1990

1. Face Validity: Siren Song for Teacher Testers

W. James Popham
UCLA and IOX Assessment Associates

1

Face Validity: Siren Song for Teacher Testers

W. James Popham
UCLA and IOX Assessment Associates

The sirens of Greek mythology were a seductive set of women who, by singing melodies that apparently topped even those of Diana Ross and the Supremes, could lure mesmerized men to their doom. Greek mythology, it is clear, was solidly sexist, for the sirens used their supernatural singing talents to entice only unsuspecting males into trouble. Gender-equity considerations were conspicuously absent from the forays of Greek fablemakers. Sexism aside, however, it is certain that the sirens of yesteryear knew how to sing some truly enticing tunes.

FACE VALIDITY'S ALLURE

In today's current frenzy to develop teacher assessment devices that tap truly important dimensions of a teacher's skills, astute observers will recognize a melody subtly reminiscent of the an-

cient sirens' top 10 hits. The seductive refrain to which I refer is *face validity* or, as it can be more pedantically described, *verisimilitude.* With ever-increasing frequency, the architects of teacher tests are striving to create assessment instruments that simply reek of face validity, that is, assessment approaches consonant with the actual day-to-day requirements of teaching. Face validity is being touted by some as a genuinely indispensible element of new, more defensible teacher assessment devices. Paper-and-pencil tests, particularly those of the multiple-choice genre, are regarded by these new face-validity enthusiasts as assessment tools of a benighted past in which teacher tests yielded inferences of only debatable validity.

Many educators' experience with multiple-choice teacher tests has been based on the *National Teacher Examinations* (NTE) developed by the educational Testing Service (ETS). NTE tests have been available for some time. They deal not only with general knowledge and pedagogy, but also with an array of special subject fields such as chemistry and French. Originally constructed to assess the consequences of teacher preparation programs, NTE tests have recently been used in various states as part of licensure systems for prospective teachers. One suspects that when critics disparage multiple-choice teacher tests, they are generally thinking of the sorts of examinations that they imagine the NTE to represent.[1]

As indicated earlier, a dominant reason that today's teacher testers are scurrying from multiple-choice teacher tests is that examinees' responses to such tests do not resemble what goes on in teachers' classrooms. Instead, a new cadre of teacher testers is currently striving to create assessment approaches simply swimming in face validity, that is, assessment approaches unquestionably parallel to the activities in which classroom teachers must engage.

Webster's Dictionary defines an object possessing verisimilitude as one "having the appearance of truth." Therein, of course, is the attraction of face validity. It *looks* so appropriate. The appeal of verisimilitude in testing is compelling. But is that appeal warranted in the case of teacher testing?

---

[1]There are, to be sure, concerns abut NTE-type tests other than their lack of face validity. Some critics contend that such paper-and-pencil tests must be replaced or augmented with performance tests.

LICENSURE AND CERTIFICATION

Prior to looking more carefully at the pros and cons of face validity, we need to consider the various kinds of teacher tests to which face validity may be germane. Before doing so, however, it will be useful to engage in a bit of preliminary term tidying, for there is the potential for substantial confusion in the way that educators employ two key terms, namely, *licensure* and *certification.* Historically, many states have awarded teaching certificates to prospective teachers at the close of their teacher-education programs. Thus, the use of a test in conjunction with this process would typically lead us to describe such a test as a "teacher-certification" test.

Yet, in recent months the efforts of Lee Shulman and his Carnegie-supported associates (Shulman & Sykes, 1986) to devise what they refer to as "certification" tests has forced us to be more circumspect in using teacher-test descriptors.[2] Shulman employed the expression "certification test" to describe a test used with experienced, incumbent teachers. His use of the adjective *certification* coincides with the idea of a *certified* public accountant, this is, a professionally sanctioned, superior accountant. (Not all accountants, of course, are certified.) A certified teacher, to Shulman and his colleagues, is an incumbent teacher who has demonstrated clearly superior competence. Only a modest proportion of American teachers would, therefore, achieve such a state of certified excellence. Shulman would prefer to describe end-of-teacher-training tests as *licensure* tests.

