DOCUMENT RESUME

ED 253 586 TM 850 132

AUTHOR Popham, W. James; Yalow, Elanna S.
TITLE Standard-Setting Options for Teacher Competency Tests.
PUB DATE Apr 84
NOTE 16p.; Paper presented at a joint session of the American Educational Research Association and the National Council on Measurement in Education (New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Beginning Teachers; Boards of Education; Competence; *Cutting Scores; Educational Policy; Elementary Secondary Education; *Minimum Competency Testing; *Public School Teachers; Screening Tests; Standards; Superintendents; Teacher Education
IDENTIFIERS National Teacher Examinations; *Standard Setting; *Teacher Competencies

ABSTRACT

Teacher competency tests are being used more frequently to assess the knowledge and skills of prospective teachers. Educational policymakers face the dilemma of setting passing standards for these tests which will satisfy the desire for meaningful quality standards for teachers while meeting the necessity for making available a reasonable number of teachers to staff the schools. Several standard-setting studies were carried out by IOX Assessment Associates. These studies were designed to assemble information for use by standard-setters such as boards of education or high level educational officials. Preference data and performance data are necessary to establish realistic standards. The methods of data collection and recommendations for use of the information are discussed. (DWH)

STANDARD-SETTING OPTIONS FOR TEACHER COMPETENCY TESTS*

by

W. James Popham
UCLA & IOX Assessment Associates

Elanna S. Yalow
IOX Assessment Associates

Educational policymakers are increasingly employing teacher competency tests as vehicles to assure an incredulous citizenry that public school teachers possess requisite knowledge and skills. Such tests are now being used both as screening examinations, that is, examinations which must be passed by prospective teachers in order to be admitted to teacher education programs, and exit examinations, that is, examinations which must be passed by teacher education graduates in order to be certificated.

In the rush toward reliance on teacher competency tests as quality assurance devices, however, a dilemma arising from the use of these devices has not been fully recognized. More specifically, the problems arising from the setting of passing standards for such important examinations have not been satisfactorily addressed. To illustrate, if passing standards are set high enough to placate the public, substantial numbers of prospective teachers, particularly minority candidates, will be barred from the teaching profession. If passing standards are set low enough to allow a resonable number of applicants to become teachers, the public may rightfully dismiss teacher competency tests as empty rituals.

---

*A symposium presentation at a joint session of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 23-28, 1984.

Given the delicacy of the dilemma faced by policymakers who wish to satisfy simultaneously (1) the desire for meaningful quality standards for teachers and (2) the necessity to make available a reasonable number of teachers to staff the schools, it is not surprising that special attention has been centered on the procedures whereby passing standards, that is, cut-scores, are set for teacher competency tests. In the following analysis an attempt will be made to isolate some of the key procedural choice points available in the standard-setting process, then offer recommendations specific to the setting of standards for teacher competency tests.

During recent years a number of first-rate analyses of the standard-setting enterprise have been offered (Shepard, 1980; Livingston and Zieky, 1982). Few of these, however, have focused specifically on the setting of standards for teacher competency tests. Given the particulars of the teacher competency assessment operation, there may be some factors that warrant different procedural choices for standard-setting on tests of teacher skills than for tests aimed at other examinee populations.

In our analysis we shall draw on our experience from several standard-setting studies carried out by IOX Assessment Associates (IOX). The first of these called for the setting of standards on the National Teacher Examinations (NTE) for the state of Kentucky (IOX Assessment Associates, 1983). The NTE was being considered in Kentucky as an exit examination to be completed at the close of a prospective teacher's preparation program. Over 600 expert panelists participated in the identification of recommended passing standards for the NTE during late 1982.

The second standard-setting study involved the establishment of passing standards for the Pre-Professional Skills Test (P-PST) during 1983 in Texas (Popham and Yalow, 1983). The P-PST was to be used as a screening examination to admit students to state-approved teacher preparation programs. Approximately 300 expert panelists and over 1,200 college students participated in the P-PST study.

