DOCUMENT RESUME

ED 471 654 TM 034 683

AUTHOR Tanner, David E.

TITLE To Err Is Human: Decision Errors and Candidates' Competency.
PUB DATE 2002-00-00
NOTE 13p.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; Decision Making; State Programs; Teacher Certification; Teacher Competencies; *Teacher Competency Testing; Teacher Evaluation

ABSTRACT

The competency testing of teacher candidates has become almost universal in the United States. This paper explores some of the issues associated with competency testing, especially as they relate to teacher testing in California. The practice has its origin in concern that some of those who choose to pursue teaching careers in the elementary and secondary schools may lack the competencies that effective teaching requires. Stated in terms of decision errors in scoring competency tests, the fear has been that some of those moving into full-time teaching may represent a population of false positives, teachers who lack necessary skills and abilities but who are certified nevertheless. The contingencies under which the competency test is actually administered in California, however, suggest greater worry over the possibility of false negatives, a concern that some of those who are denied certification may actually be qualified. (Contains 2 figures and 10 references.) (Author/SLD)

To err is human: Decision errors and candidates' competency

David E. Tanner, Ph.D

The Department of

Educational Research, Administration, and Foundations

California State University, Fresno

Abstract

The competency testing of teacher candidates has become almost universal in the United States. The practice has its origin in concern that some of those who choose to pursue teaching careers in the elementary and secondary schools may lack the competencies the effective teaching requires. Stated in terms of decision errors, the fear was that some of those moving into full-time teaching may represent a population of false positives, teachers who lack necessary skills and abilities but are certified nevertheless. The contingencies under which the competency test is actually administered in California, however, suggests greater worry over the possibility of false negatives, a concern that some of those who are denied certification may actually be qualified.

The ubiquitous element in college and university admissions procedures and in certification activities, is some sort of testing. Although employing test scores as a selection criterion is often criticized and major educational research organizations uniformly condemn using a test score as a sole selection criterion, test scores remain a fixture in many selection decisions. Many institutions, particularly colleges and universities, face the need to classify and screen large numbers of applicants, and do so with accuracy *and* efficiency.

The value of test data, of course, is in what they reveal about the level of a candidate's aptitude, ability, or command of some essential skill or knowledge. Because mental traits, abilities, and knowledge cannot be measured directly, quantifying them involves some risk of measurement error. Errors in mental measurement increase the potential for errors in the decisions that are made regarding who have the requisite level of the measured characteristic, who will most likely succeed in advanced study, or who will become the most successful candidates. The point of this paper is to examine some of the issues associated with decision error, particularly as they relate to teacher certification in California, although they generalize to many other selection and screening situations.

Error in Assessment

The point of collecting candidate data is to identify what is relevant to decisions about candidates' qualifications. This requires the predetermination of some sort of reference to which assessment data can be compared in order to reach a decision. When what the candidate must know or be able to do can be specified in discrete terms minimal competencies are defined. They may take the form of formal standards or some less formal set of criteria. The processes for determining them are the domain of standard setting.[1]

Particularly when it is difficult to determine the minimal level of some required ability (What level of reasoning is sufficient?), or when there are greater numbers of qualified applicants than there is opportunity to accommodate them, a normative reference is usually adopted. Norm-references specify as a required standard, some ranking within the group of all who fit a defined

category; 'the top 20% of high school seniors,' or 'applicants scoring in the upper quartile,' are examples.

Whether the standard is normative or criterion-based, data are gathered so that an informed decision can be made regarding which candidates qualify. Both approaches impose the risk of error, but the risk may be greater when objective criteria are used since the difficulty of accurate assessment of the candidate's ability can be compounded by any error that might occur in identifying the appropriate standard. In spite of the increasing popularity of assessment grounded in authentic outcomes, assessment procedures can rarely be truly comprehensive. Inferences must be drawn from a sample of candidates' responses regarding the larger universe of their skills, abilities, and knowledge. Furthermore, there may be some component of what is measured that has its origin in a characteristic other than the characteristic of interest. The students' command of the language becomes confused with their analytical ability. Their talent for calculation becomes a component of what is inadvertently termed problem-solving ability. Such occurrences increase the potential for decision errors when candidates are classified.

