# Considerations related to setting cut scores for teacher tests

Terry Hibpshman
Kentucky Education Professional Standards Board
October, 2004

## A note about the organization of this paper

This is a paper intended to advise policymakers and other interested parties about the probable consequences of alternatives for setting cut scores for teacher certification tests. It deals with a question – the use of tests to determine who shall be licensed to teach – that involves a complex nexus of policy, legal, and technical issues. The paper, in addition to presenting analysis about the consequences of changes in cut scores, also presents a review of previous research and comment on the subject and the results of original data analysis relative to the Kentucky testing regime

Many authors in the legal, educational policy, psychometric, and research communities have a great deal to say on this subject, and this presents a problem. The conventions for citing previous work and for describing the technical characteristics of the studies contained herein are likely to be tedious and distracting for the audience – education policymakers – for whom the paper is intended. Yet it would not be appropriate to just assert the truth of matters which require proper documentation, and this presents a conundrum: how to produce a readable paper without denying others proper credit for their work, or subjecting the methodology of my own work to appropriate criticism?

My solution is to organize the paper into three sections. In Section 1, I discuss the issues with a minimum of scholarly impedimenta. In section 2, I present the research, legal, and policy background as represented in the vast literature on the subject. In section 3, I present the methodology used in original statistical studies conducted by EPSB relevant to the use of teacher certification tests in Kentucky. Readers who are solely interested in the policy issues can thus confine their reading to the first section, with the assurance that the analysis and conclusions contained therein have an adequate basis.

# 1. Executive Summary

## Background

It is a well-established principle that the various states have an interest in the qualifications of persons in many different occupations and a responsibility to assure that such persons are competent. All states require that physicians, attorneys, and many other types of workers - including teachers - obtain a license before they can legally practice their profession. The licensure process is almost exclusively a matter for individual states, and although states often collaborate in the setting of standards or the development of licensure procedures, each is free to determine who will be licensed, and on what basis. In Kentucky, authority to issue teacher certificates is vested in the Education Professional Standards Board (EPSB). The purview of EPSB includes not only the certification of teachers but also the regulation of teacher training programs, disciplining of teachers, and direct provision of teacher education and professional development via web-based instruction.

The use of professional licensure tests is also well established. All states require licensure tests for at least some occupations, and 45 of the 50 states currently require tests for teacher certification. The use of teacher certification tests has a long history dating back to the nineteenth century, but the current testing regime had its beginnings in the decade of the 1980's, when national commissions came to the conclusion that the overall quality of teachers nationally was less than desirable. Teacher certification tests were seen as a means of raising the quality of teaching by assuring that teachers had minimum levels of literacy and content knowledge. The use of tests for these purposes began in the South and quickly spread nationwide.

That states have a right to require certification of teachers is rarely disputed, but particular requirements related to certification – including especially testing – are sometimes controversial. Testing by its nature must distinguish between groups of individuals, those who pass and those who do not, and persons who fail often believe that the tests did not fairly measure their teaching ability. This is understandable. Preparation for teaching takes several years and involves the expenditure of a great deal of personal resources, and the State requires that an individual who wants to teach meet a number of requirements before he or she can even be admitted to teacher training. To someone who has met all the requirements and expended personal resources with the expectation that a teaching certificate would result, it is a rude shock to have the way barred at the end by failure to pass a test. Both the validity of tests and the method of their administration are subject to challenge, and an individual who has failed to be certified has an incentive to challenge them.

Thus, while teacher certification is a legitimate function of state government and the use of tests is a legitimate and popular mechanism for selecting candidates for certification, the question of how tests may best be used for this purpose ultimately has legal implications. As the use of tests for teacher certification and for other high-stakes purposes has proliferated around the country, so have lawsuits. The lawsuits have served to define both the limits of use of tests and methodologies for developing testing programs.

The legal implications of test use, along with technical considerations, ultimately must drive how we set cut scores. Although the setting of cut scores is an arbitrary matter in the sense that we must select from among a number of alternatives with varying and sometimes contradictory consequences, the existing case law as well as the statutory and regulatory base on which it rests require that the process of determining cut scores not be done capriciously and that we have a reasoned and professionally defensible rationale for the levels we select. This requires that we have an awareness of three bodies of information: the statutory and regulatory base related to test use, technical considerations related to the development and use of tests, and guidelines established in case law for the application of these other two bodies of knowledge.

The statutory and regulatory basis

No statute specifically regulates the use of tests for teacher certification or for any other purpose, but because test scores affect the lives of people, general constitutional and statutory provisions apply. Two provisions account for most of the legal cases related to the use of tests either for occupational licensure or for other purposes, the 14th Amendment to the Constitution, and sections VI and VII of the Civil Rights Act of 1964.