The final standard-setting study that has informed our thinking on these issues was a recently concluded project carried out for the Charleston (South Carolina) County School District. In that study standards were set for a language skills test to be used with tenured teachers (Schaeffer and Collins, 1984).

During these three standard-setting endeavors we found ourselves faced with a number of choices. Having thought about those options at some length, we wish to share our conclusions with those conducting similar standard-setting procedures for teacher competency tests.

Data to Be Gathered

As we conceive of a standard-setting study, that descriptor is really a misnomer. Those conducting standard-setting studies rarely set standards. Rather, we assemble information that is designed to help those who ultimately do set the standards. Typically, these standard-setters will be Boards of Education (state or district) or high level educational officials (state or district superintendents). Nonetheless, for convenience, we characterize our efforts as standard-setting studies.

The first choice that designers of standard-setting studies must face is, "what kinds of information should be assembled for those who will ultimately make the standard-setting decision?" Generally speaking, there are two types of data potentially relevant to the deliberations of

standard-setters. The first of these consists of preference data, that is, the judgmentally rendered preferences of individuals regarding the number or percentage of items that should be answered correctly if an examinee is to pass the test. Such preference data can be collected from a variety of different constituencies, for example, teachers, students, citizens, etc. The second kind of information of potential relevance to the deliberations of standard-setters is performance data, that is, the actual performance of examinees on the test's items. Performance data are often gathered during field-tests of an examination on which performance standards are to be set, such field-tests being conducted to gauge the psychometric quality of the examination's items.

In the schemes proposed by several writers, e.g., Popham, 1981, the suggestion is usually proffered that both preference data and performance data be given standard-setters. After all, why not let standard-setters see not only various judges' recommended passing standards, but also the actual test performance of examinees? We'd like to review the wisdom of that suggestion.

It is our experience that, in general, passing standards based on performance data (such as the mean performance of field-test examinees) will be lower than the passing standards recommended by experts. What role, then, should the lower performance-based estimates play in policymakers' deliberations regarding standards?

We believe that there are two major considerations which should influence the extent to which the typically lower performance data should influence standard-setters to set reduced cut-scores that, if based only on preference data, would most likely be higher.

Gravity of mistakes. First, there is the magnitude of false-positive consequences. If there is substantial danger associated with passing an individual who doesn't possess the necessary skills, such as would be the case with competency tests for airline pilots or brain surgeons, then the significance of performance data should be minimized. If expert aviators have recommended a passing standard for prospective pilots of 95 percent correct on a Foggy Landing Emergency Procedures Test, we would not wish standard-setters to relax that standard merely because an inept group of prospective pilots scored substantially lower, hence yield performance data which would incline us to opt for a lower cut-score.

Confidence considerations. The second factor influencing the extent to which we should rely on performance data in our standard-setting decisions is based on the confidence placed in the preferential data. There are two major considerations leading to the confidence we can have in the preference data. First, there is the caliber of the individuals rendering preferences. The more expert that these individuals are, the more confidence we have in their preferences. If, for example, we were having professors of pedagogy render their opinions regarding appropriate performance standards, we would doubtlessly have more confidence if those professors possessed years of experience than if they were first-year, fresh out of graduate school professors. Second, there is the subject matter itself. How confident are we that there is a sufficiently solid knowledge base in the subject matter at hand that experts, no matter what their degree of expertise, can arrive at defensible preferences? For example, we undoubtedly believe that the subject matter associated with plumbing is better understood than the

subject matter associated with clairvoyance. Thus, we would be less apt to use performance data as a check if the task were to determine passing standards for a test of one's plumbing proficiency than we would be if we were creating a clairvoyance competency test.

In the case of teacher competency tests, we believe that the chief factor in reaching a defensible passing standard should be the judgments of knowledgeable experts. Performance data should play a meaningful, but lesser role in determining the standards. We take this position in view of the two factors described earlier, that is, (1) the magnitude of the false-positive consequences and (2) the confidence placed in the preferential data we might collect.