Components of a score that are irrelevant to the named construct reflect measurement error. Because problem-solving, for example, cannot be assessed the way one assesses height, or age, there is a great potential for measurement error. Lucky guesses on a multiple choice exam, a grader's scoring mistakes, a level of test-wiseness that allows test-takers to detect correct responses when they lack the construct which is the focus of the exercise, even the candidate's health can contribute to measurement error. These situations compound the effects of errors in measurement that might occur because the test items are of poor quality or insufficient quantity, because test-takers don't understand the directions, or because the data are tabulated improperly. Any of these factors may inflate, or deflate a score and in doing so, they may give rise to decision errors.

Decision Errors

In an ideal assessment world where there is no error, the population of competent candidates would be entirely separate from those who are incompetent (Figure 1a). This is actually the inference when a cut-off score separates those who are competent from those who are not. But because of measurement error, the scoring distributions of the competent and the incompetent can overlap, giving rise to decision errors.

Decision errors fall into two categories. False positive errors, also called alpha (α) errors or type I errors, occur when one who lacks the relevant characteristic is judged to possess it. Some of those judged to be competent are actually not competent—often those at the lower end of their distribution (Figure 1b). False negative errors, also called beta (β)or type II errors, occur when candidates who possess a necessary characteristic at the level required are judged instead to be wanting. Some of those judged to be not competent are actually competent—typically those at the upper end of that distribution.

Place Figures 1a and 1b About Here

While error inflates and deflates the scores of members of either group, only those scores nearest the cut-off score are of concern, since they represent the candidates who are most likely to be misjudged. In particular, errors that deflate the scores of those candidates in the lowest ranks of the competent and the errors that inflate the scores of those in the highest ranks of the incompetent provide the greatest probability of misclassification.

The Horns of a Dilemma

Part of the difficulty in making candidate decisions is that false positive and false negative classifications are inherently related. The number of false negatives can be minimized by lowering the cut-off score, but the companion will be more false positive classifications. Correspondingly, false positives can be minimized by raising the cut-score, but only by

increasing the likelihood of false negative decisions. Therefore, a critical element in standard-setting procedures, is deciding which of false positive or false negative errors pose the greater threat.

Choosing One's Error

Which of type I or type II error is more acceptable depends upon circumstances and is manifest in the rigor of the standard. Those judging the competency of prospective surgeons are more willing to accommodate type II than type I errors. The safety of prospective patients requires that standards be fixed relatively high. On the other hand, when landed immigrants seek naturalization as United States citizens they are given an oral examination of their understanding of U.S. government. Decision-makers may determine that the occasional false positive classification (the individual really does not quite understand the system) is preferable to denying someone citizenship who has a reasonable understanding of the political system, but does not speak English well enough to respond to all of the nuances of the questions.

In many other instances, the issues are less clear-cut and questions about which type of error is more palatable are grounded in politics as much as in measurement issues. But the preference for one type of error over another is implicit in the rigor of the standard. Teacher certification in California represents a case-in-point. Twenty years ago, much of the impetus for educational reform came from criticisms that teacher candidates are among the least academically talented of university students (Nelli, 1984; Weaver, 1981). In the wake of the charges, nearly every state adopted competency tests with particular attention on candidates' academic, rather than their pedagogical ability.

As long as concern about teacher competency continues to be a driving force behind the use of screening instruments (Barth, 2000; Hextall, Mahoney, and Menter, 2001; Song and Christiansen, 2001) the suggestion is that false negative errors are less objectionable. For the sake of the learner, it would seem better to occasionally exclude someone who may actually possess the required literacy skills, than provide a credential to someone who does not.

Besides being less objectionable from the standpoint of educational reform objectives, there is a compensation built in that allows the possibility of correcting a false negative decision: the test-taker may repeat the test. Besides allowing for a decision error to be rectified, it also allows a second chance opportunity to candidates who may have worked in the interim to develop the measured competencies. On retesting, candidates may be able to meet standards they failed to meet in an initial trial.