Given the attention that Shulman's work is receiving these days, it seems that his licensure/certification distinction is apt to be used with increasing frequency in the field, hence, it is the distinction that will be employed in the remainder of this chapter. In other words, a *licensure* test will refer to tests given to prospective teachers at the end of their training programs. In contrast, a *certification* test will refer to tests given to incumbent teachers who aspire to be recognized for their advanced level of competence. (It should be noted, however, that in states where teachers have traditionally received certificates to teach, it may be difficult to persuade local educators to adopt the descriptor licensure test.)

---

[2]Shimberg (1981) drew this distinction between licensure and certification tests some years ago. See also Murray's (1986) essay dealing, in part, with this distinction.

DISTINCTIVE FUNCTIONS OF TEACHER TESTS

At the present, there are numerous varieties of teacher tests used in the U.S. To equate these diverse tests would be akin to considering a Lear jet and a San Francisco trolley as equivalent modes of transportation. We need to do some sorting out of the various species of teacher tests in order to decide which, if any, need to be face valid.

We can distinguish among teacher tests most conveniently by considering their functions. Let's look, therefore, at six relatively distinctive functions served by today's teacher tests.

Teacher-Education Screening Tests

One function of a teacher test is to screen applicants for admission to teacher-training programs. Technically, this use of the phrase "teacher test" is inaccurate. Clearly, if a test is being used to determine whether or not students are admitted to a teacher-education program, those students are not yet teachers. However, because the examinees have clearly set out in pursuit of teacherhood, it seems only a mild misnomer to consider such tests as members of the teacher-testing family. An example of such a teacher-education screening test would be the *Pre-Professional Skills Tests* (P-PST), tests of reading, writing, and mathematics distributed by ETS. Teacher-education screening tests characteristically focus on such subject matter. These sorts of screening tests are designed to determine whether an examinee is sufficiently literate to perform satisfactorily in a teacher-education program and, if successful in that program, thereafter as a classroom teacher.

Teacher-Licensure Tests

A second variety of teacher test is one used at the close of a teacher-education program to determine if examinees possess sufficient knowledge and/or skills to be granted a teaching license. As noted earlier, this is a function currently served by the NTE in many states. Generally, the focus of teacher licensure tests is on the examinee's mastery of a subject field or, perhaps, pedagogy. The content of teacher-licensure tests, as is the case with all varieties of teacher tests, is determined on a state-by-state basis. In California, for example, the *California Basic Education Skills Test* (CBEST)

covers the same reading, writing, and mathematics content as the P-PST (the test from which the CBEST was originally derived). Thus, it would seem, teacher-licensure tests can cover the full gamut of content tapped by teacher tests, that is, subject matter, pedagogy, and basic skills.

In some instances it is useful to think of teacher-licensure tests as "initial" licensure tests because such tests may be *provisional* or *permanent* depending on the regulations of the particular state involved. In Connecticut, for example, the NTE are used at the close of an examinee's teacher training in order to grant an initial certificate to teach, a certificate that must be renewed by the state within 2 years. Other states grant more "permanent" licenses to teachers on the basis of an end-of-training licensure test.

Confirmatory Teacher-Licensure Tests

A third type of teacher tests is one employed to confer a permanent teaching license on those who have previously been only provisionally licensed to teach. In Connecticut, as previously indicated, a confirmatory licensure test will be administered to provisionally licensed teachers during their 1st or 2nd year of teaching.

Not all states issuing provisional teacher licenses rely on a formal teacher test to confer permanent licenses. In most states this function is accomplished chiefly by in-class observations and/or administrator' judgments regarding the neophyte teacher's competence. However, when a formal test is used as part of a process to confirm a teacher's provisional license, that test is clearly a distinctive species of teacher test.

In the main, confirmatory licensure tests deal with pedagogical content as opposed to subject matter content or the 3Rs. Typically, confirmatory licensure tests are used in concert with other indices of a teacher's skill, for example, classroom observations.

Career-Ladder Teacher Tests

A fourth function of teacher tests arises in connection with the educator career-ladder systems that have been established by an increasing number of our states. In Tennessee, for instance, in addition to relying on a host of other evaluative data, those who evaluate career-ladder candidates use teachers' scores on (a) a basic literacy test and (b) a test of pedagogical knowledge dealing with curriculum, instruction, and evaluation.