We believe that the certification of a teacher as competent who doesn't actually possess the tested subject matter knowledge or basic skills does not constitute an error of monumental magnitude. There are, fortunately, other mechanisms for checking on the teacher's capability to deliver effective instruction, e.g., supervisorial judgments. Thus, we need not rely exclusively on experts' recommendations, and can be guided to some extent by performance data, particularly insofar as such information can prove useful in helping us control the numbers of individuals entering the teaching field.

Regarding the second consideration, that is, the confidence we can place in preferential data regarding passing standards for teacher competency tests, we are uneasy. It appears to us that the knowledge base associated with most teacher competency tests is too fragile or, perhaps more kindly, too unclear, to warrant all that much confidence in judges' standards-preferences. Thus, again we find ourselves leaning toward performance data as a check on judges' standards-preferences.

Yet, having concluded that performance data should be used to monitor the judges' advice, there is something almost unprofessional about discounting the judgments of educational experts regarding "how good is good enough." If we allow the current test performances of teachers to govern our passing standards, then today's status quo is apt to become tomorrow's. Hence, with some misgivings, we still recommend that in the setting of passing standards on teacher competency tests, we weight preference data from qualified experts much more heavily than performance data.

Preference Data Options

In the gathering of preference data, we encounter a number of choices which must be resolved as we set up our data-gathering procedures. These choices involve decisions regarding (1) the individuals to be included as judges, (2) the materials we will supply to the judges, (3) the circumstances in which judgments will be rendered, and (4) the nature of the data-gathering questions posed to the judges. We shall address each of these briefly.

Judge selection. Standard-setters, in our experience, are anxious to receive recommendations regarding performance standards from all relevant constituencies. Because the standard-setters can, with relative impunity, disregard such recommendations, they have little to lose by receiving them. In a sense, the more recommendations that are supplied, the better. Dissimilar recommendations allow standard-setters to rely on the advice of those whose opinions they most value.

In the case of teacher competency tests, the most likely judges are teachers, citizens, and higher education representatives. The higher education group can be further subdivided into professors, education

administrators (e.g., education department chairs), and college-wide administrators (e.g., presidents). We have employed all five of these groups in our own standard-setting studies. Although their recommendations are often similar, there are occasionally meaningful differences between, for example, the passing standards recommended by teachers and by citizens. The citizens, predictably, often opt for higher standards than the teachers.

We recommend that the entire array of these judges, if economically feasible, be involved in the standard-setting activity. If any of those groups must be excluded because of cost considerations, we favor jettisoning the education administrators and college-wide administrators.

Materials supplied. What sorts of materials should judges review in order to render recommendations regarding appropriate passing standards? The options that usually come to mind are these:

1. The actual test items
2. Descriptions of the test items (plus illustrative sample items)
3. Performance data from examinees who have taken the test
4. Recommendations of others

We favor having judges review the actual test items. This requires the creation of a security-monitored environment where officials can make certain that the security of the items being reviewed is not compromised. The logistics of setting up such a review session are not all that simple. Frequently, these kinds of sessions involve travel costs for judges and, because the judges often are called on to review many items, an honorarium as well. Nonetheless, we have found no legitimate substitute for having judges actually ponder the test items themselves in the course of making their cut-score recommendations.

One procedure that has often been employed in such in-person standard-setting conclaves is the iterative approach devised by Jaeger (1981). Having used Jaeger's modified delphi procedure in another context, we are not enthralled with its virtues. In essence, Jaeger's approach provides judges with incrementally increased information relevant to the setting of standards, not the least of which are the preferences of other judges who are participating in the standard-recommending process. Our experience with Jaeger's method is that the approach tends to reduce heterogeneity of recommendations, but not bring about drastic changes in judges' mean preferences. Assuming that the judges in Jaeger's approach are offering recommendations to the actual standard-setters, we think that those standard-setters would be better served to see the unadulterated, if divergent, preferences of judges rather than the homogenized product of an exercise in social psychology.

