimate the consistency of accurate classification. Tests with higher reliability estimates and extreme cut-scores (e.g., 10% or 90% pass rates) tend to yield more classification agreement (Subkoviak, 1988). Nonetheless, in most certification testing, where the cut-score is not as extreme, the agreement rate is less than the reliability of the test, underscoring the importance of ensuring a very high reliability estimate.

Predictive Validity Considerations

There is some debate over the necessity to document evidence of predictive validity during the validation process. Downing & Haladyna (1996) claim that criterion-related validity should be examined for high-stakes tests, but do not recommend the form this validity should take, such as either concurrent or predictive validity. The Standards for Educational and Psychological Testing (AERA et al., 1985, 1999) offer slightly more direction for test developers. In the section, Standards for Employment and Credentialing, in the 1999 edition, it is stated that, "Reliance on local evidence of empirically-determined predictor-criterion relationships as a validation strategy is contingent on a determination of technical feasibility" (p. 159, Standard 14.3). The conditions that should be considered to assess feasibility include: (1) the stability of the job, (2) a relevant and meaningful criterion measure exists, (3) a representative sample exists, and (4) the sample is of adequate size. Furthermore, the authors of the standards suggest that local studies should be conducted only when little empirical evidence has been accumulated to study the predictive validity of the test in question. If local studies are conducted, researchers should consider certain possible contaminants and artifacts, as well as range restriction and missing data issues. It appears that, given the difficulty in conducting predictive validity studies that yield accurate results, the authors of the standards concluded that these studies should only be conducted under suitable conditions, and that conclusions should not be drawn from one local study.

Many educators, however, believe that predictive validity evidence is essential in teacher testing. These individuals would argue that if teacher licensure tests are designed to identify those meeting minimum standards for teaching, with the implication that those not meeting these minimum standards will be more likely to be at risk of serving the public, then reasonably, these exams should predict to some extent performance as a teacher. While a job analysis is an essential step in the test development process, it is not sufficient for producing accurate certification exams.

Item Validity Concerns

While most validity studies emphasize the gathering of information at the macro level, Downing and Haladyna (1997) remind us that micro-level validity studies are just as important. Furthermore, it is Anatasi and Urbina (1997) who remind us that it is our obligation to gather evidence about a test to account for all potential sources of error. Therefore, information must also be gathered at the item quality level to insure that the validity decision is correct. Downing and Haladyna have produced a series of recommendations related to the types of evidence that need to be gathered at the item level. Table 1 is drawn in part from this work and lists several critical types of evidence related to item level information. While Downing and Haladyna were speaking specifically to industry level licensure testing, their issues and recommendations hold true in teacher certification testing field.

Each of the nine points raised in Table 1 is crucial in the process for establishing the purpose, function, and appropriateness for each item to be included on a test. Specifically, the following concerns must be addressed as part of the process. First is the concern about the need to provide a full coverage of content. In other words, have the developer and/or contractor provided sufficient guidelines and specifications necessary to assure that an adequate items pool is available? One must be careful to insure that a sufficient number of items, keyed directly to content specifications, is available for the desired level of reporting of results. This, of course, necessitates the development of a very precise set of test specifications. Far too often either the test specifications are too broad or missing and/or the available pool of items is insufficient to cover the level of specificity dictated by the test specifications. If the test specifications

Table 1

Model of Item Validity Evidence: Qualitative Evidence

	Type of Evidence	Evidence Needed
1	Content Definition	Documentation of the method(s) used to select item content
2	Item Content Verification	Content experts' credentials; records of content-expert review process
3	Cognitive Behavior	Documentation of system used and its rationale; reports of any research using system
4	Item Writer Training	Documentation of methods, principles, written materials, and sample items
5	Test Specifications	Documentation of systematic link of test content to test specifications/test blueprint
6	Key Validation/ Verification	Policy and procedures for key verification; documentation of key validation results
7	Adherence to Item-Writing Principles	Evidence of compliance with rules and documentation of process used to review items
8	Item Editing	Credentials and experience of editors; editorial and style guidelines, documentation of edit/review cycle
9	Bias/Sensitivity Review	Documentation of bias/sensitivity review; rationale for policies; credentials of reviewers

Meet the appropriate level of detail, a trained person working from just these specifications should be able to develop an alternate form of a test that would result in essentially the same score for examinees (Anatasi & Urbina, 1997).

A second concern is the technical qualifications of the people writing the items. Specifically, are the item writers qualified to write items in the domain of skills and knowledge being tested? While professional item writers are desirable, it is even more desirable to have professional item writers who were trained and (previously) employed in the field for which they are writing the items. This adds substantial credence to content nature of the individual items far beyond mere adherence to professional item writing guidelines. Related to this concern for professional item writers is a concern for trained item editors. These persons should be very qualified in the construction (i.e., production) of items and item-types. These editors should insure that all items have been constructed following accepted item writing principles and guidelines.

