1990


George F. Madaus
Boston College

Follow this and additional works at: http://digitalcommons.unl.edu/burosassessteaching

Part of the Educational Administration and Supervision Commons, and the Educational Assessment, Evaluation, and Research Commons

http://digitalcommons.unl.edu/burosassessteaching/9

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Assessment of Teaching: Purposes, Practices, and Implications for the Profession by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Legal and Professional Issues in Teacher-Certification Testing: A Psychometric Snark Hunt

George F. Madaus
*Boston College*

But the Snark is at hand, let me tell you again!
'Tis your glorious duty to seek it!
To seek it with thimbles, to seek it with care;
To pursue it with forks and hope;

But the Judge said he never had summed up before;
So the Snark undertook it instead,
And summed it so well that it came to far more
Than the Witnesses ever had said

But the valley grew narrow and narrower still,
And the evening got darker and colder,
Till (merely from nervousness, not from goodwill)
They marched along shoulder to shoulder

—Lewis Carroll (1975)
*The Hunting of the Snark; An Agony in Eight Fits*
The validation of teacher-certification tests has become a snark hunt. The snark pursued by test contractors however, is not nearly as elusive, nor fanciful as Carroll’s shadowy beast. Their snark is a “legally defensible” test. Test contractors are—merely from nervousness (not from goodwill)—marching along shoulder to shoulder with jurists. I will argue that hunting the shadowy snark through the dark wood of the court room cheapens the conceptualization of validity and delimits the conduct of validation studies to minimalist exercises designed to obtain a positive result.

I start from the basic premise that the most important feature of any test is the accuracy of the inferences or decisions made from the score. Countless generations of measurement students have been taught the old saw: A test can be reliable but not valid, but if it’s valid, it has to be reliable. Here I argue that a test can be legally defensible but not valid. However, if the results from a well-designed validation process, on balance support the inferences or decisions made on the basis of the test, legal defensibility should be satisfied. Legal defensibility is important. Tests must be able to stand court scrutiny. However, I contend, we have the cart before the horse—whatever courts will accept drives current validation practices. Legal precedent is circumscribing inquiry about the essential question underlying all testing—“how correct are the inferences made from a person’s score?”

The vulnerability of testing companies engaged in this legal snark hunt consists in their readiness to view their task no more broadly than that of satisfying the desire of state agencies for legally defensible certification tests. Some contractors are in danger of becoming ambidextrous yes men for states. Testing companies, of course, insist on test validity. However, faced with the exigencies of the applied world—legal precedent, contract obligations, tenacious timelines, and budget limitations—they offer excuses to absolve themselves from gathering construct and criterion-related evidence related to their products. They have built a tunnel through a mountain of precepts about validity which are regularly taught students in test and measurement classes. They take the law of validity literally, but argue that the letter is elastic, plead-

---

1There are notable exceptions. ETS refused to bid on the lucrative Texas and Arkansas recertification testing program because, as Gregory Anrig publicly pointed out, the NTE tests should not be used for decisions regarding retention or termination. The Guidelines For Proper Use of NTE Tests, states that such decisions: “about in-service teachers should be based on teaching competencies as determined directly by the supervisory and evaluation procedures of the employing school district” (NTE Policy Council, 1985, p. 7).
ing for leniency from basic precepts and strictures. The companies deny that their product purports to measure a complex construct—a subset of competence in the classroom.

This behavior of test contractors is a reaction to the following factors:

- the mandates of policy makers for a quick, visible, quantifiable, and administratively convenient fix to a perceived problem with the quality of the teaching corps
- a naive confidence on the part of policy makers and the public that multiple-choice tests can improve the quality of the teacher corps
- a lack of understanding on the part of policy makers and the public, concerning the kinds of evidence needed to support the inferences they wish to make about teacher candidates
- the bureaucratic nature of the state agencies charged with implementing the mandate of policy makers
- the funding level and lead time available to implement a systematic validation effort
- the applied, commercial nature of the testing industry
- the selection of contractors through the competitive, low-bid RFP process
- the allegiance of testing companies to the state agency that awards the contract, rather than to examinees who later pay to sit the exam, or the general public who often base their perceptions of educational quality on test scores
- the evolution of a set of tried and true procedures for gathering content validation evidence that redounds to confirmation rather than disconfirmation
- the vested interest of policy makers in vigorously defending their programs when legally challenged
- the competitive, adversarial nature of litigation that colors arguments on both sides about scientific-technical issues
- the existence of legal precedent about what is sufficient to justify the use of employment tests

I was asked to address the professional and legal implications of teacher-certification testing. This chapter, therefore, consists of two distinct sections. In the first section, I examine the legal challenges to teacher-certification tests. I offer examples of how testing experts operate within the legal arena. In this first section validity
issues are interpreted from the legal rather than psychometric tradition. I begin by describing two types of legal challenges to teacher-certification tests and illustrate the part experts from the psychometric community play in these legal challenges. I conclude the first section with a description of the formula for a legally defensible teacher-certification test that has emerged from court decisions.

In the second section I leave behind the legal tradition and examine the validity issues surrounding teacher certification testing from the professional point of view. I begin by offering examples of the inferences and decisions made by various publics from scores on teacher certification. I then develop the psychometric implications for validation that flow from these inferences, and the problems associated with them in the applied situation. Next, I discuss issues related to the mix of content, construct, and criterion-related evidence that I feel is necessary if we are to understand what we are measuring and basing important certification decisions on. Finally, because the cut-score is the trigger for any inference or decision about teachers, I raise issues related to its validity.

Before I begin, two caveats. First, my treatment of litigation is strictly that of layman. My credentials consist of a single week in law school before leaving for the army, a decision I've never regretted. Second, I have a bias about high-stakes tests. If the use of a test has the potential to harm, or has serious consequences for individuals, the test should be subjected to a thoroughgoing validation before it is used operationally. Teacher-certification tests certainly trigger this bias. Large numbers of examinees are affected, particularly large numbers of minorities. From my perspective, the protection of the individual takes precedent over the state's interests until the state can demonstrate that the test, although not error free, does a reasonably good job of identifying individuals who in all likelihood lack the necessary skills or knowledge to be minimally successful in the classroom. The last clause anticipates my later argument that teacher certification tests are really about a candidate's competence in the classroom.

THE BASIS FOR LEGAL CHALLENGES

I shall concentrate on two theories under which teacher-certification tests are challenged: (a) the Fourteenth Amendment to the Constitution; and (b) Title VII of the Civil Rights Act of 1964. I do not attempt to analyze the extensive case law on which these theo-
ries are based. Instead, I will use plaintiffs’ briefs in *Allen v. The Alabama State Board of Education*, (1985), and *The Georgia Association of Educators v. State of Georgia* (plaintiffs’ brief, 1987) to illustrate the way they were recently used to challenge two teacher certification-testing programs.²

The Fourteenth Amendment to the Constitution

The Equal Protection and Due Process clauses of the Fourteenth Amendment to the Constitution provide the first grounds for legal challenges to teacher-certification tests. Either can be used on behalf of an examinee to challenge the use of a test for entry to the profession as an arbitrary, irrational, or unreasonable governmental action. Thus, for example, in *Schware v. Board of Bar Examiners* (1957) the court ruled that “A state cannot exclude a person from the practice of law or from any other occupation in a manner or for reasons that contravene the Due Process or Equal Protection Clause of the Fourteenth Amendment” (p. 238–39).

The Equal Protection Clause

The Equal Protection clause applies when the government engages in any form of grouping to determine differential government treatment of individuals. Once grouping occurs, courts ask, “What is the composition of the group?” and “What is done to the individual as a result of being placed in the group, that is to say, does it result in the person being stigmatized or denied a basic right such as freedom of speech or the right to vote?” Answers to these questions determine which of three forms of scrutiny of evidence the court will consider in examining the case: (a) If a racial group is affected adversely, or if a fundamental right guaranteed under the constitution is denied on the basis of being placed in a particular group, the claim receives strict scrutiny. Where racial grouping occurs it is necessary for plaintiffs to show that grouping was intentionally race based. Further, the court looks to see that the government is pursuing a compelling state interest and that no

²The reader is warned that I have worked for the Plaintiffs on both of these cases and that although I have tried to report objectively on the “facts” my possible bias should be noted. I have used the following documents in what follows. *Allen v. The Alabama State Board of Education Plaintiffs*” and Plaintiff-Interveners’ Joint Proposed Findings of Fact and Conclusions of Law, Civil Action 81-697-N in the United States District Court for the Middle District of Alabama Northern District.
reasonable alternative is less discriminatory. (b) If there is grouping not based on race, or if no fundamental right is denied, the case receives a relatively simple scrutiny. The question the court addresses under this analysis is whether the government is pursuing a legitimate governmental interest and doing it in a rational, non-arbitrary, noncapricious way. (c) Intermediate judicial scrutiny falls between simple and strict. It is used in cases where a well-defined nonracial group is affected or where a fundamental right isn’t involved, and the issue has important social significance. Courts have applied the intermediate scrutiny criterion in sex-discrimination and school-finance cases.

The Due Process Clause

The Due Process Clause applies when there is governmental deprivation of a life, liberty, or a property interest. For example, veteran teachers have a property interest in maintaining their jobs. Recertification testing could deprive them of this property interest. There are two types of due process: substantive and procedural. Substantive due process asks whether a governmental action is just, or fundamentally fair. Arguing that substantive due process has been violated in a teacher-testing case involves demonstrating that the use of the test is so arbitrary and so capricious that it bears no relationship to the state’s objective. This is obviously extremely difficult to prove. Procedural due process asks whether the government went about its decision making in a systematic manner designed to minimize mistakes. Procedural due process also affords citizens an opportunity to have advance notice of, and an opportunity to influence, important government decisions that affect them.

Allen v. The Alabama State Board of Education (1985) is illustrative of a Fourteenth Amendment challenge. Alabama required that candidates for initial teacher certification pass a test of professional knowledge and a test specific to their certification area. Plaintiffs challenged the state’s testing regulation by arguing that the tests employed:

were fundamentally flawed, varied from sound psychometric guidelines, and violated basic principles of fairness. Consequently, defendants’ classification of teacher applicants who have failed to achieve their cut-off score on these examinations is irrational and violates the equal protection and due process rights of the class. (p. 31)

The plaintiffs, relying heavily on precedent from Debra P. v. Turlington (1979),—the Florida minimum competency graduation
testing case—challenged the testing program in part on pro-
cedural due process grounds:

[the] defendants hastily concocted and implemented these tests as an absolute requirement for teacher certification in a period not exceeding 18 months. The tests were in fact used as a requirement for certification only some three months after their construction, the first opportunity to provide students notice of their content [italics added]. Under these circumstances defendants' implementation schedule of the test provided the plaintiff class insufficient notice, thereby violating their rights to due process under the fourteenth amendment to the Constitution of the United States. (Plaintiffs & Plaintiff-Intervenors’, 1986 p. 39).