Teacher-Certification Tests

As indicated earlier, Shulman and his associates have set out to develop assessment prototypes suitable for use by a national teacher-certification board bent on bestowing special recognition on superior teachers. Shulman's approach rests on the assumption that many key teaching acts represent a teacher's pedagogically appropriate use of suitable subject matter (Shulman, 1986). Accordingly, the tests to be developed by Shulman and his colleagues seem destined to assess admixtures of both subject matter content and pedagogy. Shulman's group is eager to devise assessment schemes that are fundamentally different from conventional paper-and-pencil teacher tests. He hopes to rely far more heavily on the measurement of actual or simulated performance than is the case with most extant teacher tests.

Teacher-certification tests are at a particularly early stage of development. Preliminary assessment ploys devised by Shulman et al. will doubtlessly need to be revised, based on numerous tryouts, before they are perfected. New and better assessment tools are difficult to build. Nonetheless, a programmatic effort has been initiated to create certification tests for American teachers. The nation's educators will view with interest the endeavors of Shulman and his cohorts.

Teacher-Relicensure Tests

The final variant of teacher tests has been installed, thus far, in only three states, namely, Arkansas, Texas, and Georgia. Regulations in these three states oblige incumbent teachers to pass state-mandated tests as a condition for license renewal. Because incumbent teachers, as a consequence of results on such tests, could be excluded from teaching, these relicensure tests have received substantial media attention.

In Arkansas, the teacher-relicensure test deals with basic skills reading, writing, and mathematics. In Texas, the teacher-relicensure test measures examinees' mastery of rudimentary reading and writing skills. In Georgia, teachers are obliged to pass subject-matter oriented teacher-relicensure tests.

It is unclear whether additional states will require incumbent teachers to demonstrate mastery of basic skills as a condition for relicensure. In Arkansas and Texas, a predictably small percentage of teachers were denied license renewal as a consequence of the

relicensure tests. That small percentage of teachers, however, interacted with many thousands of pupils each year. Arkansas and Texas policy makers conceived of the teacher-licensure test as a literacy test, that is, a test designed to identify teachers who were insufficiently literate to function as a classroom teacher. In Georgia, state policy makers were more concerned with teachers' mastery of subject matter.

A half-dozen varieties of teacher tests have now been described. Some of them deal with subject matter, some with pedagogy, and some with basic skills. For certain teacher-testing functions it is possible to develop tests that assess the *interaction* between certain pedagogical skills (e.g., task analysis) and particular subject content.

Although there are similarities among the six kinds of tests discussed, the functions served by each are meaningfully different. In Table 1.1 we see the six varieties of teacher tests alongside the types of content that has been used for such tests (● = already in place) or that could appropriately be used for such tests (○ = potential).

Let's turn now to the notion of face validity and the degree to which it is relevant to different types of teacher tests.

TABLE 1.1
Six Varieties of Teacher Tests
and Present/Potential Appropriate Content

	Appropriate Content			
Function	*Basic Skills*	*Subject Matter*	*Pedagogy*	*Pedagogy Subject Interactions*
Teacher education screening test	●	○	–	–
Teacher licensure test	●	●	●	○
Confirmatory teacher licensure test	–	–	○	○
Career-ladder teacher test	●	○	●	○
Teacher certification test	–	○	○	○
Teacher relicensure test	●	○	○	○

Note. ● = Already in place, ○ = Potential, – = Inappropriate

WHAT FACE VALIDITY ISN'T

Face validity is not regarded as a bona fide member of validity's blessed trinity. In the 1985 revision of the *Standards for Educational and Psychological Testing* (American Educational Research Association American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1985), as in the 1974 and earlier version of the *Standards,* we find no endorsement of face validity. The 1985 *Standards* accurately reflects an increasingly broad consensus in the measurement community that *validity* refers to the defensibility of an inference that is based on an examinee's test score. Thus, the 1985 *Standards* recommends that, to gauge the validity of score-based inferences, we assemble *content-related, criterion-related,* and *construct-related* evidence of validity. Face validity (or even face-related evidence) is not touted in the *Standards* as a fourth form of evidence pertinent to score-based inferences because it bears *no necessary relationship* to the validity of inferences we draw from an examinee's score. A test can yield valid score-based inferences even if it doesn't possess face validity. A test can yield invalid score-based inferences even if it does possess face validity.