Although allowing repeated attempts to pass a competency test can correct for false negatives, what adjustment is made for the continuing possibility of a false positive? One option is to adjust the cut-off score with each repetition of the test. Millman (1989) explained that when the required score remains fixed, the risk of a false positive decision over successive administrations of the test accumulates to the point that, as repetitions increase, there is a very high risk of false positive classifications. In a practical demonstration of this potential, Huynh (1990) used high school exit exam data to show that multiple retakes with a fixed cut-off score reduce false negatives to near 0 and false positives to a probability of nearly 1 with repeated attempts at the test.

Although the very appearance of competency tests is witness to concerns that some of those who teach are not competent (that is, they represent false positive classification errors), a fixed cut-off score for the CBEST suggests that the original decision may have been modified. The criterion is the same for repeaters as it is for first time test-takers, in spite of the fact that repeaters have seen some version of the instrument and have greater familiarity with the testing procedure than those who are taking the test for the first time.[2]

Two other elements of CBEST procedure underscore this position. First: Only the failed portion need be repeated suggesting that although a false negative can be corrected by repeating the test, there is no parallel opportunity to correct for a false positive. Second: In spite of the fact that the three sub-tests are intended to provide independent measures of discrete skills, to some degree, higher-than-minimum scores in one area are allowed to compensate for scores below criterion in another.

Implications

Whether intended for those who aspire to teach, to the bar, or to graduation from high school, assessments provide the data from which classification decisions are made. Although our focus has been on testing data and classroom teacher certification, any type of measurement data is subject to error which places the resulting classification decisions at risk. Decision errors are related in that one type can only be minimized by increasing the potential for the other. The type that is less objectionable depends upon circumstances. When there is a substantial risk to some innocent party, for example, the tendency is to favor false negatives. However, the education reform movement suggests that there is sometimes a discontinuity between the rationale for the assessment and the mechanics of its implementation. Teacher competency tests were imposed because of concerns that teachers lacked basic literacy skills. Expressed in terms of decision errors, the concern was that some of those who become credentialed teachers represented false positive classifications and the point of the testing appears to have been to adjust selection procedures to be more tolerant of false negatives. But the conditions of test administration and data interpretation indicate a relatively greater sensitivity to false positive decisions.

At this point it seems unlikely that procedures will be adjusted to compensate for what are almost certainly significant numbers of false positive classifications, but such procedures are certainly available. Besides sliding the cut-off score, Millman (1989) has noted that it is common to average the repeat scores with the original score, or designate a "no-decision" band in the area where most decision errors are made.

In teacher candidates we have an assessment situation in which procedures have been adjusted to minimize false negatives. Although their discussion is beyond the scope of this paper, teacher shortages, and scoring patterns that correlate with the candidate's ethnic group appear to be the driving issues in those decisions. Competency testing may allow decision makers to determine whether candidates have the skills and understandings that reflect literacy,

but what appears to be an institutional bias favoring false positive decision errors may undermine the original intent.

Footnote

[1]The processes employed for determining a criterion which reflects the division between passing and failing is an important issue which is beyond the scope of this paper. For those with an interest in standard setting, papers by Geisinger (1991), Reid (1991), and Plake (1991) all in the same issue of *Educational Measurement: Issues and Practice* will be helpful.

[2]Test-takers have seized the moment. It isn't uncommon for those failing CBEST to take it several times.

References

Barth, P. (2000). Honor in the boxcar: Equalizing teacher quality. Washington, D.C.: ERIC Clearinghouse on Teaching and Teacher Education. ED 458350

Geisinger, K.F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10* (2), 17-22.

Hextall, I., Mahoney, P., and Menter, I. (2001). Just testing? An analysis of the implementation of "skills tests" for entry into the teaching profession in England. *Journal of Education for Teaching, 27*(3) 221-39.

Huynh, H. (1990). Error rates in competency testing when test retaking is permitted. *Journal of Educational Statistics, 15* (1), 39-52.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18* (6), 5-9.

Nelli, E. (1984). A research-based response to allegations that education students are academically inferior. *Action in Teacher Education, 6*(3), 73-80.