Plaintiffs in Allen vs. State Board developed their Fourteenth Amendment claim further using opinions from Debra P. (1981). First, they asserted that:

Just as successful completion of a high school program creates a “state-created understanding” that one will receive a diploma, successful completion of a state-approved teacher training program creates such an understanding that one will receive a teaching certificate. (Plaintiffs’ & Plaintiff-Intervenors’, 1986, p. 33–34)

Second, using language directly from Debra P. plaintiffs asserted that a state-required examination used to deny persons a classification to which they have a property right or a limited interest must be “a fair test of that which is taught in its classrooms.” (644 F.2d 408) Third they cited the following ruling from Debra P. to argue that Alabama had violated due process: “If the test covers material not taught the students, it is unfair and violates the Equal Protection and Due Process clauses of the United States Constitution.” (644 F2d 408). Plaintiffs argued that because the testing program had a disparate racial impact, their Fourteenth Amendment claim should receive strict scrutiny. Finally, given disparate impact the defendants have the burden of proving that the examination was in fact a fair test of matters actually taught in the classrooms of their State-approved teacher-training programs and that there was no reasonable alternative that would have a lesser discriminatory impact.

By now readers will recognize the argument for curricular-related validity evidence (Madaus, 1983). Whether the Alabama tests had curricular validity or not isn’t nearly as interesting as the argument that what is taught in teacher-training institutions and measured by multiple-choice items is, in fact, necessarily related
to performance in the classroom. I examine this line of argument in more detail later.

Title VII of the Civil Rights Act of 1964

The most widely employed and powerful legal theory used to challenge teacher-certification tests is a claim under Title VII of the Civil Rights Act of 1964. Title VII mandated nondiscrimination in employment by reason of race, sex, or national origin. The Act also created the Equal Employment Opportunity Commission to enforce Title VII. It is important to remember that Title VII is an employment statute and, therefore, covers only employment testing. (The bulk of testing in this country cannot be challenged under Title VII.)

There are two threshold criteria that must be met by plaintiffs to mount a successful Title VII challenge. First, plaintiffs must establish that the State is an employer. (Although state employment issue has no direct psychometric implications, unless this first criterion is met, testing issues simply will not get heard.) Defendants in teacher-certification-testing cases argue that the State is merely a licensing board, not an employer; that teachers are employed by local school districts. Therefore, they contend that the State's certification tests can't be challenged under Title VII. However, in United States and North Carolina Association of Educators v. North Carolina (1977) and The Georgia Association of Educators v. State of Georgia (1987) the courts ruled that the state plays a pervasive role in the operation of the schools and in the employment of teachers. For example, in the Georgia case the court ruled that:

Although the public schools in Georgia are operated by local school boards, the activities of the local boards are part of a state-wide educational program which is administered by the State and by the Georgia State Board of Education. The State and, in particular, the Board of Education, exercises authority over the local school board in matters that range from curriculum and program requirements to funding, construction of facilities, and long-range planning by local school systems. Under these circumstances, it cannot be determined as a matter of law that defendants are not "employers" for purposes of Title VII. (C86-2234A at 3)

If the defendant is an "employer" under Title VII, the plaintiffs must meet the second precondition to Title VII coverage. They
must show that the tests have substantial adverse racial impact. The importance of this second test cannot be over emphasized. A shoddy, invalid teacher-certification test that fails minorities and nonminorities at essentially the same rate as nonminorities cannot be challenged under Title VII. Further, without a Title VII claim it is extremely difficult to challenge a defective, but nondiscriminatory, test under a Fourteenth Amendment due process or equal protection theory.

Once plaintiffs demonstrate adverse impact, the burden of proof shifts to the defendants to show the tests are “job-related” (Albemarle Paper Co. v. Moody, 1975). Finally, if the tests are shown to be properly validated, the plaintiffs have an opportunity to prove that “other tests or selection devices without a similar undesirable racial effect, would serve” the defendants’ purpose. (Albemarle Paper Co. v. Moody, 1975)

DETERMINING ADVERSE IMPACT

The determination of adverse impact is made by comparing the selection rate for the minority group with that of the nonminority group. There are two ways of showing disparate impact: (a) the “80% rule” and (b) the “two or three standard deviation rule.” There is, of course, extensive case law interpreting these rules, and learned treatises about them (e.g., Baldus & Cole, 1980; Kaye & Aickin, 1986). I do not attempt to review that literature. Instead, I use a recent challenge to a teacher-certification-testing program—Georgia Association of Educators v. State of Georgia (1987) —to illustrate the role testing experts played in the determination of disparate impact. The Georgia case affords the reader a concrete glimpse of how testing experts operate in the adversarial world of law.

The 80% Rule

The “80% rule” was adopted by the EEOC as a “rule of thumb” for assessing adverse impact. The Question and Answers on Uniform Guidelines on Employee Selection Procedures (UGESP) states:

A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rates for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-
fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. **Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms** [italics added] or where a user’s actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. [U.S. Equal Employment Opportunity Commission 29 C.F.R. § 1607.4D]

The 80%, or 4/5 rule, was not intended as a legal definition of discrimination but as a practical device to guide the EEOC in the use of their enforcement resources (Schlei & Grossman, 1983). Further, as the italicized section reveals, the 80% rule is not a hard and fast requirement.

In *Georgia Association of Educators v. State of Georgia* (1987), plaintiffs pointed to data over an 88-year period, in which Whites took the Teacher Certification Test (TCT) a total of 57,503 times and passed 48,986 times, for a pass rate of 85.2%. Blacks took the TCT 19,826 times and passed 8,329 times, for a pass rate of 42.0%. Because the Black pass rate was only 49.3% of the White rate \( \frac{42}{85.2} \times 100 \) — considerably less than 80% — plaintiffs argued that their Title VII claim was justified (Plaintiffs’ Brief In Opposition, 1987, p. 45). Thus, in Georgia the 80% rule seemed at first glance, straightforward. But looks are deceiving. Two issues muddy the waters; both involve the testing community.

The first issue revolves around the formulation of the comparison groups. Two groups were affected by the Georgia TCT: (a) the “Renewal Certificate Group,” consisting of experienced Georgia teachers seeking recertification; and (b) the “Initial Certificate Group,” consisting of candidates applying for Georgia certification for the first time. Robert Rentz, an expert for the State Department of Education, reported that his calculations showed that the pass rate for Blacks in both groups was more than 80% of the pass rates for Whites. He concluded, “no ‘substantially disproportionate racial impact’ is evidenced” (1987a, p. 4). On the basis of the Rentz analysis, defendants argued that plaintiffs had no claim under Title VII.

Plaintiffs countered that an analysis by their expert, John Poggio (1987a) showed that Rentz’s numbers were substantially incorrect because of “egregious multiple counting of records” (Plaintiffs Brief in Opposition, 1987, p. 35). In other words, a single person who passed a single test was counted more than once by Rentz; often five or more times. Poggio also challenged Rentz’s inclusion
of other cases in forming his groups. Rentz (1987b) in reply, argued that the multiple counting was the result of a single person using the test result to apply for more than one certificate, and therefore was legitimate.

Poggio used different inclusion rules to formulate his groups—a single person taking a single test was counted once—and found for all examinees irrespective of initial or renewable status, a ratio of Black pass rate to White pass rate of 74.6%; less than the magic 80% criterion for a Title VII claim. He then calculated the ratio of Black to White pass rates for the Initial and Renewable groups separately and found the ratios to be 69.6% and 82.8%, respectively.

This interchange between testing experts over the 80% rule isn't offered to resolve either claim, nor as a second edition of How To Lie With Statistics (Huff, 1954). Instead it illustrates how an important legal trigger can color an expert's approach to an otherwise apparently "objective" set of data.

The second issue surrounding the 80% rule is closely related to the group-definition issue. Given multiple retakes, when do you calculate your ratios to go to court? The ratio will slide with each retake. Because Georgia Blacks failed the TCT in greater numbers than Whites, the pool retaking the test for recertification became more and more Black over time. And with each Black passing a retake the 80% rule becomes harder to meet. The State, of course, would include as many retakes as possible before going to court. Plaintiffs would like to establish a temporal base line beyond which retakes are not considered in the calculations.

Before leaving the 80% rule, one additional observation is in order. It takes no genius to recognize that the 80% rule is a function of where in the distribution the agency sets the passing score. Field-test data, or data from the first or second administration of

---

3Poggio's analysis shows that 5,782 matches in the Rentz Initial Certificate file were the result of multiple counting; the comparable figure for the Renewable Certificate file was 5,715 matches. Poggio, offered as one example of thousands of cases of multiple counting a Black who took the TCT Science test on only one occasion and passed it and was counted as a "pass" six times. (Poggio, 1987, p. 5)

4Thus Rentz (1987b) argued that the person counted six times as a pass for the science test cited by Poggio could be used by an "otherwise qualified individual in applying for 15 different certificates, any one of which might subsequently qualify that person for different job openings" (p. 1).

5My belief is that the Poggio analysis best reflects the situation of how Blacks are affected by the TCT.
the test, permit fine tuning the cut score so that the ratio of the minority pass rate to the nonminority pass rate is greater than 80%. No one admits to this practice, but several plaintiff lawyers I have talked to believe that available data on the minority and nonminority pass rates are used to adjust the cut score after the fact in order to avoid a Title VII claim.

What's wrong with adjusting the cut score to raise the pass rate for minorities? Nothing, if the program is viewed cynically as nothing more than a public relations exercise to assuage a distrustful public that "incompetent" teachers will not be allowed in the classroom. After all, why get involved in very costly litigation if it can possibly be avoided by judicious choice (no pun intended) of a cut score from the known distributions of the two groups? However, if the question of the correctness of the inferences, decisions, or descriptions made from candidates' test performance is taken seriously, then clearly the operational cut-off score should be based on empirical considerations rather than on the vagaries of the 80% rule. I shall return to the question of validity and the cut score later.

The Two or Three Standard Deviation Rule

As we saw in the preceding section, the 80% rule is merely a guideline which admits of exceptions. The courts also consider tests of statistical significance in determining whether a selection device has a disparate impact. The Fifth and Eleventh Circuits have adopted the "two or three standard deviations" test first employed in 1977 by the Supreme Court in *Castaneda v. Partida* (1977). In recent cases these Circuits have approached the question of disparate impact solely from a standard deviation analysis (Plaintiffs Brief in Opposition, 1987).

*Castaneda* was a case involving jury selection. The Supreme court ruled that a difference of two or three standard deviations (in reality standard errors) between the percentage of the minority group in a large population (the general population) and the percentage of minorities actually selected for jury duty is generally indicative of discrimination.

Let's examine how this rule was operationalized by the experts in the Georgia case. Rentz (1987a) reported that Blacks in the Initial group constituted 12.7% of the larger population who attempted the TCT, but only 10.8% of the smaller sample who passed: a difference of 1.9 percentage points. He then argued that
If the differences obtained here (1.9) were translated into standard deviations based on the sample size for the sample passing the TCT (20,660), the result would be 8.26; but if that same difference were translated using the same sample size as *Castaneda v. Partida* (e.g. 870), the result would be 1.68. (p. 4)

For the Renewable group the difference was 2.4, or 8.77 SDs using an N of 21,370, but only 1.76 SDs if adjusted to a sample size of 870. In other words, Rentz considered the N of 870 in *Castaneda v. Partida* as normative. He reasoned that “almost any difference can be found to be statistically significant if the sample size is large enough. The question of whether or not 1.9 is reflective of ‘substantially disproportionate racial impact’ is not a statistical question” (Rentz, 1978a, p. 4) Rentz was certainly correct on two points: (a) large samples can ensure statistical significance, and (b) the question of what value constitutes substantial racial impact is not statistical. However, his conclusion that for both initial and renewable groups there was “no statistical evidence that persuades me that there is a substantially disproportional racial impact resulting from the requirement to pass the TCT, within the framework of . . . *Castaneda v. Partida*” (p. 5) is itself a value judgment.

