DOCUMENT RESUME

ED 246 095 TM 840 368

AUTHOR	Elliot, Scott M.; Patterson, Sandra
TITLE	Establishing Standards for Licensing and Certification Tests: Theory vs. Practice. Draft.
PUB DATE	Apr 84
NOTE	16p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE	Speeches/Conference Papers (150) -- Information Analyses (070) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Cutting Scores; Elementary Secondary Education; Occupational Tests; Scoring; *Teacher Certification; Teacher Qualifications; Test Construction
IDENTIFIERS	*Angoff Methods; *Test Revision

ABSTRACT

Exams which are periodically updated and revised also require a revision of the cut score to ensure that examinees who take the original form of the test, and those taking the revised form, are scored on an equivalent basis. This paper discusses available standard setting procedures as applied to establishing minimum passing scores for licensing and certification examinations. It describes the methods and procedures used in reviewing and revising cut scores for a teacher licensing test. These methods provide a practical, legally defensible, and technically sound approach to reevaluating cut scores for revised licensing and certification tests. They are based upon a modification of the Angoff procedure applied to tests undergoing revision. (Author/DWH)

Establishing Standards for Licensing and Certification Tests: Theory vs. Practice

Scott M. Elliot
Area Director, Licensing and Certification

Sandra Patterson
Project Coordinator, Licensing and Certification

National Evaluation Systems, Inc.
30 Gatehouse Road
P.O. Box 226
Amherst, Massachusetts 01004

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1984.

Establishing Standards for Licensing and Certification Tests: Theory vs. Practice

ABSTRACT

While there is a growing body of literature addressing methods and procedures for establishing minimum passing scores, there is little information available describing procedures for reexamining cutscores for revised tests. This paper discusses available standard setting procedures as applied to licensing examinations and describes the methods and procedures used in reviewing and revising cutscores for a teacher licensing test. The procedure described is based on a modification of the Angoff (1971) procedure applied to tests undergoing revision.

Establishing Standards for Licensing and Certification Tests: Theory vs. Practice

There is a considerable body of literature discussing approaches to setting standards on certification and licensing tests. The available literature provides useful information to test developers. However, most of this literature assumes that cut scores are set once and remain unchanged for the life of the testing program. This paper discusses available standard setting procedures, their applicability to establishing minimum passing scores for licensing and certification examinations, and describes procedures used in reviewing and revising cut scores for a teacher licensing test.

Standard Setting Procedures

One of the central concerns in developing a test for licensing purposes is establishing a cutscore or minimum passing score for the examination. Because of the legal and political environment surrounding decisions to issue a license based on an established cutscore, this represents one of the most significant points in the test development process. A number of standard settings are available to the practitioner, including Nedelsky (1954), Angoff (1971), Ebel

(1972), Jaeger (1978), Contrasting Groups and Borderline Groups (Zieky & Livingston, 1977). However, only a few of these methods have actually been used in setting cutscores for teacher certification testing.

Legal guidelines (EEOC Guidelines, 1978) require that cutscores established for personnel tests must bear an empirical and logical relationship to the job. However, until relatively recently most tests for licensing teachers did not establish cutscores that were either empirically derived or that systematically bore a relationship to successful performance on the job. For example, in a number of states the use of the NTE with an arbitrarily set cut score was legally challenged. Cases challenging the use of the NTE with arbitrarily set cutscores include: United States v. North Carolina (1975); Baker v. Columbus Municipal Separate School District (1976); and Georgia Association of Educators v. Jack P. Nix (1976). Where a cut-off score is used to determine those candidates that are qualified or unqualified, the user of the test must give sufficient proof that the cut-off was not established in a capricious or arbitrary manner.

In South Carolina in 1977, it was found that the use of the NTE resulted in adverse impact against blacks. However, the state decided to investigate the test, validate it in South Carolina, and set cutscores in a systematic, empirical fashion. The result was that some of the NTE tests were validated and approved for use in South Carolina.

Underlying most methods used are the procedures designed by Nedelsky (1954) and Angoff (1971). These procedures have been modified, consolidated, lengthened and abbreviated for use in several states. They are described and discussed below.

Nedelsky (1954). The Nedelsky model is one of the earliest "formal" cut score procedures. Nedelsky's (1954) approach requires a panel of judges to review each test item and determine which of the available response options the "lowest D student" should be able to reject as incorrect. Judges then record the reciprocal of the number of the remaining responses adjacent to the item.