Face validity constitutes the perceived legitimacy of a test for the use to which it is being put. Thus, if a teacher test is seen as appropriate for its particular function, it is said to possess face validity. Perceptions regarding the legitimacy of a test are held by a variety of individuals, not the least of whom are the examinees who are obliged to take the test. In a sense, therefore, face validity of teacher tests is a political rather than a psychometric consideration.

Today's teacher testers yearn for face-valid tests because they wish those tests to be perceived as appropriate by the numerous constituencies that might have a say regarding whether and how the tests are used. If state legislators enact laws calling for the creation of a teacher test, those legislators want a test that they and the public perceive to be a sensible approach to the testing of teachers. Such legislators, for example, would view with dismay a teacher test composed of ink-blot stimuli about which examinees were required to create fantasies. People simply don't see the relevance of such assessment approaches to the process of teaching.

As far as most of us are concerned, then, teacher tests ought to *look* like teacher tests. Teacher tests ought to look like teacher tests, however, only if other things are equal. And that, of course, is the nub of the problem. Those teacher testers who worship at the face-validity altar may fail to recognize that without other evi-

dence of a score-based inference's validity, face validity creates a potentially false impression of a test's appropriateness. For such misguided teacher testers, face validity may not only be seen as necessary, it may be regarded as sufficient for the defensible testing of teachers. As will be demonstrated, that view is in error.

VALID INFERENCES AND FACE VALIDITY

Earlier it was indicated that face validity and the validity of a test's score-based inferences were not necessarily linked. Let's illustrate that point with a few examples drawn from existing teacher tests.

Consider an assessment approach drawn from a recently installed career-ladder teacher test in which applicants to a state's career-ladder program were obliged to assemble a portfolio of materials representative of certain aspects of their educational efforts. The portfolio was to include lesson plans, teacher-made tests, teacher-developed practice exercises, and so on. Because this portfolio assessment scheme dealt chiefly with teacher-generated materials of obvious relevance to instruction, it was generally regarded by outsiders as face valid. It was not widely known, however, that the state's teachers union had pressured career-ladder officials into making public the detailed scoring criteria that were to be used when the portfolios were being judged. As a consequence of the state's publicizing the portfolio-judging criteria, all career-ladder applicants, thereafter, assembled portfolios certain to earn the maximum possible points. Portfolios were created by assiduously attending to union-supplied guidelines and models. The result was, predictably, a flock of highest possible scores from which no valid inferences could be drawn regarding teachers' ability to generate such materials. A face-valid teacher assessment approach had been modified so that no sensible inferences could be drawn from examinee's performances.

This portfolio-judging debacle should not be used as evidence that teachers should not be informed of the criteria by which their performances will be judged. Teachers have a right to know the standards that will be applied to their test results. For instance, if teachers are asked to generate a brief written composition, it is perfectly sensible to let them know (in advance of the actual test) what criteria will be employed to judge the compositions. The difference between these two examination procedures is pivotal. In the case of the portfolio test, examinees can prepare their re-

sponses (with consultant help, if desired) in advance of providing those responses (the portfolios) to the examiners. In the case of the composition test, examinees must demonstrate in security monitored conditions that they can adhere to specified criteria by applying those criteria when generating an *original* response. If divulging of examination criteria for a teacher test distorts the defensibility of inferences we wish to draw from examinee's performances, then we ought not employ that teacher-testing strategy.

In another state we currently find a form of teacher performance test employed as part of a confirmatory teacher-licensure system. The state's beginning teachers are observed in their classrooms on several occasions to discern whether sound instructional practices are being employed. As far as the public is concerned, the assessment procedure is perceived to be legitimate. Teachers are being judged as they carry out classroom teaching responsibilities. Yet, because in all instances the observations are scheduled well in advance, that is, the teacher is aware of the specific date and time when observers will be present, we find nearly all teachers deliver polished lessons, many of which have been rehearsed at length with the aid of videotape recordings, consultants' reactions, and so on. The amount of performance variation is trivial. Almost everyone wins. Inferences about a teacher's actual classroom performance based on this face-valid approach are of little value.