Poggio (1987a) countered Rentz’s acceptance of the N of 870 as sacrosanct:

it makes absolutely no sense in statistical terms . . . to take the number of standard deviations computed on the basis of disparities for a sample of over 20,000 test-takers, and “translate” them into the number of standard deviations that would result if the same percentage difference has been observed in much smaller samples . . . . In as much as the difference in black/white pass rates on the TCT was in fact *established* on the basis of the performance of over 20,000 persons, the meaning of that difference properly *must be assessed* in terms of the performance of over 20,000 people. (pp. 12–13)

The sample size issue apart, Poggio criticized Rentz for making the wrong comparison. He argued that the TCT is a selection procedure, and that the EEOC’s *Uniform Guidelines on Employee Selection Procedures* defines adverse impact for such procedures in terms of the difference between the selection rate of a particular racial group (in Georgia, Blacks) and the selection rate of the group with the highest selection rates (in Georgia, Whites). Poggio argued that the issue in *Georgia* was different from the issue in *Castaneda*. In *Castaneda* the issue was whether the percentage of minorities selected for jury duty was representative given the per-
centage of minorities in the general population. In the Georgia case Poggio identified "the extent to which the black pass rate differs from the white pass rate" as the central issue (Poggio, 1987a, p. 13). In other words, Poggio argued that the basic question in employment testing is not one of representation but of whether the test treats the two groups more or less equally. On closer examination the representation and the comparative selection methods are identical if the total population of examinees is made up only of Blacks and Whites. However, if the total population includes members of other racial or ethnic groups, the results from the two approaches can differ.

Poggio computed the number of standard deviations between the Black and White pass rates (Table 7.1) to answer what he saw to be the relevant statistical question, that is, "how likely is it that if selection were unbiased as to race, the Black pass rate would be as far below the white pass rate as the data show it is in this case?" (p. 14). Poggio pointed out that his comparative pass rates are well in excess of the Supreme Court's Castaneda rule and are indicative of discrimination. He argued that even for the Renewal group, where the pass rate is slightly above the 80% rule (82.8%), the disparity between the two pass rates is statistically significant. Finally, he pointed out that even using the Rentz representation comparison (but without the 870 transformation) the number of standard deviations obtained is never less than 9.71.

Rentz (1987b) countered Poggio's critique with an assertion that an analysis of proportional representation is the appropriate comparison. He did not, however, develop the argument of why it was appropriate. Instead he argued that the important statistic to consider is the absolute size of the difference between the population percentage (in this case Blacks taking the test) and the same group's proportional representation in the sample selected (i.e., passing the test), not the translation of this difference to standard deviations.

Rentz's argument is worth closer examination. From a review of four cases similar to Castaneda he extracted differences between

| TABLE 7.1 |
|------------------|-----------|
| • All administrations | 119.70 SDs |
| • All Examinees; cumulative pass rate | 84.12 SDs |
| • Initial certification group; cumulative pass rate | 77.12 SDs |
| • Renewal certification group; cumulative pass rate | 38.86 SDs |

*Note. Derived from data in Pozzio 1987a*
population percentages and the sample selected percentages. These are shown in Table 7.2. He concluded that an absolute difference of 19—give or take a few percentage points—is what the courts should be sensitive to. That this absolute difference is a better criterion than an arbitrary “standard deviation rule” which fluctuates with sample size.6

Rentz then computed the equivalent differences in population and sample percentages for the Initial, Renewable, and All Examinee groups. These data are presented in Table 7.3 along with the percentage difference from Castaneda. The samples Rentz used in computing his differences for the Initial and Renewal groups were substantially different from Poggio’s samples.7 From these data Rentz argued that

None of the results in the present case are close to the magnitude of the results in Castaneda v. Partida. The differences I obtained for certificate applicants are quite small, 1.9 for Initial and 2.4 for Renewal applicants; . . . The difference in representation for Dr. Poggio’s test takers is 11.1, not as small as that for applicants, but below the average of 19 for the four examples cited in the Supreme Court opinion. The 11.1 difference is even below the lowest result of the four cases cited [italics added]. (Rentz, 1987b, p. 7)

Notice Rentz’s use of the difference of 19 as a criterion.

6Rentz pointed out that a difference of 40.1 obtained in Castaneda translates to 29 standard deviations but if the sample size were to increase ten fold from 870 to 8,700 the standard deviation would be 92. He then asked, rhetorically, if the degree of underrepresentation is more in the second case than the first, and answers no; the number 40.1 represents the degree of “underrepresentation” not the size of the standard deviation.

7For example for the Initial group Poggio (1987a) reported 7,557 Blacks took the test and 5,003 passed (p. 20), whereas Rentz (1987b, p. 5) recorded 2,735 Blacks in the population of test takers of which 2,214 passed. The two experts were working from different tapes, with different inclusion rules.
The problem with Rentz’s absolute percentage difference criterion is that it is an altogether relative, population-dependent index.8 It is somewhat analogous to the test dependence of the traditional item-discrimination index. The percentage of the minority group in the overall population always sets the upper bound for the difference between the population percentage and the sample selected percentage. Thus if you were to adopt Rentz’s “difference of 19” proposal, once the percentage of Blacks in the population falls below 15% you are automatically out of the Castaneda ball game. Table 7.3 dramatically illustrates this population dependence; for the Initial certification group the highest difference you could possibly achieve is 12.7—and then only if no Black passed the test.

Taking the ratio of the sample selection percentage to the population percentage removes the population dependence inherent in the Rentz criterion. Table 7.4 presents these ratios for Castaneda, the four related cases, and the All Examinee group which Poggio developed and Rentz used in his argument. Ratios for the Initial and Renewable groups are not shown in Table 7.3 because of the large discrepancy in the sample size between the Poggio and Rentz versions. Examination of the ratios in Table 7.4 reveals that in Castaneda for every two Mexican Americans in the population only one was selected for jury duty; for the All Examinees group in Georgia the ratio is also close to 2 to 1. Both in Castaneda and Georgia the ratio is lower than that in Turner v. Fouche. This analysis, of course, begs the question of whether the representation approach or the comparative selection approach is the correct comparison to make, a question ultimately for the courts, not for measurement experts, to decide.9

---

8I am greatly indebted to John Poggio for the analysis that follows of the Rentz difference of 19 criterion.

9Using the All Examinee figures you can show that this ratio technique is identical to the 80% rule when there are only two groups. Both Rentz and Poggio reported
The analyses and arguments in the Rentz and Poggio affidavits offer us a glimpse into the world of the expert witness. We see two respected, highly qualified professionals attempt to interpret, in the best possible light, what appears at first blush to be simple court guidelines on what constitutes disparate impact. The stakes are extremely high in this interchange. Make no mistake! It is not an academic debate over what statistic to use. Without disparate impact, there is no Title VII claim, and questions about the validity of the test cannot be raised. And as things stand now, absent a Title VII claim, the contractor and agency are home free. There is no other forum, or independent auditing agency to which examinees can bring questions about test validity. I return to this issue later.

Can the profession arrive at a consensus on some of the issues posed in the Rentz/Poggio interchange? There could probably be agreement on the effect of sample size on statistical significance and power associated with the “standard deviation rule.” We could probably agree on the correct definitions for the population and the sample selected given more than two racial groups in the overall population. We might even be able to agree on whether the absolute difference criterion of Rentz, or the ratio of sample selected to population criterion described in Table 7.4 is appropriate. But how large the difference or ratio should be is clearly a value question. Similarly, although the 80% rule is free of the population size problem associated with the standard deviation rule, the actual percentage constituting disparate impact is also a value question, particularly when the ratio of selection is close to 80%.

In the final analysis, a judge using his or her own personal calculus—colored by legal precedent, his or her training, temperament, history, intellect, personality, and predilections—must interpret the disparate impact data from contending experts, and decide if an employment test is treating minorities in essentially the same way as it treats Whites. How did the district judge in Georgia interpret the Rentz/Poggio data? Although not ruling on the disparate impact issue directly, he denied the State’s motion for partial summary judgment because “the plaintiffs have raised numerous genuine issues of material fact regarding the statistical analysis performed by Dr. Rentz and relied upon by the defen-

that for this group the selection ratio of Blacks to Whites is 49.31%. The ratio of Blacks selected to Blacks in the population of test takers is .566. The ratio for the Whites is 1.145 (85.5/74.4). Dividing .566 by 1.145 gives you 49.31%.
TABLE 7.4

<table>
<thead>
<tr>
<th>Group</th>
<th>Population %</th>
<th>Sample %</th>
<th>Sample/Pop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castaneda</td>
<td>79.1%</td>
<td>39.0%</td>
<td>.493</td>
</tr>
<tr>
<td>Turner v. Fouche</td>
<td>60.0%</td>
<td>37.0%</td>
<td>.617</td>
</tr>
<tr>
<td>Whitus v. Georgia</td>
<td>27.1%</td>
<td>9.1%</td>
<td>.336</td>
</tr>
<tr>
<td>Sims v. Georgia</td>
<td>24.4%</td>
<td>4.7%</td>
<td>.193</td>
</tr>
<tr>
<td>Jones v. Georgia</td>
<td>19.7%</td>
<td>5.0%</td>
<td>.253</td>
</tr>
<tr>
<td>All Examinees</td>
<td>25.6%</td>
<td>14.5%</td>
<td>.566</td>
</tr>
</tbody>
</table>

Derived from data in Rentz 1987a & b

dants” (Order of Court, 1987, p. 4). In other words, he still had questions in his mind about disparate impact that have to be argued further. (The Georgia case was subsequently settled out of court.)

THE EMERGENCE OF THE LEGALLY DEFENSIBLE TEST

So much for legal theories used to litigate teacher-certification testing. I turn now to the impact of litigation on test validation. Reviews of employment-testing litigation in general (Novick, 1982; Wigdor, 1982) and teacher-testing litigation in particular (McCarthy, 1985; NTE Policy Council, 1985) are available elsewhere. Therefore, what follows is my rendering of the emergence from court rulings of the recipe for “legally defensible” teacher-certification tests and the ingredients of the recipe. To reiterate, it’s my contention that the precondition of “legal defensibility” drives applied validation efforts to the detriment of a careful consideration of the evidence needed to sustain the inferences and decisions made from the test scores. The form and technique to construct a “legally defensible” test has almost completely overshadowed the essential question of the meaning behind the test score. What is the form and technique of “legal defensibility”?