The relevant language of the 14th Amendment is found in Section 1:

> No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

Whether or not teacher certification is a property right depends on whether an individual is already certified. The relevant case law on the subject holds that a license granted by the state may be deemed a property right when it can be withdrawn for cause. Case law also makes it clear that the expectation of a future benefit – as when an individual who has never been certified is seeking certification – is not a property right. Thus certification tests, because they are given before an individual is granted a particular certificate, cannot form the basis for a cause of action for deprivation of a property interest under the 14th amendment. Such cases can arise on liberty grounds. The enumeration of those rights that individuals have includes among other things the right to practice one's chosen profession or to earn a living, and when lawsuits arise over teacher tests under the 14th Amendment, these are likely to be the grounds. Note that the terms of the 14th Amendment do not allow an individual to bring action merely because a state policy is unfair: it must deprive them of either liberty or property.

Title VI of the Civil Rights Act of 1964 states in part:

> SEC. 601. No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance.

The provisions of Title VII are too lengthy and complex to repeat here, but in brief they forbid employment practices that discriminate on the basis on an individual's race, color, religion, sex, or national origin. The difference between Title VI and Title VII is that Title VI relates to any program that receives federal financial assistance, while Title VII relates to employment specifically.

Many lawsuits have arisen over the use of tests in employment situations, including the use of teacher tests, under the provisions of Titles VI and VII. The use of teacher tests is subject to the findings of these cases when they relate to teacher tests in particular, and also when they relate to the use of employment tests as a general issue.

The principal regulatory basis for determining how tests may be used in employment situations is given in the Code of Federal Regulations, Title 29, Volume 4, part 1607. This very complex regulation, which is used uniformly by a number of federal regulatory agencies, including especially the Equal Employment Opportunity Commission and the Department of Labor, specifies in detail a number of requirements for test use including both procedural and technical rules. The most important provision and by far the most often used in legal action is the "four-fifths rule":

> 1607.4.D. A selection rate for any race, sex, or ethnic group which is less than four-fifths . . . (or eighty percent) of the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, although a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

This rule provides a point at which test usage may be viewed as discriminatory: when members of a protected group pass the test at less than 80% of the rate at which the group (usually white) with the highest average scores passes. This is the principle of "disparate impact." Although this principle is not definitive, and additional considerations (which will be discussed below) apply, it makes it possible for a plaintiff to establish a prima facie case for discriminatory effect.

Another provision of 29 CFR is important:

> 1607.5.H. Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force.

Thus 29 CFR requires two things: that a test be anchored in some reasonable expectation about job performance, and that there not be evidence of disparate impact. The language of 29 CFR is very explicit about how these things should be accomplished, but as we will see, the courts usually give test administrators a great deal more latitude than the regulation would seem to grant.

The basis for teacher testing in Kentucky is controlled by Title 16 of the Kentucky Administrative Regulations, part 6:010, which is authorized by Kentucky Revised Statutes 161.030(3) and (4); and by other regulations to be detailed below relative to teacher preparation. 16 KAR 6:010 specifies which tests will be required for each of the various certificates issued by EPSB, and sets cut scores. Most persons who desire to be

certified in Kentucky must take both a pedagogy test and specialty test in their content area.[1] A passing score is deemed valid for a period of five years, and an individual who fails may retake the test multiple times. Most of the tests used by EPSB are developed and administered by the Educational Testing Service (ETS), although two were developed by EPSB.

It is important to remember that in addition to certification testing, EPSB specifies other prerequisites for entry into teaching. The first set of these is elaborated in 16 KAR 5:020, the administrative regulation that controls admission to teacher preparation programs. 16 KAR 5:020 requires *inter alia* that approved teacher preparation programs have a plan for selecting teacher candidates that includes a test of academic proficiency. Acceptable measures of academic proficiency include the Praxis I tests developed and administered by ETS; college admissions examinations such as the ACT, SAT, or GRE; other unspecified assessments that measure academic proficiency; and grade point average of college work completed prior to program admission. The regulation sets a minimum acceptable grade point average and cut scores for most of the tests deemed acceptable.

After an individual has been granted initial certification (via a "provisional internship certificate," valid for only one year), he or she is required to serve for one year as an intern. Governed by 16 KAR 7:010, the internship program requires that teacher interns be observed and assessed three times during the internship year by a committee consisting of a principal, a teacher certified in the intern's content area (a "resource teacher"), and a teacher educator. Persons who receive less than satisfactory scores on these assessments fail the internship, and are not issued a professional certificate. In most cases, unsuccessful interns may repeat the internship one time, and there is a process for administrative appeal of a negative internship assessment.

EPSB thus requires a three-stage assessment process, each stage of which is designed to eliminate unsuitable persons from the teaching profession. Although the third, internship, stage is not explicitly designed to weed out persons with academic and content knowledge deficiencies, knowledge of content is one of the criteria on which interns are judged, and an individual whose content knowledge is deficient would be expected to fail the internship for that reason.