The cutscore is then determined by totalling the reciprocals obtained. The data across judges is used as a basis for determining the final cut score. Depending upon the particular application, the mean or median cut score for the group of judges may be used as the cut score. A number of modifications of the Nedelsky (1954) procedure have been employed, most notably the substitution of "minimally competent person" for "lowest D-student."

Angoff (1971). In the Angoff (1971) method, expert judges review each test item and indicate the probability that a person with minimum competency can give the correct response. The Angoff procedure is easy to explain, easy to understand and easy to administer. It is less time consuming than Nedelsky's (1954) and can be used on open-ended items. Angoff (1971) describes the methodology as follows:

> . . .ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (Angoff, 1971, p. 515.)

In some applications, judges are asked to directly estimate the percentage of individuals they would expect to answer the item correctly.

Jaeger (1978). The Jaeger (1978) procedure maximizes the involvement of educational constituencies. In the North Carolina application, 700 persons convened in groups of 50 to proceed through the standard setting model. The procedure is as follows:

Judges were first required to take the exam that they would later rate. For each item judges were asked one of the following questions:

(1) Should every high school graduate be able to answer this item correctly?

(2) If a student does not answer this item, should s/he be denied a high school diploma?

Judges next received the results of the above survey questions as well as actual performance data. With this information, judges were asked to review and revise their initial judgments as they considered necessary.

The procedure then calls for recalculation of the judges' ratings, redistribution of the new ratings, and another judgement. Judges then received information on the proportion of students who would have passed or failed, as determined on the basis of the recommended cut-off scores.

Median scores were calculated by group (type and constituency), and the passing score was then set at the minimum median score calculated for a group.

This process is technically straightforward and involves iterative reviews, and the inclusion of normative student data.

Revising Cut Scores

Teacher licensing tests, particularly in those fields where procedures and content are subject to frequent change, require updating on a periodic basis. A number of conditions in the testing environment may require a revision of the examination. These conditions include:

- o changes in the job content
- o changing definitions of the minimally competent professional
- o available data on actual examinee performance
- o changing consequences of failing the examination
- o changes in the political climate surrounding the testing program (more stringent or lax standards)
- o changes in the legal environment
- o changes in the number of licensed personnel required in the field

Regardless of the rationale for revising a test, the revision will require a reconsideration of the cutscore in some fashion. Revising the cutscore for a teacher licensing test ensures that examinees taking the original form of the test and those taking the revised form are scored on an equivalent basis.

Traditionally, adjustments in cutscores required due to test revision (e.g., item replacement, item revision) have been accomplished through some form of equating. While equating may ensure that scores are adjusted for changes in test form difficulty, equating will not account for qualitative changes in test content. To account for changes in content, expert judgement must be called upon to reevaluate standards for the minimally competent professional. For example, in a teacher licensing test a set of 10 revised items may be substituted for an initial set of 10 items; while these items may be of equivalent difficulty statistically (p-values, logit values), differences in content my require differing levels of proficiency; the initial set of items may contain content that may require 70% mastery to be considered minimally competent and the set of replacement items may require 80% mastery to establish minimum competency.

Because revisions in test content, particularly in the case of criterion-referenced tests, result in changes that can not soley be accounted for through statistical equating, the cutscore needs to be reevaluated using expert judgment. In selecting an appropriate method for reevaluating the cutscore through expert judgment, a number of additional factors need to be considered. Most importantly, the methodology employed should maintain a strong "link" to the original test since much of the original content is retained on the revised form. This can be accomplished in two primary ways. First, if the original cutscore was determined by a panel of experts initially, this same panel can be recalled to reevaluate the cutscore to maintain consistency. Second, the methodology employed for reevaluating the cutscore should, in some form, rely on data collected from the administration of the original test. In addition to these considerations, the approach selected for cutscore revision should be both practical to administer and easily interpretable by the panel of experts selected.

After exploring the issues discussed above, the cutscore procedure suggested by Angoff (1971) was selected for the standard setting effort described below. The Angoff (1971) approach, as applied below presents a practical, easily interpretable method for revising a test standard. Moreover, as applied in this situation, the panel of experts that worked on the original test development, can be reassembled and data on item performance (item p-values) can be used as a basis for making judgments.