A legally defensible test is the product of the quasi-legal, quasi-scientific approach of the court to employment-testing litigation (Schlei & Grossman, 1983). The Second Circuit in Guardians Association of the New York City Police Department v. Civil Service Commission (1981) observed that while Title VII forces courts to consider employment testing, it is not primarily a legal subject:
Employment testing] is part of the general field of educational and industrial psychology, and possesses its own methodology, its own body of research, its own experts, and its own terminology. The translation of a technical study such as this into a set of legal principles requires a clear awareness of the limits of both testing and law. It would be entirely inappropriate for the law to ignore what has been learned about employment testing in assessing the validity of these tests. At the same time, the science of testing is not as precise as physics or chemistry, nor its conclusions as provable. While courts should draw upon the findings of experts in the field of testing, they should not hesitate to subject these findings to both the scrutiny of reason and the guidance of Congressional intent. (pp. 169-79)

It seems jurists and legal scholars are as uncomfortable as many of us in testing are with courts ruling on issues of test validity. However, unfortunately, by default the court has become the arbiter of the EEOC Guidelines and the Standards for Educational & Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1985).

Two paths to the validation of teacher-certification tests have emerged from this quasi-legal, quasi-scientific crucible of the court room. Both must show that the test is job related. The first, which I call the curricular validity approach, comes out of the United States v. South Carolina (1978) case involving the use of the NTE. As we saw plaintiffs in Alabama challenged that state’s teacher tests partly on the grounds that it lacked curricular validity. The second approach, which I label the content validity route, has evolved from a number of court decisions interpreting the concept “job relatedness” in employment testing. The two approaches are not qualitatively different. Both involve judgments by panels about the match between test items and a domain. The approaches differ in domain definition.

The Curricular-Validity Approach

In United States v. South Carolina (1977) the Justice Department and others challenged South Carolina’s use of the NTE program tests for initial teacher certification, and for determining, in part, the salary schedule of experienced teachers. Plaintiffs proved disparate impact, shifting the burden under Title VII to the defen-
dants to demonstrate validity (i.e., job relatedness). To meet this burden of proof, the South Carolina commissioned ETS, the developers of the NTE, to conduct a validity study. The study sought to demonstrate "content validity by measuring the degree to which the content of tests matches the content of the teacher training programs in South Carolina" (United States v. South Carolina, 1977, p. 1112). The ETS study involved 456 faculty members from twenty-five colleges and universities in South Carolina. The participants, convened as members of panels and looked at the tests under controlled conditions and made judgments about the relationship between the tests and the curricula of teacher-training institutions in the state. (United States v. South Carolina 1977)

Relying on the following passage from Washington v. Davis (1976), a police employment case, the lower court endorsed in principle the ETS decision to validate the NTE against the academic training program rather than actual job performance:

"A positive relationship between the test and training-course performance was sufficient to validate the former, wholly aside from its possible relationship to actual [job] performance as a police officer [italics added]. . . .

Nor is the construction foreclosed by either Griggs or Albemarle Paper Co. v. Moody, 422 U.S. 405, 95 S.Ct. 2362, 45 L.ed.2d 280 (1975); and it seems to us the much more sensible construction of the job-relatedness requirement. (426 U.S. at 205–251, 96 S.Ct at 2053, p. 1113)

The Supreme Court affirmed the lower court decision without comment. As I read this ruling, a contractor wishing to build a legally defensible teacher-certification test simply has to show that the test items correspond to skills and knowledge found in the state’s teacher-training curricula. This isn’t particularly difficult to do. The composition of the panels, and the techniques used to solicit from them the match between test items and teacher-training curricula are straightforward. Further, the methodology is relatively cheap, fast, and—if the test contractor had any foresight at all—confirmatory.\(^\text{11}\)

\(^{10}\)The court also cited Washington v. Davis (1976) to dismiss the plaintiffs constitutional challenge under the 14th Amendment: "The Supreme Court has held that a substantial relationship between a test and a training program—such as is found here [S. Carolina]—is sufficient to withstand challenge on constitutional grounds" (p. 1108).

\(^{11}\)It is interesting to note that in the mid 1940s for financial reasons the NTE probably at the request of teacher educators increased the professional information
However, I know of no recent defense of a teacher-certification test that relies solely on this approach. Now when the curricular-validity approach is used it is generally folded into the content-validity approach described in the next section. Nonetheless, it is worth looking more closely at the legal reasoning about validity found in United States v. South Carolina.

In Washington v. Davis (1976) a test of verbal skills and communication ability was used to screen applicants for a 17-week police academy training program. The Supreme Court accepted the lower court’s reasoning that “the lack of job performance validation does not defeat the test, given its direct relationship to recruiting and the valid part it plays in [the training regimen]” (426 U.S. 229 at 236).

The court seemed to brush aside the fact that the South Carolina examinees taking the NTE for initial certification had already successfully completed their teacher training. Instead, the court observed that the NTE, although not measuring teaching skills, “was a measure of the extent to which prospective teachers have mastered the content of their teacher training programs” (p. 1108). The court also pointed to the plaintiffs acknowledgement of the importance of teacher-training programs when they proposed that graduation from “an approved program alone is sufficient to protect the public interest” (United States vs South Carolina, 1987).

The court also opined it was proper to use the NTE scores of experienced teachers for salary purposes. In doing so the court chose to ignore the NTE Policy Board’s position that their program tests should not be used for such purposes:

The current NTE Program tests were developed to provide information about candidates’ academic knowledge and skills, typically acquired through a teacher-training program. They do not provide a direct evaluation of teaching performance. For this reason, NTE tests should not be used . . . directly or indirectly, for decisions regarding retention or termination. Such decisions about in-service teachers should be based on teaching competencies as determined directly by the supervisory and evaluation procedures of the employing school district.

Similarly, with the exception of master teacher or career ladder plans, NTE tests should not be used for decisions regarding compensation.

to 40% of the common exam total score, and the advisory council shifted from local school administrators to teacher training personnel. For a complete history of the development of the NTE, see Wilson (1984).
The current NTE tests measure knowledge and skills needed by the beginning teacher; more is required of the teacher in service. (NTE Policy Consul 1985, p. 7–8)

I applaud as eminently sound and fair ETS’s position on the use of the NTE program tests with in-service teachers. Nonetheless, there is a curious logical problem with the reasoning behind it that might explain the court’s ruling. Why should the knowledge and skills measured by the NTE be job-related for preservice teachers but not for in-service teachers? The court made no such distinction in terms of the degree of knowledge possessed. As far as the court was concerned what was good for the goose was good for the gander: “[T]he State could reasonably conclude that the NTE provided a reliable and economical means for measuring one element of effective teaching—the degree of knowledge possessed by the teacher” (United States v. South Carolina, 1977 p. 1109).

It seems to me that the court did not go far enough in its analysis of the Washington v. Davis precedent. The reasoning found in the dissenting opinion makes several very telling points. First, the defendants in Washington v. Davis at least offered a correlation between the admissions test and the final examination grades from the police training course to support their validity claim. There is no reference in United States v. South Carolina to any correlation between NTE scores and scores in either professional education courses, or in specific subject related academic courses.

Second, even if such correlations had been offered, the ruling seems to take for granted that teacher-training curricula are relevant to one’s later teaching performance. The dissenting justices in Washington v. Davis noted that: “Sound policy considerations support the view that, at a minimum, petitioners should have been required to prove that the police training examinations either measure job-related skills or predict job performance” (Quoted in Schlei & Grossman 1983, p 126) They argued further that a correlation between the admission test and the grades in training is supportive evidence only if: “(1) the training averages predict job performance or (2) the averages are proven to measure performance in job-related training” (Dissenting opinion Washington v. Davis, quoted in Schlei & Grossman 1983, p 124)

Grades in teacher training, like grades in other professional

---

12Although I have quoted the most recent NTE guidelines, similar statements can be found in earlier versions and would have been policy at the time of the South Carolina case.
training programs, do not correlate highly with later job performance. (see Haney, Madaus, & Kreitzer, 1987, for a review of this literature) Further, we simply take for granted that the curricula of teacher-training programs are related to teaching performance. I believe that aspects of teacher training may very well be relevant to job performance. Nonetheless, the question of whether those relevant aspects of the training experience can be captured by a secondary indicator—a multiple-choice test—remains unanswered. And, this is precisely the validity issue.

The dissenters in Washington v. Davis also had an excellent intuitive recognition the underlying construct at issue in that case: "[T]here is a substantial danger that people who have good verbal skills will achieve high scores on both tests due to verbal ability, rather than "job-specific ability."" (426 U.S. 270)13

In my opinion, in United States v. South Carolina the ruling simply side stepped job-relatedness and the job-specific ability issue raised by the dissenters in Washington v. Davis. Nonetheless, the South Carolina decision has provided contractors with one clear line of legally defensible "validity" evidence.

The Content-Validity Approach

A second, and more common route to a legally defensible teacher-certification test is to show job relatedness through content validation. First, I review the essentials of the major cases from which this approach to validation has emerged to describe for the reader the essentials of the approach. To illustrate how the approach was actually implemented and then challenged in court, I outline the steps taken by the test contractor National Evaluation Systems in validating Alabama's teacher certification tests, and briefly list the plaintiffs objections to each step in that process.

Unless plaintiffs can show intentional discrimination in the use of a test, the test will not be an issue in constitutional actions. (Schlei & Grossman, 1983). Thus, as we have mentioned earlier, the cornerstone of most challenges to an employment test is a Title VII claim. The first important case was Griggs v. Duke Power Co. (1971), where the Supreme Court ruled that "If an employment

13Interestingly Mitchell (1985) reviewing the NTE also raise a similar construct-related question about parts of the Core Battery. He suggested that different parts of the test may measure, "general intelligence, scholastic aptitude, overall academic achievement, and multiple-choice test item reasoning skills" rather than "mastery of particular domains of curriculum such as professional education" (p. 1066).
practice which operates to exclude Negroes cannot be *shown to be related to job performance* [italics added], the practice is prohibited" (p. 178). The Court went on to interpret the intent of Congress in enacting Title VII:

Nothing in the Act precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force *unless they are demonstrably a reasonable measure of job performance* [italics added]. (p. 180).

The key question in subsequent cases became "what constitutes acceptable evidence that the test is job related?" In *Albemarle Paper Co. v. Moody* (1975) the Supreme Court ruled that "[a discriminatory test must be] *predictive of or significantly correlated with* [italics added] important elements of work behavior which comprise or are relevant to the jobs or jobs for which candidates are being evaluated" (422 U.S. 431). Phrases like "predictive of" and "correlated with" imply criterion-related validation, at least for someone like myself with an educational measurement background. But as I experienced first hand in *Allen v. Alabama*, defendants in teacher-certification-test cases argue that criterion-related validity evidence is not necessary; that content-related evidence is sufficient to show job relatedness.14

Two 1981 cases seem to be the source of the content-validity approach to job relatedness: (a) a Ninth Circuit ruling in *Contreras v. City of Los Angeles* (1982), an employment case involving accountants; and (b) a Second Circuit ruling in *Guardians Association of the New York City Police Department v. Civil Service Commission* (1981). I shall use the *Guardians* case to illustrate the reasoning underlying the content-validation approach.