There is one final consideration related to the use of tests for certification in Kentucky. 16 KAR 2:180 provides that an individual who is otherwise qualified but fails one or more certification tests may, upon receipt of a request from a public district willing to employ the individual, be issued a one year conditional certificate. The requesting district is required to submit a support plan for remediating the prospective teacher's deficiencies, and the certificate may not be renewed after the one-year issue period is over.

## Technical Considerations

Procedures for development of educational and psychological tests are well established. The field of knowledge devoted to test development, known as *psychometrics*, first developed in the late nineteenth century and has undergone considerable refinement since. The field has developed rapidly since the early 1970's,

[1] Special education teachers do not take a separate pedagogy test.

when electronic computers first became widely available, and as a result authors in the field often talk of "classical test theory" (i.e., the theory used most widely before about 1970), and "modern test theory." Classical test theory in general can be viewed as a special case of modern test theory, just as Newtonian mechanics can be seen as a special case of Relativity.

Either the classical or modern approach is valuable and both are still in use, but there are differences in emphasis on the types of inference necessary to demonstrate the value of a test. Because modern theory is much more complex than classical theory, and because persons outside the field usually have some background gleaned through survey courses in college or from some other nontechnical source, conceptions of test construction requirements held by laymen – and particularly federal regulators and the courts – tend to be framed in terms of classical test theory. This sometimes causes difficulties when persons imagine that a particular feature of test construction is canonical when in fact it may not be.

The "Bible" of test construction is the *Standards for Educational and Psychological Tests*, published jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. This reference describes recognized minimum requirements for the development of tests, appropriate ways of using tests, and methods for establishing that tests are well constructed. The *Standards* are widely respected, are referenced in 29 CFR, and appear in a number of court decisions.

Test construction as established in the *Standards* and by the customary practices of psychometricians requires that two features of a test be established, reliability and validity. Although the literature and common usage of these terms seem to imply that there are a number of different types of reliability and validity, logically and mathematically these two measures of the quality of tests reduce to two ideas:

1. A test score should be a consistent measure of some trait (reliability)[2]
2. A test should have proven value for some particular purpose (validity)[3]

The highly technical procedures for estimating reliability are not germane to our purpose and will not be discussed here. Two matters relating to reliability are of importance. First, reliability is an essential precondition for establishing validity: mathematically, a test's validity is bounded by its reliability. Secondly, an index derived

[2] One often hears reliability described as a measure of a test's repeatability. This is not precisely true. The extent to which a test score is repeatable is actually known as *stability*. Formally, test reliability is a measure of the amount of a test score that is accounted-for by *true score variance*, the underlying trait measured by the test. In general, a reliable test will be stable, but there are exceptions, such as tests of mental health functioning, where a highly reliable test may produce scores that vary widely over time, as the mental health status of an individual changes. In most cases we want test scores to be stable, but in the case of mental health tests, a high level of stability would greatly reduce the value of the test. In the case of teacher tests, too much stability might not be a good thing either, since we might want a test that responds to changes in teacher performance with time.

[3] Validity is often inaccurately described as the ability of a test to predict performance on some criterion, but in fact it is often very difficult to establish what, if anything, a test score predicts. This is one of the fundamental problems in testing for employment purposes, where the nature of adequate performance is always at least a bit murky. Prediction of performance on a criterion is one type of inference that establishes validity, but is by no means always either a necessary or sufficient condition.

from reliability studies, the *standard error of measurement* ($S_{EM}$), is of importance in setting and evaluating cut scores. The $S_{EM}$ is a measure of the amount of uncertainty in a test score that describes a region into which an individual's true score[4] can be expected to fall. This becomes important in setting cut scores, since wherever we set the cut score, there will be some number of individuals who might otherwise pass the test, but fail to do so for reasons not related to the trait measured by the test. Of course there are also persons who should really have failed the test, but pass for reasons having nothing to do with the trait of interest. These two types of misclassification are known respectively as *false negatives* and *false positives*. Given the imperfect nature of tests, misclassifications of these two types are inevitable. In principle we want to minimize both types of misclassification, but in practice there is often a tradeoff between them, so that if we choose a method that minimizes one, we cannot simultaneously minimize the other. This tradeoff is often a consideration in the process of setting cut scores.

Historically, three different types of evidence have been used to make inferences about test validity. We might infer that a test is a valid measure of some domain of knowledge because we have evidence that it accurately represents the content of the domain (content validity). We might infer that a test is a valid measure of something useful because it predicts some performance, either on the job or on another test (criterion-related validity). Or we might believe that a test is a valid measure because we can demonstrate that it is a good exemplar of some theoretical entity (construct validity). The rules in 29 CFR seem to imply that all three types of evidence must be adduced in order to establish a test's validity, but in fact this view is neither accepted by the majority of practicing psychometricians nor required by the courts.[5] The problem, as established in countless analyses by psychometricians, is that criterion-related validity is virtually impossible to establish in many cases because adequate measures of performance are difficult to obtain, and sometimes it is difficult to arrive at an operational definition of good performance. This is especially the case with teaching, where there are numerous conflicting ideas about the nature of good practice, mostly expressed as philosophical statements rather than measurable behaviors of teachers.