A third crucial concern relates to the documentation of each item as it is developed. Every item to be included in a pool for a [teacher] certification test must be fully documented. This is particularly important in view of the fact that graduates from a multitude of [teacher] training institutions will be obligated to meet certification requirements by being successful on such a test. Far too often item writers, without a satisfactory background in the licensure or certification field, are guilty of writing items that may inappropriately exemplify a single perspective where a wide variety of perspectives may exist (such as in the field of teacher education). Thus, the requirement that every item be documented (i.e., referenced to a set of major current works) helps to prevent such a singularity effect. Similarly, outside experts must be utilized to examine the keys (and distractors) for the items. Such a use of external experts provides further assurance of the appropriateness for the item-to-content match. Any item included on a certification test must have total agreement by the experts as to the correctness of the key; debate, at this level, is not appropriate.

While much has been written about the need to insure a bias/sensitivity review for all certification tests, its importance cannot be understated. Either the test developer or contractor must assume responsibility to put together an appropriately representative panel to review every item from a fairness (sensitivity) perspective. However, this is not sufficient. It is also important to examine items from a statistical perspective employing one of the many DIF techniques (cf. Holland & Wainer, 1993). Angoff (1993) also warns that the mere statistical presence of a DIF effect should not be judged as sufficient and items found to exhibit such should be returned to a bias/sensitivity review panel for their consideration.

The preceding concerns, for the most part, focused on content-to-item issues and/or item development and review issues. It is just as important, however, that we examine some of the various statistical characteristics of such items. While any of a number of good textbooks on test and test item design (e.g., Linn & Gronlund, 2000) provide generic conventions for such concerns as difficulty (p value) and discrimination, it is very difficult to identify specific benchmarks for such values. What is an adequate level of item difficulty for licensure test? Should an average item difficulty be below, at, or above a potential decision point? In IRT models, items are chosen to maximize the amount of information about the examinees at the decision point (Hambleton, 1989), but in tests employing classical techniques this is not possible. Furthermore, since, for the most part, standards are set after the development of initial tests, selecting items at a hypothetical (potential - possible) decision point prior to the setting of a standard almost seems to be an oxymoron.

One would assume that the general rules related to item difficulty for a NRT would not necessarily be appropriate for most certification tests (i.e. mean item difficulty equal to .50, range of item difficulties from .10 to .90, information maximized in the middle of the distribution, etc.) (Linn & Gronlund, 2000). Since [teacher] certification tests are designed around the content of training, one would anticipate that a properly trained examinee would answer most items correctly. The intent and purpose is to establish pre-entry skill level (or knowledge level) of the examinee relative to a predetermined (standard) level of this skill and/or knowledge. The ultimate goal is not to differentiate among examinees but to differentiate the level of achievement (skill/knowledge level) attained by an examinee relative to a prescribed level. This is not particularly difficult if one employs an IRT model that allows items and item pools with known levels of difficulties to be utilized (Hambleton, Swaminathan, & Rogers, 1991).

We have similar problems with the concerns related to item discrimination. In classical analysis we typically employ point biserial correlations to help us ascertain the contribution of an item to overall test scores (Linn & Gronlund, 2000). Items with negative point biserial correlation values are readily removed from our pool. Similarly, items with distractors (incorrect options) that have positive point biserial values are removed. However, finding a benchmark that suggests a minimum positive value is not easy. What is too low a value? Is .10 too low, or .15, or .20? And what happens if different parties disagree? One party says .20 and another party says .10 or is this simply splitting hairs? In an IRT model (2-parameter or 3-parameter) items may be selected according to a determined range of the discrimination value (e.g., .75 to 1.5) but no clearly defined benchmark exists for this either (Hambleton, Swaminathan, & Rogers, 1991). In the 1-parameter model the issue of discrimination is moot since all items are assumed to have the same discrimination index.

A final item level check can be done through an analysis of trace line each option (Haladyna, 1994). Under these guidelines, a properly functioning correct option has a trace line that is monotonically increasing as the level of achievement increases. Conversely, the distractors should decrease monotonically as student achievement rises. Such an analysis will indicate nonfunctioning distractors (e.g., those selected by limited numbers of students) as well as properly or improperly functioning trace lines for correct options (e.g., nearly flat, indicating extremely limited discrimination or nearly vertical, indicating too much discrimination).