What does the content-validity approach to showing job relatedness entail? The Second Circuit, after an interesting discussion of content and construct validity, distilled from the EEOC Guidelines 14After a 6-week trial the district court did not rule because a prior settlement that the State, under intense political and media pressure attempted to disown, was upheld by the Eleventh Circuit (No. 86-7215, 1986). The settlement is now in effect, and experts from both sides will attempt to build a test within its framework. Among other things, both sides agreed to a stricter version of the Golden Rule Settlement in terms of item inclusion on subsequent tests. (Recently the state dropped the teacher exam entirely)
five attributes of an exam that, notwithstanding its disparate racial impact, had sufficient content validity to be used for employment decisions:

The first two concern the quality of the test’s development: (1) the test-makers must have conducted a suitable job analysis, and (2) they must have used reasonable competence in constructing the test itself. The next three attributes are more in the nature of standards that the test, as produced and used, must be shown to have met. The basic requirement, really the essence of content validation, is (3) that the content of the test must be related to the content of the job. In addition, (4) the content of the test must be representative of the content of the job. Finally, the test must be used with (5) a scoring system that usefully selects from among the applicants those who can better perform the job. (from employment Practices Decisions, p. 16979)

To understand how contractors in teacher-certification cases have employed these guidelines to build legally defensible tests it is necessary to examine more closely the steps in the job-analysis phase of the content-validation process just described.

The EEOC Guidelines call for an assessment “of the important work behavior(s) required for successful performance and their relative importance” (§ 14(C)(2)). In the Guardians case, the Civil Service Commission went through a five-step procedure designed to meet this job-analysis requirement. First, the work behaviors were identified by extensive interviewing of job holders and supervisors. Second, the list was reviewed by another, smaller panel of job holders and supervisors to add any tasks omitted, or to eliminate duplicate tasks, or those so specialized that an entry-level person would not be expected to perform them. Third, a questionnaire was widely distributed to job holders asking them to rate each of the tasks on the basis of its frequency of occurrence, its importance, and the amount of time spent in performing it. The tasks were then clustered into related activities. Finally, each cluster was analyzed by a separate panel of job holders to identify the knowledge, skills, and abilities for the cluster as a whole. This final step defined the test domain and formed the blueprint for item

---

15In Guardians, 49 police officers and 49 supervisors were interviewed in step one; 7 of each reviewed the list in step two. 5,600 police officers were sent the questionnaire.
writing. These five steps are the skeleton of the job-analysis phase of the content-validity approach to job relatedness.\textsuperscript{16}

Given the framework provided by Guardians, let us examine the seven steps the contractor in the Allen v. Alabama case went through to implement the content-validity approach, and the objections plaintiffs raised. Needless to say, the defendants strenuously contended each objection. The steps and the objections are taken directly from Plaintiffs' and Plaintiff-Intervenors' Joint Proposed Findings of Fact and Conclusions of Law (1986).

1. \textit{Outline Development:} To begin to identify the content of each the 35 certification fields, the contractor reviewed state department guidelines and school based curriculum materials for grades K to 12. This review lead to the generation of a topic outline for each certification area. Each outline included 100 to 200 topics.

   Plaintiffs argued that the contractor failed to review curricular material from the state's teacher-training institutions when generating topics for the Basic Professional Studies test (BPS). The BPS was a generic test of professional knowledge and skills which all teachers had to pass regardless of certification area. Thus, plaintiffs argued that the BPS topics lacked any relevant curricular basis. Plaintiffs also argued that the lack of records of any part of any of the topic reviews made an audit of the process impossible.

2. \textit{Objective Development:} To define the limits of each topic and clarify its intent, each topic was rewritten by the contractor as an objective. Curriculum committees composed of eight to ten Alabama educators reviewed and modified the contractor's list of objectives to make them accurate, comprehensive, sufficiently specific, and unbiased.

   Plaintiffs argued that although the members of the panels were described as "experts" in their respective fields, the criteria for their expertise were unavailable. Plaintiffs also contended that the verbs used to define many of the objectives were broad and not operational. Therefore, many of the objectives that survived this stage were ambiguous. Further, the committees did not review the objectives relative to teacher-training curricula. Finally, no information was provided on the racial composition of the panels.

\textsuperscript{16}It is worth noting that the plaintiffs in Guardians challenged the care in which each step of the job analysis was conducted and the Court observed that "With a job analysis of questionable sufficiency, the City then proceeded to the test construction stage."
3. **Job Analysis:** To further determine the job-relatedness of each objective that emerged from step 2, a list of the objectives for each certification area was mailed to current job holders (e.g., teachers, principals and support staff) throughout the state. They were asked whether they had taught or used the content of the objective during the current year or in the past year. They were asked the amount of time they spent using the objectives and the extent to which they considered the objectives to be essential to their teaching or instructional support area. From the "time spent" and "essentiality" scales the contractor derived a composite index for each objective. Based on this index, the objectives were characterized as "Preferred" (P) "Accepted" (A) (moderately job related), and "Not-as-Job Related" (NJR).

Plaintiffs spent a great deal of time assailing the rationale and methodology of the job analysis. They argued that a job analysis should have focused on the functions and tasks teachers are expected to perform, not simply the subject matter objectives perceived to be needed on the job. They argued that the contractor employed two incorrect rating scales; no opinions were sought from teacher educators, supervisors, administrators etc.; many of the objectives were too broad to rate properly; nine of the reviews were based on a response sample under fifteen, six under ten and one as low a two; the composite index was flawed for a number of conceptual and methodological reasons; minority opinions were ignored.

4. **Objective Selection:** The results from step 3 were presented to the subject-area curriculum committees to select the final set of objectives for which test items would be written. The committees chose 35 to 54 objectives per exam.

Plaintiffs argued that the objective-selection process was vague and undocumented; hence, the process could not be audited. Plaintiffs pointed out that all the objectives were presented as job-related to some degree; objectives with an NJR composite ratings were presented to the committee as not as job related rather than as non-job-related. Since some of these NJR objectives survived the selection process and were included on the test, the job analysis was meaningless because it failed to identify and isolate non-job-related objectives.

5. **Item Writing:** The contractor developed a pool of approximately 150 multiple choice items for each content area; from this pool 120 items were to be included on the test, of which 100 were
to be scorable. The contractor checked the items for content validity, accuracy, sufficiency, clarity, and consistency. An item review conference then gave the curriculum committees an opportunity to review and revise the items in the pool.

Plaintiffs argued that the quality of the item writing and the editorial review was poor; the size of the pool was inadequate forcing the inclusion of questionable items; the item review process did not allow sufficient time for a careful consideration of all items; some changes recommended by the curriculum committees were not made; the items were not field tested; separate bias review panels were not convened to screen the items for potential stereotyping or offensiveness; and no statistical item bias analyses were carried out.

6. **Content Validity Review:** New review panels were formed to determine the content validity of each item and to set the cut-score for each test. Each panel member, working independently, was asked to rate each surviving item in the pool as "valid" i.e., an adequate measure of the objective in question, or "not valid." If the item was rated "not valid," the panelist was asked to indicate one of four reasons for the rating (inaccurate content; not a measure of the stated objective; tricky, ambiguous or misleading; or bias). The panelist had the option of not rating an item that contained unfamiliar content by choosing the "I don't know" response. For each item rated "valid," cut-score data were generated by having the panelist decide whether or not entry level teachers should be able to answer the item correctly. Plaintiffs argued that the original rating forms and the data tapes from this review were not available for auditing; and that simply because an item was judged to measure an objective categorized as job related did not make the item itself job related. They attacked the contractor's cut-score methodology, arguing that among other things, it lacked support in the research literature, and that it produced artificially higher cut-scores than established methods.

7. **Test Assembly:** Data from the previous step guided the assembly and preparation of the tests for the first operational administration.

Plaintiffs contended there were a large number of flawed items on the final test (i.e., the stem did not set a clear problem, the distracters were ambiguous, more than one distracter was correct); there were a large number of negatively stated items; items
appeared that measured NJR objectives; items appeared that were never reviewed by subject matter panels; that there were a number of miskeyed items; and item statistics from the first eight administrations signaling possible problems were ignored.

What's wrong with this legal-defensibility approach? Nothing, as long as satisfying the courts is the only objective. However, the essential question is whether the inferences, decisions, and descriptions made about a candidate from his or her performance on the test are inadequately supported by this legalistic approach to validation. Following this tradition of validation in fact defines the kinds of explanations of a score that are sought.

In the remainder of this chapter I examine the inferences that policy makers, the media, and the courts make from teacher certification tests. Second, I describe the mix of evidence needed to support such inferences and critique the shortcomings of present validation techniques. Third, I discuss the tension between exigencies of the applied situation and the need for more thorough validation efforts before the test is used. Fourth, I outline the three types of validity evidence that must be collected during the validation process. Finally, because the cut score triggers the inference or decision, I examine the issue of the validity of the cut score.

THE VALIDITY OF TEACHER-CERTIFICATION TESTS

According to the 1985 Standards for Educational and Psychological Testing, "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (AERA, APA, NCME, 1985, p. 9). Any analysis of validity, therefore, must begin with a description of the actual inferences that people make from a person's score on a teacher certification test.

Inferences and Decisions From Teacher-Certification Tests

How do people interpret a candidate's performance on a teacher certification test—what inferences do they make about the person? Consider the following:
• From an editorial in the Mobile Register:17 U.S. Middle District Judge Myron Thompson has sentenced hundreds of Mobile public school students to an inferior education with his incredible order that the local school board give regular teaching posts to 51 teachers who had failed to pass a basic competency test. . . 

Thompson said, in effect, “No! It doesn’t matter if they can teach or not. Put them in the classrooms!” (“Filling Classrooms,” 1983, p. 4-A).

• From a news story in the Montgomery Advertiser concerning the State Board’s rejection of a settlement it had earlier agreed to in the Allen v. Alabama case: Wednesday, Attorney General Charles Graddick18 blasted the settlement, saying it would lead to “the dumping of 650 incompetent school teachers into the state’s public education system” (Cork, 1985, p. 2A).

• From an editorial in the (Montgomery) Alabama Journal on the Eleventh Circuit’s reinstatement of the settlement in Allen v. Alabama:
The settlement was an utter sell-out of the teacher competency testing program, which sought to do nothing more than keep incompetents—regardless of race—out of the classroom. There is nothing wrong—and nothing discriminatory—about the state’s expecting those who would teach to show certain levels of expertise in their particular fields and a degree of general knowledge. (“Living with error,” 1987 p. 14)

• From the Austin (Texas) American-Statesman: Requiring competency tests for teachers, the [Perot] committee estimated, would rid the public schools of 20,000 incompetent teachers. (Copelin, 1984)

• From a letter to the editor of the Montgomery Advertiser: Putting incompetent teachers into the classroom could prove to be education’s Vietnam. Just as the United States could not win a war with only a half-hearted effort, neither can Alabama win a war on illiteracy with incompetent teachers. (Trotter, 1985, p. 12A)

• From the Amarillo (Texas) News-Globe: Teachers would have to pass a test by June of 1986 to show that they were competent to teach. (“School Reform,” 1984, p. 2A)

---

17This editorial refers to the use of the NTE by the Mobile Alabama public schools to fire experienced teachers who failed to achieve the cut score of 500 or above. In an out-of-court settlement in York v. Mobile the use of the NTE was discontinued. ETS went on record as opposing the school board’s use of the NTE for termination decisions.