A test is never really valid in any abstract sense. Validity in modern terms represents a chain of inference establishing that a test provides useful information for a particular purpose. In the case of teacher tests, the established purpose is to select individuals who have a minimum level of academic proficiency and content knowledge to be presumed capable of delivering education to children in the public schools. This is to be distinguished from a similar but conflicting purpose, selection of the best candidates from a large pool of applicants for a limited number of positions. Either purpose attempts to improve the overall quality of persons selected by the test, but the former does so by setting a minimum standard of quality, while the latter attempts to select individuals with maximum levels of the quality of interest.

It is widely believed that academic proficiency and content knowledge are essential in determining who will be a good teacher, but in fact research studies provide at best weak support for this idea. When teacher basic skills test scores have been used as predictors of teacher performance, few studies have shown any strong relationship, and some studies have shown no relationship at all. Similarly, studies that have attempted to

[4] A purely theoretical but highly useful quantity, the score a perfectly reliable test would produce.
[5] If all three types of inference were in fact a requirement, few tests would ever pass muster.

relate content knowledge to teacher performance (most of them confined to mathematics and science) have shown modest results at best. The reasons for these results are unclear, but two possibilities present themselves. First, it is a principle in statistics that restriction of the range of one variable in a correlational study usually produces measures of association that underestimate the relationship between the two variables. To understand why this is true, consider that if all individuals have the same level of ability no mathematical relationship between ability and any other variable could ever be demonstrated. This would explain why test scores and teacher performance show such weak relationships: teachers are a population selected by various procedures such as college entrance examinations and GPA requirements that assure they have higher levels of academic skills than would be true for an unselected group of persons from the general population. Secondly, the problem could be with the measures of teacher performance. As mentioned above, this is a very difficult thing to measure or even to define, and it is not given that the measures of teacher performance used in previous studies have been especially reliable or valid. This would limit the findings of these studies because reliability and validity have direct bearing on the relationship of a criterion with other variables.

An additional consideration is more political than technical. Because of the commission reports of the 1980's there has been a public perception that the quality of teachers is generally poor, and much of the impetus behind implementation of the current testing regime can be traced to public demand for improvement of teacher quality. When first implemented, teacher basic skills tests resulted in large proportions of failures in some places, which further inflamed public opinion. It is arguable whether the problem of poor academic skills among teachers was as great nationally as press reports at the time seemed to indicate, but the public perception was important and inescapable. Public perception is not a validity issue in a technical sense, but it must be considered in any test development effort. However well founded formal validity studies and cut score procedures happen to be, they will usually be unconvincing if the public perceives teacher quality to be a serious problem and the cut scores too lenient.

Statistical tables that relate test scores to the expected level of a criterion exist, but because of the weak evidence of the relationship of teacher test scores to teacher performance (and because we have no clear definition or measure of teacher performance in any case) we cannot use these tables to select the point at which some proportion of candidates would be expected to reach some desired level of teacher quality.[6] But because validity in modern terms is taken as a chain of inference about the usability of a test for specific purposes, we have other alternatives.

The solution to this problem is the Angoff procedure, a method developed in 1971 by William H. Angoff and refined over the intervening years by other psychometricians. This method capitalizes on the information available in the process of content validation of a test. Although we may not have a clear definition of what constitutes good teacher performance, it makes sense that teachers and those who train and employ teachers would have a good idea of what is a minimally acceptable level of knowledge and basic academic skills for persons who teach. The Angoff procedure works by assembling a panel of such experts for each test, presenting test items for their consideration, and

[6] And note that even if we had an acceptable criterion, we would still have the problem of determining what level of the criterion were optimal.

asking them to estimate the proportion of persons with minimally acceptable skills in the given content area who would be expected to get each item right. A criterion is set in advance for the proportion of items that must be judged job relevant in order for the test to be deemed a valid measure of performance. After all items have been rated, the judgments of the raters are combined to determine a cut score for the whole test.[7] The principle thus operationalized is the creation of a cut score that would indicate that an individual has a minimally acceptable level of knowledge. The assumption behind the use of this procedure is that the state desires to assure that persons entering the teaching profession have some minimum level of content or basic skills knowledge, but does not wish to greatly restrict the pool of available candidates.