Cut Score Revision Procedures

Overview. The ten tests which underwent revision were initially developed in 1980 to assess teacher candidates' content knowledge in their fields of specialization. All teachers in the state for which the tests were developed are required to pass the content knowledge tests prior to receiving a license to teach. In order to ensure that the tests remain up-to-date, they are reviewed approximately every three years by a panel of content experts. In the Fall of 1983, 10 of the tests included in the program were reviewed to update the tests based on (1) changes occurring in the field and (2) information gained from the results of the first three years of test administration.

For each field reviewed, the original test development committee consisting of practicing educators (public schools and higher education) was convened. (In some cases additional committee members were added to replace original committee members unable to participate or to provide added expertise to the committee.)

Updating procedures. Updating of the tests was accomplished by having the committees review the pool of questions on the basis of topicality and based on item statistics obtained from test administration data. The committees identified out-dated information or terminology within individual test questions, then revised or replaced the questions identified as non-topical.

Before conducting the review, committee members were provided a brief training session. The purpose of the review was explained and the criteria used to judge the topicality of a test question were discussed. The committee members were first asked to go through the item review booklet, item by item, to determine if the content, terminology or correct answer for any item was not up-to-date. If an item was judged to be non-topical the committee members were instructed to circle the item code number in the item review booklet and briefly describe the nature of the problem. The committees discussed each of the items identified as non-topical and the necessary corrections were recorded into a "Master" item review booklet. Then each individual test question was reviewed in light of the results of the first three years of test administration; items requiring revision were identified and revised or replaced. Committee members were referred to the item statistics which provided the results for each question that had appeared on a test form. The reviewers were asked to pay particular attention to those questions for which more than 95% or less than 30% of the students answered the question correctly or for which a greater number of examinees selected an incorrect response than the correct response. These questions were reviewed to determine if revisions were required. If the committee felt the question would be improved by revising it, the item number was circled in the item review booklet and the reason for labeling the item as requiring revision was noted. The committee then discussed each item identified by one of its members as requiring revision and the necessary corrections were recorded in the "Master" item review booklet.

Minimum passing score review. In order to gain the information necessary to re-evaluate the minimum passing scores for the examination, each committee member was asked to independently judge the difficulty level of each item based on the procedures suggested by Angoff (1971).

Each committee member was instructed to envision the minimally competent entry-level teacher; the individual who has acquired the basic knowledge in the field to meet the minimum basic standards for teaching. After reading each item carefully the committee members were then asked to decide on the percent of minimally competent entry-level teachers whom they felt should answer the item correctly. To assist the committee members in making a decision, they were referred to the item statistics which indicated the percent of examinees who answered the item correctly for the first three years of test administration; this information was provided only for those items that had been included on a test form. Committee members were cautioned that changes made to the items during the updating portion of the review may increase or decrease the difficulty level of the items. They recorded their decisions on the minimum content knowledge rating form which listed each item by its code number.

The results were analyzed according to procedures suggested by Angoff (1971). The mean (difficulty) rating across judges was computed for each item. The sum of the mean ratings provided the preliminary cutscore for the revised examination. For example, one exam with five items having mean ratings of 80%, 75%, 85%, 80%, and 70% would have a preliminary passing score of 390 or 78%.

```
Example:  .80
          .75
          .85
          .80
          .70
         3.90
```

The results of the first administration were examined and the preliminary standards were lowered by 3 Standard Errors of Measurement (SEM) to minimize false negative errors and to maintain consistency with the procedures used in setting the original cutscores.

The procedure described provides a practical, legally defensible, and technically sound approach to reevaluating cutscores for revised licensing tests. Other "field-accepted" standard setting methods may be equally adaptable to establishing standards for revised licensing tests, and research examining other approaches is currently underway. Regardless of which approach is employed, when criterion-referenced licensing tests are revised, the cutscores should be reassessed by expert judges in a manner that maintains a close link with the original test.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971

Baker v. Columbus Municipal Separate School District, 329 F. Supp. 706 (1971).

Ebel, R.L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Georgia Association of Educators v. Jack P. Nix, 407 F. Supp. 1102 (1976)

Jaeger, R.M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the meeting of the North Carolina Association for Research in Education, Chapel Hill, North Carolina, 1978.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

United States v. North Carolina, 400 F Supp. 343 (E.D.N.C. 1975), 425 F Supp. 789 (E.D.S.C. 1977).

Zieky, M.J., & Livingston, S.A. Manual for setting standards on the basic skills assessment tests. Princeton, N.J.: Educational Testing Service, 1977.