18Graddick, a candidate for Governor in the Democratic primary, made the settlement and the certification testing program an issue in his campaign.
From the decision in *United States v. South Carolina* (1977) discussing the issue of false positive and false negative classifications:
If there is a teacher shortage, a relatively high minimum score requirement may mean that some classrooms will be without teachers, and it may be better to provide a less than fully competent teacher than no teacher at all. But to the extent that children are exposed to incompetent teachers, education suffers. (445 F. Supp. 1094 at 1115)

From a January 8, 1980 Alabama State Board of Education resolution:
[T]he State Superintendent of Education and the State Department of Education staff proposed to develop for Board approval . . . criterion-referenced . . . exit professional competency tests in each of the teacher certification areas specified by Board policy, to measure the specific competencies which are considered necessary to successfully teach in classrooms in Alabama schools. . . . (Plaintiffs’ and Plaintiff-Interveners’ 1986, p. 15)

From the test contractor’s *Registration Bulletin* for the Alabama Initial Teacher Certification Testing Program:
An individual’s performance on a test is evaluated against an established level of competence. . . . As stated above, the test items were reviewed for minimum content knowledge competence that (a) practicing Alabama teachers must have in order to successfully teach in the classroom, or (b) if in an instructional support personnel position, must have in order to be successful in the position. The pass/fail scores clearly define those individuals demonstrating sufficient competence in the teaching or instructional support field (Plaintiffs’ and Plaintiff-Interveners’ 1986, p. 53).

Last, but far from least, from an August 23, 1987 news story in the *Savannah News-Press* headlined *School Firings Praised*:
Although five Chatham County teachers are among the ranks of 327 teachers state-wide unable to continue teaching in Georgia, U.S. Education Secretary William Bennett assured the state Saturday that it is setting a good example by barring the teachers from the classroom because they failed a new certification test.

“It is a great thing for the state of Georgia,” Bennett said. “It is a declaration that the state is not willing to have incompetent teachers in the classroom.” (p. 1C)

There is no shortage of analogous examples. What’s clear is that people make inferences from teacher-certification—test scores about a teacher’s competence. Editorial writers, newspaper reporters, politicians, blue-ribbon committee members, State Board
members, judges, ordinary citizens, even test contractors, and yes, the then U.S. Education Secretary all seem to agree that teacher-certification tests measure some aspect of teacher competence. Further, this competence has a referent—classroom performance. Whatever it is, it's necessary for successful teaching.

Additionally, note the implicit—sometimes explicit—a priori conviction that the test in fact correctly distinguishes between competent and incompetent teachers. People accept the test as valid without ever seeing or taking it. As feminists have come to realize, whoever names the world owns it (McFague, 1982) Many people have a literalistic mentality and consequently, the test is simply what the contractor or agency says it is. Naming the test reifies it. If it's called a teacher-certification test, a priori it's an indicator of competence. When people reify a test they are no longer like the Wizard of Oz who knew green glasses made Oz green, they believe that Oz is green (Turbayne, 1962). Naming something can also affect attitudes at a profound level. This affective component explains why many people find it difficult to understand why plaintiffs challenge a teacher-certification test that "obviously" weeds out "incompetents."

Proponents of teacher-competency testing are quick to point out that the competence being measured is necessary, but not sufficient, for successful classroom performance. Passing the test does not mean the teacher will be successful; the test simply doesn't purport to measure all of the knowledge and skills necessary to teach. These tests only measure certain aspects of the job. However, this argument leaves something important unsaid. If the test is a valid measure of some subset of necessary, but not sufficient, knowledge, skills, and abilities, and a person fails it, then it follows that the probability is high that the person is incompetent. Therefore, the person cannot be minimally successful in the classroom. In other words, at least for those labeled incompetent, successful classroom performance is still the ultimate criterion. I now turn to lines of evidence that are needed to substantiate inferences about competence.

Lines of Validity Evidence

Court decisions apart, what kind of evidence is needed to sustain an inference that a candidate is incompetent to perform success-
fully in a classroom and, thereby, to justify the subsequent decision not to certify the person? The 1985 *Standards* point out that traditionally there are three categories of validity evidence: content-related, criterion-related, and construct-related evidence. The *Standards* also state that rigorous distinctions between categories are not possible, and that

An ideal validation includes several types of evidence, which span all three of the traditional categories. Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable quality.¹⁹ (p. 9)

Validation can no longer be a question of one predominant type of validity evidence nor of one predominate method to gather that type of evidence. Rather, the question should be how the three types of validity and various methods of gathering evidence about them should be joined so as to produce a solid, overall line of evidence in support of whatever particular type of inferences or decisions is under consideration.

As we saw, contractors presently are arguing for the single line of content validity evidence to justify the use of teacher-certification tests. How is this all consuming line of evidence gathered? The approach constitutes a closed system of soliciting teacher judgments at key steps in the test construction process (i.e., job analysis, content validation, standard setting). Content-related evidence of job relatedness ultimately comes down to what panels of teachers—about whom we know very little—consider or feel are the skills, knowledge, and abilities needed to function successfully in the classroom (Madaus 1986; Madaus & Pullin, 1987).

The question then becomes “How strong is this judgmental line of evidence?” As far back as 1934, Tyler, describing the practice of

¹⁹Defendants in *Allen v. Alabama* (1985) emphasized the last clause and linked it to the following sentence from the Professional and Occupational Licensure and Certification chapter to argue that criterion-related evidence wasn’t needed and a strong single line of content related evidence was sufficient; “Investigations of criterion-related validity are more problematic in the context of licensure or certification than in any employment settings” (p. 63).
validating IQ tests against teacher's judgments of student intelligence, pointed out that using this judgment approach exclusively means that the test is never more valid than the teachers' judgments. More recently, Haney, Madaus, and Kreitzer (1987) pointed out that item bias is the only area of testing where judgmental methods have been closely compared with empirical methods. They pointed to Jensen's (1980) conclusion based on an extensive review that, "claims of test bias cannot be supported by subjective judgments regarding the item content" (p. 371). They concluded that the literature on judgmental versus statistical methods of bias detection is a real source of disquiet when one tries to justify the judgmental approach exclusively to establish test validity. It is interesting to note that defendants in Allen v. Alabama used this same item bias literature to justify not convening separate panels to examine items for bias, offensiveness, and stereotyping.

Granted, teacher judgments carefully elicited form an important line of validity evidence. However, a single line of evidence based exclusively on opinions is, I submit, insufficient to sustain the types of inferences just described (Madaus & Pullin, 1987). I believe that the validation of teacher-certification tests must include evidence from all three traditional validity categories. The inference made from a teacher-certification test is too complex, the decision not to certify too important, and the potential harm to individuals too great, to be supported by a single line of validity evidence. Validation if done properly is hermeneutic, and to arrive at the meaning behind a test score, all three kinds of evidence are needed. Further, for any of the three lines of evidence there is no one best method for collecting the data; instead multiple methodologies must be employed to better illuminate the question being investigated.

Validation and Applied Testing

Before developing further the reasons for three distinct lines of evidence, I need to raise one other very important point about validation. Validation is an ongoing, additive process of accumulating evidence, not a single study. However, the open endedness of validation and the necessity for multiple lines of evidence using multiple methodologies immediately pose very real problems for contractors. Contractors, after all work in the applied, commer-
cial, competitive, litigious, nonregulated, business world of testing.

As just noted, a fundamental characteristic of validation is a search for the meaning behind a test score. Validation is ultimately a scientific enterprise. It calls for iconoclastic speculation, suspicion, doubt, autonomous judgment, deliberate hard mindedness, open-ended tense enquiry, and a willingness to follow the evidence wherever it may lead. Contractors, however, are basically technicians, not scientists. They are forced to have can-do mentalities, which focus more on delivering a product that will stand up in court than on searching for meaning behind a test score. Contractors have to deliver an operational test by a fixed date, sometimes as quickly as 6 to 8 months from the award of the contract. To make good contractual obligations they rely on a standard, stereotyped technology that they know the courts have accepted. There simply isn’t time to implement an integrated, multifaceted, ongoing validation strategy. Furthermore, contracts are not funded for comprehensive, ongoing validation efforts. Contractors are also acutely aware that pursuing certain lines of evidence is a risky business given the possibility of a legal challenge (Yalow, Collins, & Popham, 1986). Finally, because contractors must produce a product that is “valid” to get paid, they avoid studies that are open to disconfirmation. Simply put, asking disconfirming hypotheses about their product poses a real conflict of interest for contractors. Consequently, habit always triumphs over novelty and they cling to proven techniques and institutionalized formulations that deliver a positive result.

Obviously, in the real world of applied testing, validation cannot be altogether open ended. That’s not the issue. Nor is it an issue of whether to test or not. Tests are potentially too valuable a source of important information about individuals and institutions. The critical practical issues are: (a) How much validity evidence is enough before the test can be used operationally? and (2) How do we get around the conflict of interest inherent in validation in the applied situation?

Determining that a teacher-certification test is ready for use involves balancing ethical and practical issues as well as individual and social interests. On the one hand, contractors and agencies have an ethical duty to provide the proper kinds of evidence to support inferences of competence in the classroom. There is also little doubt that people can be harmed by a test contractor’s product. On the basis of a test score a person is permitted to teach or not; can be stigmatized with the label incompetent; can have his or
herself concept diminished; can be hurt psychologically and financially. Similarly, teacher-training institutions can be damaged by decisions made about them on the basis of test information. On the other hand tests can provide valuable information about individuals and institutions. Policy makers and the public have come to expect that test information be used to help protect society from incompetent teachers and inferior institutions. Finally, there is no such thing as a perfect test. Even the most thoroughly validated test will produce false positive and negative classifications. Inferences are always problematic. Validation offers a reasoned defense for inferences, not proof.

Given these competing factors, the question then becomes, "What mechanism should be used to decide if the test is ready to be used to make high-stakes inferences and decisions?" One analogy, albeit incomplete, is to drug testing. At what point is it safe to permit the operational use of a new drug? The decision isn't left up to the drug company. There are guidelines from the FDA on the studies that need to be done; and the results are evaluated by the FDA before approving the product for commercial use. Testing has no analogous review mechanism. When the contractor delivers the test, it is used to make important decisions. The profession needs to come to grips with this value-laden, political question of when a test is ready to be used operationally. No resolution, I submit, is to be found in the Standards, the EEOC Guidelines, the courts, nor the contractor's proclamation of validity.