The focus of validity, as indicated previously, must be on the defensibility of the score-based inferences that we attempt to draw from examinee's performances. Typically, in the case of teacher tests, we administer tests so that we can make inferences about how a teacher is apt to behave in an instructional setting. We use the examinee's score on the test as a proxy to represent aspects of future classroom performance. Even if a test focuses exclusively on a teacher's subject-matter knowledge, hence, we are more concerned with appropriateness of test's content, we still infer that the more knowledgeable teacher will dish out better content in the classroom. If, for a particular teacher test, we have reason to believe that the inferences we draw about teachers' classroom performances are not warranted, then the teacher test is of no utility irrespective of whether it possesses face validity.

Teacher tests can yield valid inferences even though they do not possess face validity. For example, biographical information often proves potent as a predictor of one's future success in many settings. Yet, even if biographical information yielded yummy predictions of a teacher candidate's future classroom performance, it

would not possess face validity. Nonetheless, if an administrator is choosing among prospective teachers and has no other predictor available, the use of biographical information is apt to yield better decisions than a table of random numbers.

To repeat, the presence or absence of face validity bears little, if any, relationship to the defensibility of score-based inferences we make when we use tests to make decisions about examinees. Does this mean, then, that face validity is unimportant? The answer to that question is emphatically negative.

FACE VALIDITY'S IMPORT

As previously noted, face validity constitutes the perceived legitimacy of a test. Perceptions of legitimacy on the part of those concerned with a test's use are important. Take, for example, the teacher relicensure tests used in Arkansas and Texas. It was sufficiently traumatic for the teachers in those states to face the loss of their teaching licenses because of unsatisfactory results on a basic skills test, imagine how much more stress they would have experienced if the test itself appeared to be educationally irrelevant.

In Arkansas, for example, test items in the mathematics, reading, and writing sections of the Arkansas Educational Skills Assessment (AESA) were all couched in educationally relevant contexts. For instance, mathematics word problems dealt with the sorts of activities in which teachers typically engage such as calculating test-score averages or managing classroom materials budgets. One effect of this attention to the face validity of the AESA was at least a partial reduction in the strident, union-spurred resistance to the test. Had the AESA not been perceived as a legitimate measure of the basic skills needed by teachers, then the Arkansas Education Association would have been more effectively able to galvanize teacher resistance to the test.

Similarly, if the legislators of Arkansas who had mandated the teacher-licensure test perceived the AESA to be irrelevant to the requirements of teaching, then they would certainly have been less willing to support the AESA when it was under fire.

Over 40 years ago, Mosier (1947) argued that, when possible, tests which possessed face validity would be decisively superior to those which did not:

> In Civil Service situations, the candidate whose score is less than he expected is inclined to attribute his low score, not to his own defi-

> ciencies but to the impractical nature of the test in relation to the job for which he is being examined. His dissatisfaction with the test results and his feeling of injustice may, of course, have real merit. We have not yet reached the era of public personnel examining where all tests are technically sound. Whether or not there is merit in his claim, the legislature, the courts, and public opinion, the court of last appeal, are more readily impressed by superficial appearances than by correlation coefficients. It becomes highly important, therefore, that a test to be used in such a situation not only be valid in the pragmatic sense of affording reasonably accurate predictions of job competence, but *have the appearance of validity* as well.
>
> This appearance of validity as an added attribute is important in terms of the acceptance of the test, not only by the persons being examined, but also by those operating officials who are charged with the responsibility for taking action based upon the test results. If sound tests are given and accurately reported, but the supervisor, interviewer, or counselor has no confidence in them, the results will not be used effectively. (p. 200)

Clearly, face validity for teacher tests is always a plus. If teacher tests can yield valid inferences *and* also possess face validity, they are certainly apt to function more satisfactorily (for any of the six functions cited earlier) than tests that yield only valid inferences but possess no face validity. Hence, if other factors are equal, face validity is a quality earnestly to be sought for teacher tests. Yet, as we well know, how often do we find situations in which "other factors are equal"?