As a <u>supplemental</u> source of preference data, we often employ a mail-out data-gathering instrument in which judges receive description of what the test items measure, one or more sample items, then are asked to recommend a "percent correct" necessary to pass the test described (and illustrated). The virtue of this mail-out procedure is that it is inexpensive and, thus, can be used with large numbers of individuals interested in offering their advice regarding the passing standards. Indeed, as a political vehicle for involving many constituencies in the standard-setting process, it is highly effective. On the negative side, however, one is less apt to trust the recommendations of judges who are basing their views only on descriptions of items and not on the items themselves. Nevertheless, we favor using mail-out requests for cut-score recommendations as a source of additional data which may prove useful to the ultimate standard-setters.

Our views are mixed about making performance data available to judges prior to asking for their cut-score recommendations. As suggested earlier, we are unwilling to rely too heavily on performance data (as opposed to preference data) in the setting of standards for teacher competency tests. Yet, we would suggest making such data available to judges, but only after they had first rendered recommendations based on the items themselves (or, in the case of mail-out materials, descriptions of items).

We favor a final, just-before-the-decision-is-made, gathering of preference data from a set of individuals we call "standards advisors." We supply these individuals, by mail, with preference data of others plus performance data, then ask the standards advisors to offer their best counsel to the ultimate standard-setters. This approach was used in the IOX standard-setting study for the P-PST in Texas (Yalow and Popham, 1983b) and in Charleston, South Carolina (Schaeffer and Collins, 1984). In Texas, for example, we mailed a 15-page standards advisor's booklet containing all of the needed information to Texas (1) college presidents, (2) education deans, (3) local school board members, and (4) school administrators. Responses were secured from all four groups, then summarized, by group, for members of the Texas State Board of Education who actually set standards for the P-PST. The standards advisor strategy presents to advisors all of the information that will subsequently be available to standard setters, then asks the advisors to use their best judgment in rendering a cut-score recommendation. Standard setters can see what others, given the array of available information, would set as a passing standard. In addition, the standards-advisors data-gathering strategy is relatively inexpensive.

Data-collection setting. As suggested earlier, we favor an in-person versus by-mail securing of recommendations from judges because judges can thereby scrutinize the items themselves. Ideally, both in-person and by-mail data can be gathered. However, if a choice must be made, we definitely recommend that fewer judges be used in person than more judges via mail.

The charge to judges. The nature of the data-gathering inquiry to judges is, obviously, pivotal. Standard setters can go to elaborate lengths to get the proper judge and the proper setting, yet ask flawed questions and, thereby, botch the whole exterprise.

For in-person judges, we recommend that item-by-item judgments as well as total-test judgments be secured. The phrasing of the specific questions for judges is critical. In our Texas study of the P-PST, for example, we agonized at length over subtle shadings of meaning before settling on this item-by-item charge to judges: "Should a student be required to know the answer to the item in order to be admitted to a teacher preparation program in Texas?" (Yalow and Popham, 1983a). It is doubtful whether the question's apparent simplicity reflects the effort expended in its formulation. Formulation of the judge's charge is more critical than usually recognized, hence should be given suitable attention by those conducting standard-setting studies.

To illustrate, the relationship between the performance standards question and questions regarding the test's content validity are often not considered. Yalow (1983) has explored that relationship:

> At first glance, the content validity of a test and performance (passing) standards for it seem to be clearly distinct concepts. Upon closer inspection, however, it becomes extremely difficult to disentangle them. Consider, for example, a study designed to determine both the content

validity of a teacher-certification examination and to establish performance-standards for that examination. If respondents were asked to make item-by-item judgments regarding the test's content validity, they might be asked to indicate something, such as whether the content of each item or the skill measured by the item was "necessary for competent performance as a teacher." Now consider the corresponding performance standard question. This question might be posed as, "Should a teacher-certification candidate be required to answer this item correctly in order to receive a teaching certificate?" These two questions are clearly related. For if the content or skill measured by an item is important for an individual in a profession (content valid), should those individuals not be required to answer the item correctly before being admitted to that profession? And if the content or skill of an item is not important, is it legally defensible to require a correct response to that item on a certification examination? Is it appropriate for individuals making performance standards estimates to base those estimates on normative judgments regarding what examinees should know, or must those judgments be limited to whether examinees need know the content in order to function acceptably?