The profession must also come to grips with the knotty issue of the conflict of interest contractors find themselves in when validating their own test. Of course the contractor must be involved in validation. The test construction process itself is an integral part of the process. However, test validation must continue well beyond the test-construction phase, and must actively pursue disconfirming hypotheses. The answer to this dilemma is, I believe, an independent test-auditing agency analogous to the FDA. Such an agency would provide contractors with guidelines for leaving a proper audit trail during the test construction phase and would audit this phase. It would provide state agencies with guidelines for validation study designs to test disconfirming hypotheses after the test construction stage, and it would audit this phase as well. These audits would guide the agency in making recommendations to the contracting agencies about whether the test is ready for operational use. The profession must take the lead here or legislatures eventually will. New York's "Truth in Testing" legislation is but an initial shot across testing's bow. An independent, profes-
sionally sponsored, nongovernmental agency is preferable to the governmental FDA model.

Currently, the closest thing to what I have in mind is the ETS internal auditing process. On assuming the presidency of ETS in 1981, Gregory Anrig initiated an audit process to assure that each of the company’s products was in compliance with the ETS Standards of Quality and Fairness. An Officer of Corporate Quality Assurance was established to oversee the audits and the implementation of any recommendations coming out of an audit. Each ETS program is reviewed at least every 3 years. A Visiting Committee annually reviews the audit process itself and examines the findings from the yearly audits. The Visiting Committee reports directly to the ETS Trustee Committee on Public Responsibility.

Interestingly, this past year the NTE program was audited. The Visiting Committee observed that

The role of the tests that comprise the NTE Programs—in qualifying candidates for admission to colleges of education, certification into the profession, and the identification for career ladder and master teaching programs—along with increasing public concern for the quality of our teaching force, guarantees them a place as one of the more controversial testing programs within ETS. (Report of the 1987 ETS Visiting Committee, 1987, p. 3)

The Visiting Committee then raised a series of questions about the NTE programs tests:

- Are the NTE Programs contributing to the improvement of teacher education and practice?
- Do the NTE Programs have a negative effect on the recruitment of minorities into the teaching profession?
- Are the NTE Programs tests valid for the inferences commonly drawn from their uses?
- Can ETS continue to rely on studies undertaken by state-level clients using the NTE Programs tests as their primary source of validity? (p. 9)

The Committee went on to recommend that:

In light of [its] concerns, and those of others, regarding current NTE Programs tests, along with the changes occurring in teacher educa-

20The Standards of Quality and Fairness are a set of standards that, in my opinion, go beyond the 1985 Standards in outlining criteria for test development and use.
tion and the teaching profession, . . . the talents, technical expertise, and resources of ETS be directed toward researching and developing new methods of teacher evaluation. [That] would: (1) respect the complexities of the work of teachers at different grade levels and in different subject or specialty areas, (2) be valid measures of the knowledge and skills teachers need to perform competently [italics added], and (3) not discriminate against minorities (p. 9).

The audit process is taken very seriously at ETS and is a commendable effort at public responsibility. However, it is not the answer for the entire industry. It lacks the authority of an external independent auditing body. Unfortunately, it certainly would be perceived by some critics of testing as self-serving. Further, it is beyond the resources of smaller contractors. Nonetheless, it provides an excellent model for what an independent agency might do.

I now turn to a closer examination of the need for construct, content, and criterion-related evidence, problems with current validation efforts, and the issue of the validity of the cut score.

Construct-Related Evidence

Inferences people make about a teacher’s competence are inferences about a construct. Competence is a construct, and therefore construct validation is essential for any teacher certification test. The construct, competence in the classroom, must be “embedded in a conceptual framework, no matter how imperfect that framework may be” (AERA, APA, NCME, 1985, p. 9). Development of a conceptual framework for teacher competence must involve a functional analysis of what minimally competent teachers actually do in their classrooms. This analysis must start from below with the work of teachers and move then to construct definition. Further, a functional analysis must be specific to the certification area in question. Teaching is not a generic profession. Only by functionally analyzing the different certification areas can the subset of skills, knowledge, and abilities common to all teaching areas be identified. A functional approach would also answer the question of whether a generic professional-skills test required for certification of all teachers makes any sense. I feel that extant generic, multiple-choice teacher-certification tests make little sense, and are simply not valid (Madaus & Pullin, 1987).

Although a theoretical, or structural analysis can also be helpful, the functional, area-by-area approach to defining job competencies should shed more light on the critical ingredients needed to
be minimally successful in the classroom. Not all of the ingredients that emerge—in fact probably very few—can be measured by the current, administratively convenient multiple-choice technology. However, by illuminating what is possible within the limits of this technology the first step in the construct validation of such tests has begun.

The functional approach goes beyond the current practice of mailing lists of objectives to teachers for them to rate, although, if done carefully, a survey of teachers' opinions on the aspects identified from the functional analysis contributes to the web of construct-related evidence. Content- and criterion-related evidence, although contributing to construct validation, is insufficient. The 1985 Standards remind us that: The process of compiling construct-related evidence for test validity starts with test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicates the meaning of the test score [italics added]. (AERA, APA, NCME, 1985, p. 10) The 1985 Standards describes the process of compiling construct-related evidence, along with possible sources of, and techniques for collecting such evidence. These suggestions must be taken seriously when validating teacher certification tests.

Content-Related Evidence

As just noted, content validation is the favorite, exclusive approach of contractors building teacher-certification tests. Content-related evidence is a necessary but not sufficient ingredient in validating teacher-certification tests. However, if we are searching for the meaning behind a score on a such a test, then we need to reexamine the techniques currently used to solicit this evidence.

Cronbach, writing in the context of program evaluation, calls for "a bundle of studies [using] different techniques to examine subquestions" (Cronbach & et al 1980, p. 73). Within the same evaluation context Cooley and Bickel (1986) pointed out that multiple approaches to examining a question deepen an understanding of the phenomenon under investigation. This is certainly good advice for those validating teacher-certification tests. Current content validation methodology has become institutionalized, a cocoon for confirmation. Disconfirming hypotheses about content relevance are not examined. To the degree that expert opinion varies as a function of the way it is elicited or quantified, inferences made from scores are suspect.

For content-related evidence to be believable, we need to learn a
lot more about things we now take for granted such as: how well panel members understand the tasks they are asked to do; how much response set influences the process as panel members become absorbed in the process; how much the meaning of directions are interpreted and prejudged by individuals according to their own framework; how easily panel members recognize flaws in items; and how time constraints affect the task. We need to learn more about how teacher opinion varies as a function of the composition of the panels, of the wording of the judgment question put to them, and of the way the results are quantified. The final set of items chosen to measure successful job performance should be robust enough to emerge consistently under different conditions of gathering judgments about their relevance. Madaus and Pullin (1987) offered a sample of “what if questions” that test the robustness of teacher judgments about test content.

There is an interesting paradox associated with the current content-validation approach. Many teacher certification test items that successfully survive the content-validation panel review are roundly criticized by test reviewers. This contradictory phenomenon spans the decades and cuts across a number of different tests. (e.g., Buros, 1938; 1953; 1959; 1965; 1972; 1978; Darling-Hammond, 1986; Koerner 1963; Madaus, 1986; Madaus & Pullin, 1987; Melnick, 1987; Melnick & Pullin, 1987; and if it is ever released the trial transcript in Allen v. Alabama; see Haney et al., 1987, for former examples and a more detailed discussion). Further, this criticism isn’t limited to one or two bad items that may have slipped by accidently. It blankets a host of items. Mitchell’s comments on the item quality of the 1976 NTE captures the flavor of this history of criticism:

The items on the Professional Education test should be carefully scrutinized by potential users. Some of the items seem to smack of professional shibboleths, others have shaky research foundations, others seem to reflect the values of the writer more than the substance of the field, others are simplistic, others are combinations of these. Some items of course, are wholly acceptable. But if the items are taken as a whole, it is in fact difficult to believe that these items adequately sample from the professional preparation provided by most teacher training programs. (p. 517)

The regularity and similarity of this item criticism, across the years and across different contractors’ tests calls into question current content-related evidence, and illuminates the need for a careful evaluation of the traditional techniques used by contractors. At the very least this history of item criticism is indicative of
the fact that typical content review panels simply aren’t sensitive to many problems inherent in the objectives and multiple-choice items they are reviewing.

Questionable items are a serious threat to validity when the inferences about a person are triggered automatically by a cut score on the test. Each item—good, bad, or indifferent—contributes one point to a person’s score. To illustrate the damage done by the presence of questionable items, plaintiffs in Allen v. Alabama constructed what became referred to in the trial as failure charts for each test.\textsuperscript{21} Persons who failed the test were identified. Their answers to the questionable items were then recorded. This gave a crude estimate of the extent to which answering a questionable item incorrectly contributed to falling below the cut score. A considerable number of candidates who were denied certification fell below the cut score due in part to answering a subset of the questionable items incorrectly. For those close to the cut score one or two questionable items may have made the difference between passing and failing. Of course we don’t know how these candidates would have answered flawless items measuring the same objective had they appeared. That’s not the point. The point is that the inference of incompetence made about these candidates was corrupted by the presence of questionable items.

In summary, content-related evidence is essential in the overall validation of teacher-certification tests. However, the track record for content-validation studies of teacher-certification tests to date is poor. The meaning behind a test score is not enhanced from results obtained using present content-validity techniques. The results do not contribute to construct validation. The methodology endorses many dubious items as job related. We need to begin to carefully reassess the methodologies that have evolved to collect such information. We must ask the extent to which the results, on which everything hinges at the present time, are method dependent. To obtain sound overall content-related evidence we need to begin to employ different techniques with different types of panels.

\textbf{Criterion-Related Evidence}

If teacher-certification tests measure important aspects of successful classroom performance, scores on such tests should corre-

\textsuperscript{21}Items were labeled \textit{questionable} if there were judged ambiguous, miskeyed, having no best answer, measuring trivial content, or context dependent for an answer. Because the items are protected by court order, and the testimony was in Camera, I am unable to offer concrete examples. Defendants challenged some of the items characterized as questionable, but not the majority.
late with criterion measures of such performance. The record here isn’t good either. Reviews of attempts to correlate scores on teacher-certification exams and performance in the classroom show that “teacher tests have little if any power to predict how well people perform as teachers, whether that performance is judged by ratings of college supervisory personnel, ratings by principals, student ratings or achievement gains made by students taught” (Haney et al., 1987, p. 199).

In the 1930s Ralph Tyler reminded the measurement community that indirect (multiple choice) indicators must be validated against direct indicators of the construct of interest. Contractors building teacher-certification tests seem to have forgotten this advice or dismissed it as impractical. Instead, they argue that adequate criterion measures of successful classroom performance simply aren’t available. Although this may be true, it is an admission that contractors know next to nothing about the construct that they are trying to measure. This admission also undercuts the accuracy of the so called job analysis portion of the content-validity approach. A proper functional analysis of the job at the outset should produce direct indicators of the construct and also suggest ways to measure them.

The national movement to hold teachers accountable should eventually lead to the development of better, more systematic techniques to evaluate educational personnel. Decisions about such things as merit pay, tenure, and career ladders cry out for such criteria. What emerges from new efforts to evaluate teachers may offer a subset of criteria against which performance on teacher-certification tests can eventually be validated.