Plake, B.S. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10* (2), 15-16.

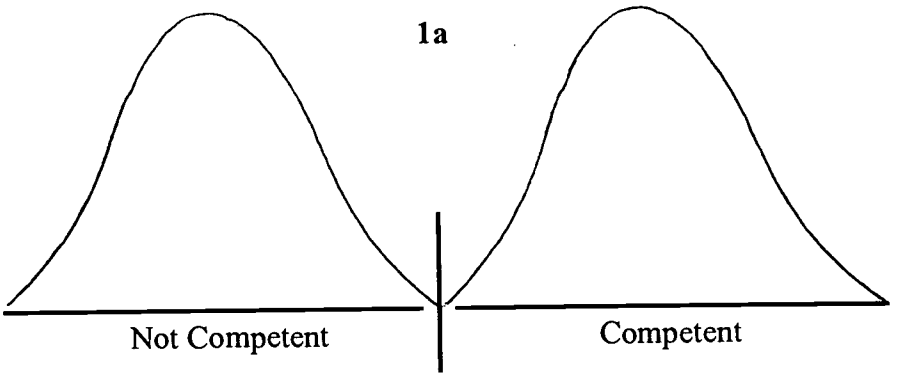
Reid, J.B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practices, 10* (2), 11-14.

Song, K.H. and Christiansen, F. (2001). Achievement gaps in pre-service teachers in urban settings. Washington, D.C.: ERIC Clearinghouse on Teaching and Teacher Education. ED 456187.

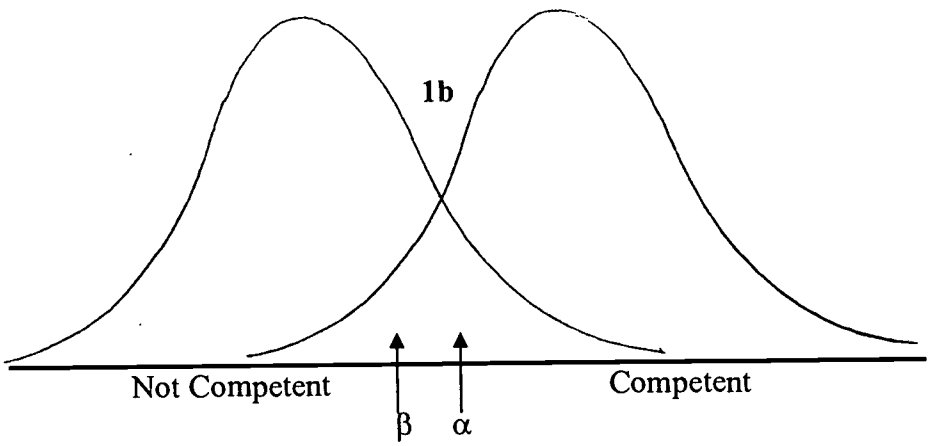
Weaver, T. (1981). The talent pool in teacher education. *Journal of Teacher Education, 32*, 32-36.

Figure 1a, 1b

The Competent and the Not Competent



The Cut-off Score



Cut-off Score

Type II Error Type I Error

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

REPRODUCTION RELEASE

(Specific Document)

TM034683

I. DOCUMENT IDENTIFICATION:

Title: To err is human: Decision errors and candidates' competency	
Author(s): David E. Tanner	
Corporate Source: California State University, Fresno	Publication Date: Nov., 02

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.

Signature:	Printed Name/Position/Title: D.E. Tanner, Professor
Organization/Address: MS #303, C.S.U., Fresno, Fresno, CA 93740-8025	Telephone: 559 278 0254 FAX: 278-0404
	E-Mail Address: davidt@csufresno.edu Date: 20 Nov., '02

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:

ERIC Clearinghouse on Assessment and Evaluation **1129 Shriver Laboratory (Bldg 075)** **College Park, Maryland 20742**	**Telephone: 301-405-7449** **Toll Free: 800-464-3742** **Fax: 301-405-8134** **ericae@ericae.net** **http://ericae.net**

EFF-088 (Rev. 2/2001)