There are other similar methods of setting cut scores, but the Angoff procedure is the most widely used, and has held up in most research studies as the method that produces the most stable results. It has generally been accepted by the courts and by professionals in the field as a reasonable means of providing evidence of validity of a cut score for the selection of teachers and other employment candidates. The Angoff procedure is the method used to set cut scores for Kentucky certification tests, and is widely used with licensure tests of many other types.

There are a few other technical issues that need to be considered when tests are used to make decisions that have serious consequences for individuals. First, it is important to remember as noted above that no test is perfect, and that there is a range of uncertainty around the true score estimated by any test score. Because we use tests to classify individuals into dichotomous categories, some persons will be misclassified. Both false positives and false negatives are unavoidable, and an effort to minimize one will usually result in an increase in the other. A false negative denies an individual who is otherwise capable of teaching the right to do so; a false positive places an unqualified person in the classroom. Either type of error may have more serious consequences than the other depending on the use of the test,[8] and our strategy must be predicated on our evaluation of which is the more serious error. One of the procedures often used to correct for unreliability is to set the cut score one $S_{EM}$ above or below the level established by the Angoff procedure. If the cut score is set one $S_{EM}$ below, the number of false negatives will be reduced; if set one $S_{EM}$ above, the number of false positives will be reduced. Note that if cut scores are set near the extreme of the distribution of test scores – either very high or very low – misclassifications will be much less of a problem than if the cut scores are set near the center of the distribution, where errors of measurement have the greatest effect.

A related and more serious consideration has to do with the performance of subgroups on our tests. It is an unfortunate but undeniable fact that minorities as a group tend to perform less well on standardized tests than do whites. The reasons for this problem are beyond the scope of the present paper, but the lower relative performance of minorities presents unavoidable complications in the process of setting cut scores. A decision to raise a cut score will be relatively disadvantageous to the group with the

[7] This is of course a stark oversimplification of the methods employed by the Angoff procedure.

[8] Many experts in both education and psychometrics view false positives as the more serious error in the case of teacher tests, while false negatives are usually thought to be the more serious in the case of tests that affect the outcomes of students.

lower mean score,[9] and as a result a greater proportion of minority than white persons will usually fail, who would have passed under the older, lower cut score. Although the lower mean scores of minorities do not necessarily imply that a test is biased, care must be taken to assure that a decision to set a cut score at any given level does not produce a larger number of false negatives for minority candidates than for whites.

Case law

Rather than detail the numerous specific court cases that have arisen relative to the use of tests in education and personnel selection, we consider here the legal principles that have been established by these cases. Analysis of some important cases will be reserved for Section 2.

It is important to understand the legal basis for court action involving tests. There are many possible grounds on which an individual could bring an action related to test use, but the grounds for most cases include only due process and equal protection. This is because an individual, in order to bring an action, must have a cause recognized in the law, and the relevant law and regulations deal with rights deriving from the 14th Amendment. The 14th Amendment prohibits specifically state action which would deprive persons of life, liberty, or property without due process of law; thus an individual, in order to bring an action, must be able to plausibly claim that he or she was deprived of either liberty or property.[10]

In addition to the 14th amendment, relevant law and regulation include Titles VI and VII of the Civil Rights Act of 1964, and 29 CFR, volume 4, part 1607.

The courts are usually reluctant to intervene in state regulation of the professions, and as a result the requirements of proof for plaintiffs are challenging. Even though 29 CFR establishes the four-fifths rule as minimum evidence of disparate impact, court decisions have established a number of additional conditions that must be met before a test program can be judged to violate anyone's rights. When a plaintiff can demonstrate that the four-fifths criterion has been met, the burden of proof shifts to the test administrator and developer to prove that use of the test is necessary and that it has an essential relationship to the requirements of the job for which it selects. If this burden is met, then the burden shifts again to the plaintiff, who has the option of proposing an alternative method of selection, but must prove that it is equally effective at selecting qualified individuals at no greater cost than the state's procedure.

The reluctance of the courts to intervene in these matters is based on a distinction made by the courts between *fundamental* rights and *nonfundamental* rights. Fundamental rights include mostly those rights guaranteed by the Constitution; nonfundamental rights include most notably for our purposes, economic rights. When a fundamental right is at issue, the courts apply *strict scrutiny* to the state's need to implement a particular program; when nonfundamental rights are at issue, the courts apply the principle of *mere rationality*. These distinctions are important because the standards of proof for the two types of circumstances are quite different. In the case of fundamental rights, there is a presumption that the state's action is unconstitutional, and the state must go to extraordinary lengths to prove that its action is necessary. In the case of nonfundamental

[9] This is a consequence of the mathematics of population distributions.
[10] Deprivation of life never enters the picture, except in death penalty cases.

rights, there is a presumption that the state is acting reasonably, and the courts are quite lenient about the kinds of evidence that would justify continuation of the state practice.