Another argument proffered as to why criterion-related evidence isn’t needed for teacher-certification tests is that such evidence isn’t collected for licensure tests in other professions such as law, medicine, or nursing. One reason for this absence is that candidates in those other professions go into a wide range of very different types of jobs. For example, a candidate who passes the bar exam might practice real estate law, corporate law, criminal law, and so forth. There are so many job paths open to the new lawyer that the criterion problem is simply insurmountable. Or so the argument goes.

The situation in education is different. A teacher applies for a very specific certification; one tied to a grade level, subject field, or speciality area. Unlike lawyers or doctors, teachers don’t get a generic license to practice. Thus, for each test particular to a certification area, individualized criteria of competence are needed and should be possible to obtain. With 30 or more separate cer-
tification areas this won't be an easy task. However, we have to begin. And despite the difficulty, I feel that criterion-related evidence, whether predictive or concurrent, is absolutely necessary to validate properly the commonly made inferences about minimally successful classroom performance. Without such evidence we are locked into the closed circle of judgmental evidence of content-relevance.

The Cut Score

Essential to any validation of teacher-certification tests are studies of the validity of the cut score. It is the cut score, after all, that triggers both the inferences about a person's competence, and the decision to certify or not. However, to the best of my knowledge, there have been no attempts to validate the cut scores used with teacher-certification tests. Presently cut scores are set by having panels make judgments about the expected performance of minimally competent teachers on each item. The results are then aggregated to arrive at a cut score. We have already discussed problems associated with the closed system of relying exclusively on panel plebiscites about objectives or items. Those same problems apply with equal force to the cut-score process.

We already know that cut scores on teacher-certification tests are method dependent (Berk, 1986, p. 163; Goertz & Pitcher, 1985; Poggio, Glasnap, Miller, Tollefson, & Jaeger, 1986). (This body of literature is a further reason to investigate the method dependence of current content-validation practices.) The practical implications of this dependence is far from trivial. Using the Goertz and Pitcher data, Haney et al. (1987) calculated the proportion of examinees in the 1982–1984 national sample of NTE candidates who would pass or fail using the highest and the lowest cut scores found across the various states using the tests. The figures ranged from 75% to 95% passing the Communication Skills subtest, 67% to 91% passing the General Knowledge subtest, and 78% to 98% passing the Professional Knowledge subtest. These data caused Haney et al. (1987) to ask "Is there really this much variation from one state to the next, in the skills and knowledge level needed to be a minimally successful teacher?" (p. 201–202) Despite questions like these about the underlying construct of teacher competence, judgmental cut-score methodology presently rules the roost. We need to validate any cut score through an empirical examination of the degree to which it correctly separates those with insufficient knowledge and skill from those with adequate prerequisites (Madaus, 1986).
It is undeniable that at some point a judgment has to be made about where to set the cut score. However, such a judgment should be informed by data on the number of possible false positives and false negatives associated with the selection. Mass screening techniques in medicine provide an analogy for what is needed. To screen hospital blood-bank donors for possible anicteric hepatitis, measures of serum enzyme are employed (Colton, 1974). To set a cut score for decisions about the blood's acceptability pathologists plotted the distributions of the serum glutamic pyruvic transaminase (SGPT) (transformed to the logarithm to base 10) of healthy individuals and those with hepatocellular damage. Informed cost-benefit judgments can then be made about where in the overlap zone to place the cut score to minimize either of these unavoidable misclassifications.

The medical example just described is, of course, nothing more than the contrasting-groups method used in some minimum competency graduation testing programs. Such a technique could be used for teacher certification tests. Naturally, the underlying scale of the construct teacher competence will not be physical. But the contrasting-groups method for cut-score setting could be employed if serious construct and criterion-related evidence were available.

There is also an analytical approach to the problem of setting a cut score on a teacher certification test. It is based on the fact that a relatively small number of persons tested are actually incompetent. Haney et al. (1987) found that estimates of the percentage of incompetents in the teacher corps are similar to estimates of incompetence associated with other professions—about 10%. When base rates of incompetence are low, even using a screening test that is highly accurate, you run the risk of doing great damage to candidates falsely labeled incompetent. I describe three examples that illustrate the point.

Recently in the APA Monitor John Bales (1987) reported on the testimony of psychologist Edward Katin to a congressional panel investigating the use of the polygraph in employment settings. Katin pointed out that a relatively small number of people tested by a polygraph will be dishonest or deceptive. Using available estimates of 85% accuracy for the polygraph, and a 10% "dishonest" estimate, he provided an example of what happens under these conditions when 1,000 employees are screened. The polygraph would identify 85 of the 100 dishonest employees, but would misidentify as dishonest 15% or 135 honest employees. Of the 220 "suspects" (85 + 135), 61% are innocent. He concluded that it is a
mathematical reality that the majority of suspects are, in fact, innocent. Nonetheless, a cloud of suspicion blankets these innocents.

Light (1987) offered another provocative example from widespread screening for drugs or AIDS. He asked the simple question “Of all the people that the drug test (or an AIDS test) classifies as having drugs (or AIDS) what proportion really has it?” (p. 51). He used the following three figures to answer it: (a) an estimate of 1% for the proportion of people who actually have AIDS or routinely take drugs; (b) an estimate of 95% for the accuracy of a screening test to detect drugs or the AIDS virus when a person really has them; and (c) an estimate of 95% for the accuracy of the test to detect the absence of drugs or the AIDS virus. Using these three estimates he found that “of all the people the test categorizes as having drugs (or AIDS), only 15% really do!” (p. 51). Even a test that is 100% accurate in detecting people who do take drugs (or have AIDS) hardly changes the situation. Under this assumption, 17% would be misidentified; 83% of all those labeled as positive would be falsely accused. Needless to say, a false accusation in such a matter can seriously harm a person.

But what have hepatitis screening, polygraphs, and drug testing, all involving physical measures, to do with teacher-certification testing? The situations are analogous in that the potential for harm to those falsely labeled incompetent is great, and the proportion of teacher candidates who are actually incompetent is probably low—10%. This is particularly true when you consider the fact that few people in testing or education could claim anything near the accuracy for teacher tests ascribed to the polygraph, urine, or blood tests.

Haney, et al. (1987) simulated what might be happening in the teacher-certification-testing situation. They assumed that 10% of the candidates are, in fact, incompetent. Based on a review of the literature they estimated a correlation of .20 between present certification tests and a measure of teacher quality corrected for unreliability. They then used the Taylor and Russell (1939) tables to estimate a 1% increase in selection efficiency in using the test over random selection.

From the state's point of view, such a small increase in selection efficiency might appear worthwhile. However, it hides the balance of correct and incorrect decisions made using the test. To clarify this balance, they carried out a simulation. They assumed an estimate of incompetence of 10%, and a correlation of 0.20 between the test and criterion measure. Further, they assumed both the test and
the measure of teacher quality were normally distributed. Next they set the cut score on the test and the measure of teacher quality at 1.28 standard deviations below the mean. Using a Monte Carlo simulation they estimated: (a) the correct acceptances (above the cut score on both variables); (b) the correct rejections (below the cut score on both variables); (c) the false acceptances (above the cut score on the test, but below on the criterion); and (d) the false rejections (below the cut score on the test, but above the cut score on the criterion). As expected, about 10% of the cases were below the cut score on the test and on the criterion. However, only 1.8 out of 100 cases were below the cut score on both measures. Moreover, they found that more than 80% of the cases below the cut score on the test were above the cut score on the criterion. In other words, more than 80% of the rejections based on the test would be false ones. The use of a test for widespread screening of teachers (given the previously described assumptions) identifies less than 20% of the candidates who truly fall below the cut score on the real criterion of interest, competence in the classroom. A success rate of 20% must be balanced against the staggering cost of an 80% false negative rate.

These three examples, I think, argue for the need for both analytic and empirical investigations of the validity of any cut score used for the widespread screening of teacher candidates. We can’t continue to rely exclusively on teacher judgments about individual items as the exclusive basis for setting a cut-score—not if we are serious about making inferences about competence in the classroom.

CONCLUSION

The courts are a world unto themselves, a world that gives contractors a relatively simple formula for validating teacher certification tests. Although contractors need to be cognizant of judicial opinions about employment tests, they should acknowledge that a legal definition of validity is inadequate. A legally defensible certification test is presently obtained at the expense of a proper search for the meaning behind scores on these tests. It isn’t easy to properly validate certification tests in the applied situation. But in the end, if the test survived a proper validation effort, it would be legally defensible.

It is an inescapable fact of life that all manner of people use the tests to make inferences about potential competence in the class-
room. On the basis of these inferences, state agencies make decisions that can adversely affect the careers of numerous candidates. The nature of these inferences demands construct, content, and criterion-related evidence for support. If we are serious about keeping incompetents out of the classroom then we must get a lot more serious about comprehensive test validation. Otherwise, teacher-certification testing is nothing more than a slick public relations ploy.

The dilemma of the ongoing nature of validation and the exigencies of the applied situation must be addressed—and quickly. Guidelines for when a test has been validated sufficiently for operational use must be developed. The profession must come to grips with the conflict of interest that is present when test contractors are the only validators of their own secure products, particularly when they approach validation exclusively from the perspective of defending their product in court. Some independent, mutually respected agency is needed to audit contractors’ work during test construction and to apply professional standards for validation after test construction is over.

I opened the chapter with the validation metaphor of the snark hunt from Lewis Carroll. I’d like to close with a paragraph from the preface to *The Hunting of The Snark* as metaphor for what the quest for a legally defensible test is doing to teacher certification testing:

The Bellman, who was almost morbidly sensitive about appearances, used to have the bowsprit unshipped once or twice a week to be revarnished; and it more than once happened, when the time came for replacing it, that no one on board could remember which end of the ship it belonged to. They knew it was not of the slightest use to appeal to the Bellman about it—he would only refer to his Naval Code, and read out in pathetic tones Admiralty Instructions which none of them had ever been able to understand—so it generally ended in its being fastened on, any how, across the rudder. The helmsman used to stand by with tears in his eyes: he knew it was all wrong, but alas! Rule 42 of the Code, “*No one shall speak to the Man at the Helm,*” had been completed by the Bellman himself with the words “*and the Man at the helm shall speak to no one.*” So remonstrance was impossible, and no steering could be done till the next varnishing day. During these bewildering intervals the ship usually sailed backwards. (p. 8–9)

Teacher certification testing currently is a rudderless enterprise, sailing backwards through a legal sea of codes and court opinion.
It's time to repair, and put in its correct place the rudder of test validity.

ACKNOWLEDGMENTS

My thanks to my generous colleagues at Boston College, Walt Haney, Diana Pullin, Joseph Pedulla, and Amelia Kreitzer for their help, ideas, suggestions, and reactions. My thanks also to Lee J. Cronbach for his insightful reactions. Errors of fact or interpretation and the opinions expressed are solely my responsibility.

REFERENCES

Contreras v. City of Los Angeles, 656 F.2d 1267, 25 FEP 866 and 29 FEP 1045 (9th Cir. 1981), cert. denied, 455 U.S. 1021 (1982).


Georgia Association of Educators v. State of Georgia, Civil Action No. C86-2234A (U.S D.C Atlanta.)


Wigdor, Alexandra. (1982). Psychological testing and the law of employment dis-