One key difference between the content validity and performance standard concerns may be that the content validity question should focus on the nature of the content or skill measured by the item; the performance standard question should focus on the actual difficulty of the item. But this is a delicate distinction to convey to raters, for it may be extremely hard to distinguish between the general content or skill and the specific item used to measure it.

In review, then, we believe that the choice points regarding the acquisition of preference data are numerous. Given the significance which we believe should be associated with judges' recommendations regarding teacher competency tests, we think it unlikely that those conducting standard-setting studies can devote too much attention to making these choices wisely.

Performance Data Options

Turning briefly to the matter of performance data, we see fewer significant options, although there are certainly some choices to be made in how we gather performance data for purposes of informing the decisions of standard-setters.

In most instances the circumstances under which performance data can be gathered prior to the setting of standards will be dictated by practical contingencies. Frequently, for example, such performance data are gathered as a consequence of a field-test of items for an under-construction test. Clearly, in such circumstances one is obliged to take what can be had in the way of data.

In other cases, as in the P-PST study in Texas, a special administration of the test (largely for standard-setting purposes) was carried out. Because such separate (for standard-setting) administrations are typically quite costly, we recommend that they be modest in magnitude. After all, the actual circumstances of the "real" testing can rarely be replicated, hence one is always cognizant of motivational differences in examinees that might have led to atypical performances. Performance data plucked from situations different than those where the test scores "really count" should invariably be taken with several grains of salt. Thus, we do not favor major allocation of resources to the collection of data under circumstances which are not identical to the circumstances when test results make a real difference to examinees.

Information and Informed Judgment

As has been noted by numerous writers regarding the setting of standards, the process eventually boils down to a matter of human judgment. We think it unlikely, given the highly ideosyncratic nature of the circumstances in which standards must be set for teacher competency tests, that uniform procedures will or should evolve.

Rather, it is the responsibility of those conducting standard-setting operations to assemble the most pertinent data available for the

particular situation at hand. That requires a context-dependent review of such choice points as we have enumerated earlier. For the kinds of standard-setting studies we have recently conducted, the foregoing recommendations hold. Given different circumstances, all bets are off.

References

Jaeger, R. M., An iterative structured judgment process for establishing standards on competency tests: Theory and application, Educational Evaluation and Policy Analysis, 1982, 4, 461-476.

Livingston, S. A., and Zieky, M. J., Passing Scores, Princeton, N.J.: Educational Testing Service, 1982.

IOX Assessment Associates, Appraising the National Teacher Examinations for the State of Kentucky, Culver City, Calif.: Author, 1983.

Popham, W. J., Modern Educational Measurement, Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Popham, W. J., and Yalow, E. S., Appraising the Pre-Professional Skills Test for the State of Texas: Final Report, Culver City, Calif.: IOX Assessment Associates, 1983.

Schaeffer, G. A., and J. L. Collins, Setting standards for high stakes tests, a paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 23-26, 1984.

Shepard, L. A., Standard setting issues and methods, Applied Psychological Measurement, 1980, 4, 447-467.

Yalow, E. S., Content validity discontent, a paper presented at the annual meeting of the American Educational Research Association, Montreal, April 11-15, 1983.

Yalow, E. S., and Popham, W. J., Appraising the Pre-Professional Skills Test for the State of Texas: Report No. 1, Culver City, Calif.: IOX Assessment Associates, 1983a.

Yalow, E. S., and Popham, W. J., Appraising the Pre-Professional Skills Test for the State of Texas: Report No. 5, Culver City, Calif.: IOX Assessment Associates, 1983b.