The validity provisions of 29 CFR are rarely effective in providing relief to plaintiffs, because courts generally give test administrators wide latitude in the types of evidence used to adduce test validity. In contradistinction to the methods prescribed by 29 CFR, which tend to be very specific about the technical methods required to establish validity, courts seem to generally accept evidence that validation efforts are reasonable, credible, and well-intentioned. In those few cases where courts have ruled against test systems on the basis of validity, the test administrator or developer had engaged in egregious disregard for any consistent and reasonable procedure for establishing the validity of the test. Courts will often accept test programs with weak validity studies if they are convinced that the developer and administrator made a good-faith effort to establish validity and have a coherent plan for remediating the deficiencies of the validity studies.

Courts are especially prone to give test developers latitude on the question of the validity of the performance measure for which the test selects. In one notable case,[11] the court noted that the definition of functional literacy the test was designed to measure was just one of many, and no one had established a generally acceptable definition of what functional literacy was. But the court concluded that the definition applied by the state was reasonable if not widely accepted, and that the test as developed did an adequate job of measuring an individual's functional literacy by that definition. This the court ruled a constitutionally acceptable use of the test.

Courts have generally held that the testing of persons to determine their suitability to serve as teachers is a reasonable public policy. When the courts are willing to consider whether a state's testing program is legitimate, their scrutiny will usually be based on one of three possible grounds, Procedural Due Process, Substantive Due Process, and Equal Protection. A brief discussion of these three considerations follows.

*Procedural due process*

When the state implements a policy that entails the possibility of depriving individuals of property or liberty, it must assure that adequate procedures are in place to safeguard the due process rights guaranteed by the 14th Amendment. Due process is the subject of a legal history much too voluminous to summarize here, but as it applies to testing generally, employment testing in particular, and to other employment requirements, it may be reduced to a few principles. Previous cases decided on this basis have revolved around the idea that the state must provide an adequate mechanism for individuals to appeal negative decisions and must provide adequate notice of changes in policy. The specific due process procedures that must be provided are not fixed: the courts weigh the salience of the harm that might be done to the individual and the likelihood that due process will result in a reversal, against the cost to the state of implementing due process procedures. Where no property right or liberty interest is involved, the state has no obligation to provide due process procedures, except as established in statute.[12]

[11] Debra P. v. Turlington, 474 F.Supp. 244 (M.D. FL 1979)

[12] See Kentucky Administrative Law, 1st Edition, for a discussion of this subject as it applies to Kentucky.

When the courts have viewed testing programs as inadequate on procedural due process grounds, the issue has usually been adequate notice of change. Testing programs represent an assessment of the performance of programs that take years to prepare candidates, and a sudden change in policy may not provide either candidates or the institutions that provide the programs adequate time to adjust. Since individuals should be held responsible only for what is within their control, inadequate notice subjects them to loss of property or liberty without due process.

*Substantive due process*

Although the courts accept teacher testing as a legitimate public policy, they insist that tests must have a legitimate purpose and that they be reasonable. This resolves in general into a requirement that test development follow a credible and professionally appropriate development methodology, and that decisions about test use – including the setting of cut scores – not be capricious. Persons should be tested only on matters that are under their control; there should be a demonstrated relationship between the content of the test and the purpose for which the test is intended; and validation efforts should have a credible rationale and should be conducted in good faith. Most notably for our purposes, courts have held that cut scores should not be set based on purely political or public relations criteria, and often scrutinize carefully the professional qualifications of persons selected for cut score panels. In two cases courts have ruled that specific testing programs were unconstitutional because the test was not validated for the state's use. In Sharif v New York State Education Department, the court ruled that the SAT could not be used to select students for scholarships; and in Groves v Alabama State Board of Education, the court held that the ACT could not be used to select candidates for teacher preparation programs.

*Equal protection and disparate impact*

Equal protection cases are especially difficult to prove. 29 CFR establishes the four-fifths rule as a presumptive test of disparate impact, but courts have ruled that disparate impact by itself does not constitute discrimination, and require the additional proofs described above. In addition to disparate impact cases, some cases have been brought on the basis of *discriminatory intent*, the idea that a state policy intentionally discriminates against some protected group. Such cases are very difficult to prove. When plaintiffs have been successful, it has usually been when it was shown that the state knowingly ignored the advice of experts who had warned of the possibly discriminatory consequences of adopting some policy.

**The Kentucky Experience**

Teacher tests have been used in Kentucky for some time, and quite a bit of data have accumulated in the EPSB database. Analysis of these data can help illuminate issues related to setting cut scores.

The first consideration is whether the results show differences for subgroups, perhaps meeting the disparate impact test of 29 CFR. An analysis of data on this subject

produced Tables 1 and 2.[13] Table 1 shows the means and standard deviations for Whites and African-Americans[14], and for both groups combined, for those tests that are currently in use by EPSB for which sample sizes permitted analysis.[15] Note that the means for African-Americans are lower than for Whites for all tests, and that the standard deviations are generally of about the same magnitude.