TRADE-OFFS IN TEACHER TESTING

If a teacher test can be created that simultaneously yields valid inferences and also possesses face validity, then it should be cherished. It should be cherished, that is, if the costs of achieving face validity are not prohibitively expensive. For example, future teacher tests might call for examinees to instruct specially assembled, randomly assigned pupils in one or more short lessons. These lessons could be videotaped so that, subsequently, the teacher's instructional prowess could be judged by a panel of experts. The costs associated with this sort of face-valid assessment approach, obviously, are far from trivial. If a state is not reasonably affluent, such assessment tactics would probably be out of the question on financial grounds alone.

It is possible that a less face-valid assessment approach, for

example, one involving the examinee's multiple-choice answers to verbal simulations of classroom situations, would be so highly correlated with the more elaborate test that it is quite clear the extra assessment dollars were being spent exclusively in quest of face validity. The key issue to be faced by decision makers in such situations is whether the boost in face validity is worth the boost in assessment costs.

There is another sort of trade-off that should be verboten to teacher testers, namely, the enhancement of face validity at the cost of validity regarding score-based inferences. To illustrate, suppose that examinees are presented with videotaped stimuli consisting of classroom vignettes in which teachers are functioning at differing levels of proficiency. In a multiple-choice version of the test, examinees select from a series of alternative interpretations the most appropriate analyses of the videotaped teachers. In an interview version of the test, an interviewer interacts with the examinee to record the examinee's appraisal of the videotaped sequence and also the examinee's rationale for that appraisal. This examinee–examiner interview is videotaped and, subsequently, the examinee's performance judged by a team of trained evaluators who view the examinee–examiner videotape.

In this comparison, the multiple-choice version appears to be a fairly conventional selected-response assessment approach whereas the interview version appears to be a far more face-valid scheme in which examinee and interviewer discuss at length the substance of classroom instruction. The interview version deals more directly with the "stuff" of teaching and, therefore, would typically get face-validity votes both from examinees as well as from those who subsequently viewed the interview videotapes. Indeed, in contrast to the multiple-choice version's prosaic selected-response strategy, the interactive interview version positively glistens.

Yet, suppose that intensive probing of both approaches reveals the dominant factor operative in determining the quality of an examinee's performance on the interview version is the examinee's skill in oral discourse. Examinees who can chatter comfortably with interviewers might secure better scores from judges irrespective of how much those examinees know about classroom instruction. For the interview version of the test, face validity has been purchased by reducing the validity of score-based inferences. That price is too great.

Yet, in today's highly publicized educational climate where teacher testers are eager to curry the favor of teachers unions, educational policy makers, and the public in general, this type of

trade-off will, for some test developers, be too appealing to resist. Even worse, teacher tests may be created that are judged favorably on the basis of face validity alone. This would be truly deplorable.

REPRISE

In this analysis it has been argued that face validity may become an all too alluring magnet for teacher testers. Whereas face validity is an important consideration in a political context, it has precious little to do with the validity of the score-based inferences we draw from teacher tests. If face validity can be attained with tolerable increases in costs for teacher tests that otherwise yield valid inferences, face validity should be sought. Both examinees and others concerned with the teacher tests will be more positively disposed toward such face-valid tests. If, however, an increase in face validity causes a decrease in the validity of score-based inferences, then efforts to enhance face validity should be foregone. Teacher testers should not become so preoccupied with the appeal of face validity that they fail to scrutinize the validity of a test's score-based inferences.

In a politicized milieu, when we are eager to secure assessment approbation from many parties, face validity is an enormously attractive commodity. But face validity, as was true with siren-sung songs, is accompanied by both promise and peril. Teacher testers dare not allow face validity's promise to mask its perils.

REFERENCES

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.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7,* 191–205.

Murray, S. L. (1986). *Considering policy options for testing teachers.* Northwest Regional Educational Laboratory. Portland, OR.

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

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

Shulman, L. S., & Sykes, G. (1986). *A national board for teaching? In search of a bold standard.* Task Force on Teaching as a Profession, Carnegie Forum on Education and the Economy. New York: Carnegie Corporation of New York.