Ultimately the responsibility for gathering all validity evidence lies in the hands of the contractor for the tests. While the test developer is most often responsible for establishing content validity for their tests (Anastasi & Urbina, 1997), the final responsibility for all validity evidence must reside with the contractor. It is their responsibility to compile all the various forms of validity evidence, to review the evidence provided by the developer or other outside authorities, to summarize such evidence and establish the validity of the use of the results of a test.

References

AACTE Task Force on Teacher Certification. (1984). Emergency teacher certification: Summary and recommendations. Journal of Teacher Education, 35(2), 21-25.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anastasi, A. & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland and H. Wainer (Eds.) Differential item functioning (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 231-266). Baltimore: Johns Hopkins University Press.

Clifford, J. P. (1994). Job analysis: Why do it, and how should it be done? Public Personnel Management, 23(2), 321-341.

Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance. (1993a). Development, administration, scoring, and reporting of credentialing examinations: Recommendations for board members. Lexington, KY: Author.

Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance. (1993b). Principles of fairness: An examining guide for credentialing boards. Lexington, KY: Author.

Downing, S. M., & Haladyna, T. M. (1996). A model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. Educational Measurement: Issues & Practice, 15(4), 5-12.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. Applied Measurement in Education, 10, 61-82.

Dubois, D. A., Shalin, V. I., Levi, K. R., & Borman, W. C. (1995). A cognitively-oriented approach to task analysis and test development. (ERIC Document Reproduction Service No. ED 402 336)

Educational Testing Service. (1984). The ETS audit program. Princeton, NJ: Author.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38290-38315.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Linn (Ed.), Education measurement (3rd ed.). New York: MacMillan.

Haladyna, T. M. (1994). Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hanser, L. M. (1995). Traditional and cognitive job analyses as tools for understanding the skills gap. (ERIC Document Reproduction Service No. ED 383 842)

Henderson, J. P. (1992). Job analysis. In A. H. Browning, A. C. Bugbee, & M. A. Mullins (Eds.), Certification: A NOCA handbook. Lexington, KY: The National Organization for Competency Assurance.

Holland, P.W., & Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Hyman, R. T. (1984). Testing for teacher competence: The logic, the law, and the implications. Journal of Teacher Education, 35(2), 14-18.

Joint Committee on Testing Practices. (1988). Code of fair testing practices in education. Washington, DC: Joint Committee on Testing Practices, American Psychological Association.

Kane, M. (1997). Model-based practice analysis and test specifications. Applied Measurement in Education, 10(1), 5-18.

Kane, M., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing test plans for licensure and certification examinations. Journal of Educational Measurement, 26(1), 17-27.

Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation of validity. In J. C. Impara (Ed.), Licensure testing: Purposes, procedures, and practices. Lincoln, NE: Buros Institute of Mental Measurements.

Linn, R. & Gronlund, N. (2000). Measurement and assessment in teaching (8th ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Madaus, G. F. (1992). An independent auditing mechanism for testing. Educational Measurement: Issues and Practice, 11(1). 26-31.

Mitchell, R., & Barth, P. (1999). How teacher licensing tests fall short. Thinking K-16, 3(1), 3-23.

National Association of State Boards of Accountancy. (1994). Report of the CPA Examination Review Board. New York: Author.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. Applied Measurement in Education, 5(4), 285-301.

Pugach, M. C., & Raths, J. E. (1983). Testing teachers: Analysis and recommendations. Journal of Teacher Education, 34(1), 37-43.

Redding, R. E. (1992). A standard procedure for conducting cognitive task analysis. (ERIC Document Reproduction Service No. ED 340 847)

Rosenfeld, M., & Kocher, G. (1998). Task and knowledge areas important for middle school teachers of language arts: A transportability study. Princeton, NJ: Educational Testing Service.

Salvia, J. & Ysseldyke, J.E. (1998). Assessment (7th ed.). Boston: Houghton Mifflin.

Society of Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures (3rd ed.). College Park, Maryland: Author.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classification. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 267-283). Baltimore: Johns Hopkins University.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement, 25(1), 47-55.

Tannenbaum, R., & Rosenfeld, M. (1994). Job analysis for teacher competency testing: Identification of basic skills important for all entry-level teachers. Educational and Psychological Measurement, 54(1), 199-211.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032078

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Technology of Teacher Licensing Testing: A Discussion of Issues and Recommendations for Practice

Author(s): Kevin D. Crehan, Robert K. Hess, Jerome V. D'Agostino

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Kevin D Crehan

Printed Name/Position/Title: Kevin D. Crehan

Organization/Address:
College of Education
UNLV
4505 Maryland Pkwy
Las Vegas, NV 89154

Telephone: 702-895-4303

FAX: 702-895-1658

E-Mail Address: Crehan@nevada.edu

Date: 10-26-00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com