Table 2 shows the relative pass rates for African-Americans and Whites for each of the listed tests. As can be seen from the table, just four of the tests would pass the four-fifths criterion with the current cut scores.[16]

It is instructive to evaluate what would happen if the current cut score were reset to some arbitrary point, perhaps the median score for the entire sample.[17] Table 3 shows the consequences in terms of disparate impact of accepting this policy. Most white candidates would still pass if the cut score were set at this level, but in most cases there would be severe consequences for African-American candidates. Notice that 4 of the 30 tests would pass the disparate impact criterion with the existing cut score, but none would if the cut score were increased to the median of the overall population of candidates.[18]

As noted above, candidates who fail to pass on the first try may be able to obtain a conditional certificate, providing them with the opportunity to teach while remediating deficiencies. There were too few subjects to allow for estimates to be computed for any specific test, but for the 133 conditionally-certified individuals in the sample, 67% passed one or more of the assessments on later testing.

Finally, it is of interest to evaluate whether, subsequent to a change in cut scores, the mean score of any test increased. This is of interest because one of the hypotheses often heard about teacher training is that raising the standards for teacher certification would attract more capable persons into teaching. We computed the mean test scores before and after cut score changes for those tests whose cut scores had been adjusted upward in recent years and found no difference in mean score for any test. This fails to provide support for the idea that the overall quality of candidates has increased since cut scores were raised.

**Analysis**

[13] The reporting of ethnicity is voluntary, and a large proportion of the individuals for whom assessment data are available do not have information on this variable. The results reported here include values only for African Americans and Whites, to make the results consistent across categories.

[14] The number of minorities in all other groups except African Americans was too small to permit the calculation of reliable estimates, and the number of African-Americans was too small to permit reliable estimates to be calculated for some tests.

[15] Tests were selected for analysis only if scores for 30 or more African-American subjects were available.

[16] These are scores for all candidates compared to the current cut score. In a few cases cut scores have been raised in recent years, and some candidates would have been subject to lower cut scores at the time when the test was actually taken.

[17] The median was selected for this test instead of the mean because for a few of the tests, the population mean is not a good estimate of the center of the distribution. Exactly one half of the subjects fall below the median score.

[18] Note that these results depend on the relative proportions of white and minority subjects tested so far. If the mix of white and minority subjects were to change, the median for the entire sample might change as well, and this would affect the results somewhat.

There are three considerations that must be encountered in our evaluation of cut scores for Kentucky teacher certification tests:

1. Have the cut scores that presently exist been set appropriately?
2. What would be the advantage of setting cut scores for some or all tests at higher levels?
3. What consequences, positive or negative, might result from resetting cut scores?

1. Have the cut scores been set appropriately?

The answer to this question depends on an analysis of whether EPSB has used an appropriate methodology for establishing the validity of the tests for the purpose of selecting teachers, and the quality of the inferences made about the validity of the tests and the selected cut score.

With the exception of two, all of the tests used by EPSB were developed by ETS. ETS provides, as a requirement of its contractual agreement, a manual describing the procedures for cut score setting for each test, including a description of how the expert panel was assembled and how the process was managed. The expert panels are assembled by EPSB staff, based on procedures in the ETS manual. The procedures call for panels of persons with more than one and less than 8 years of experience, who are certified in the content area of the test, and are selected to represent as fairly as possible the geographic, gender, and ethnicity distribution of teachers across the state. In order to assure that panels fairly represent minorities, ETS recommends that ethnic minorities be oversampled, and EPSB staff go to great lengths to do this. Panels are required to have no fewer than 10 members. All persons selected for each panel are required to complete a biographical form, which serves as part of the documentation of the qualifications of the panel members.

Note that the final decision about any cut score is made by the EPSB board. ETS creates the tests and manages the content validity studies and the Angoff procedures that lead to recommended cut scores; but the actual cut score selected for any test is set by a Board vote.

The procedures used by EPSB and ETS are in keeping with generally accepted requirements for the Angoff procedure. At least insofar as a reading of the documentation can indicate, test validation procedures were conducted appropriately in the case of all ETS tests used in Kentucky, and the cut scores for the tests were set appropriately. Since the procedure used in setting cut scores requires an analysis by the panels of individual test items, and since these tests were content-validated, we have a sufficient basis for believing that the tests are valid for the purpose for which they were intended.

The two EPSB-produced tests, the Interdisciplinary Early Childhood Education Test and the Kentucky Principal Test, were developed using procedures similar to those used by ETS. In each case panels were assembled using the same criteria as for Praxis tests, and the items in the tests underwent content validation. The item pools for the two tests were initially developed by panels of experts as well.

The cut scores for the Praxis I used for admission to teacher preparation programs were initially set using the same modified Angoff procedure as for other tests. These cut scores are recommended minimums: teacher training programs have the option of setting higher cutoffs if they wish. The cut scores for the ACT, SAT, GRE, and GPA were set by EPSB Board action, apparently based on the judgment of board members about what constitutes a reasonable measure of academic proficiency.

With respect to the ETS tests, and probably also in the case of EPSB-developed tests, it is clear that appropriate validation and cut-score-setting procedures have been followed. Documentation about the process of assembling panels is impressive, and the results of panel deliberations are well documented. It is likely that cut scores as established through this process will hold up well should a challenge arise.

The cut scores for the Praxis I, to the extent that preparation programs use the minimum levels established by EPSB, are also appropriate, and are on equally solid ground. Cut scores established above the EPSB-recommended level by the colleges may not be. Although it is reasonable for EPSB to grant preparation programs latitude in setting admissions standards, doing so does not relieve the institutions of the obligation to justify the necessity and credibility of levels set by them.

Cut scores for the ACT, SAT, and GRE are more problematic. As noted above, there have been court decisions that denied states the right to use the ACT and SAT for purposes other than their original use. The GRE, which is a similar test developed for similar purposes, may well be covered by the same rule. The *Sharif* and *Groves* decisions are not binding in Kentucky, but it is possible that federal or state courts would view them as providing persuasive authority. Additionally, there seems to be no readily accessible documentation for the rationale of the standard-setting decision in the case of these tests, and it seems unlikely that the decision was based on a technically sound methodology, since EPSB is hardly in a position to perform content validation on these measures. It is possible, however, that courts would view the history of cut score setting for these tests as a reasonable and well-intentioned effort by the EPSB to provide selection procedures for teacher education candidates, and might allow the program to continue. It would be in our interests to review this history, to determine whether we can defend our reasoning in the case of the use of these tests.

When preparation programs use tests other than those enumerated above in determining admissions, it is essential that they be able to document the inferences that led them to set cut scores. The responsibility and liability for these procedures are the responsibility of the institutions, but because EPSB established the regulation that covers the use of these tests for admissions, it would be in our interests to encourage them to use sound procedures and to document their reasoning.

Although the cut score for the GPA has the same basis as for the SAT, ACT, and GRE – board action based on the judgments of board members – I suspect that it is much less subject to question. GPA is not a test, and it would be unreasonable to expect the state to conduct a formal validity study leading to a cut score decision. The GPA is a widely-accepted global measure of student academic performance, and has a great deal of credibility both within the academic community and with the general public. The discussion of the rationale for the GPA cut score level has been detailed and is well documented. Since GPA is a widely-recognized measure of academic performance that is within the control of a teacher training candidate, I suspect that no regulatory agency or

court would view the currently-set level of 2.50 as unreasonable, in light of the state's responsibility to apply reasonable teacher program admissions standards.

2. What would be the advantage of setting cut scores for some tests at higher levels?

Note from Table 2 that the overall pass percentiles for the various tests vary markedly for both ethnic groups. For African-Americans, the pass rates vary from a low of .2 for *Biology: Content Knowledge Part 1* to a high of 1.0 for the *School Leadership Licensure Assessment*. For White subjects, the pass rates range from .59 for the *PPST Mathematics* test to a high of .99 for the *School Leadership Licensure Assessment*. One might suppose, based on these data, that the cut scores are too low for the School Leadership test, and perhaps appropriately set for the PPST Mathematics test. But a deeper analysis reveals that the situation is not so simple.

First, consider the rationale for the cut score setting procedure. As noted above, the intent of the procedure is not to select a small sample of highly qualified individuals from a large pool of candidates; it is to select persons who are minimally competent from the group of persons who register to take the test. Given the small number of persons each year who are candidates for the school leadership assessment, and the fact that this is a group of highly experienced and presumably highly motivated individuals, it is not unreasonable to suspect that a large proportion of them do in fact possess the minimum levels of skill necessary to serve as school leaders. To the extent that the test appropriately and accurately classifies persons into categories based on a minimum level of competence, it might happen on any given occasion that most people who take the test are capable of doing the job. Still, a pass rate in excess of 95% does seem a bit high.

On the other hand, considering that the PPST Mathematics test is a general test of mathematical knowledge for teachers in all disciplines, having 41% fail may be a bit high; or it might indicate that teacher preparation candidates are woefully deficient in basic mathematics skills.

The cut scores in both cases were set by experts with a knowledge of the minimum requirements for success in the relevant roles, and it was the standard of minimum performance rather than a desire to select the most capable candidates that determined their decisions. It is possible that the expert panels that set these levels were either too lenient or too severe, but we should not lightly overrule their deliberations.

The two examples above illuminate an important issue about cut scores: they can be established either as an absolute standard of performance, or they can be established on a normative basis. The use of an absolute standard makes the assumption that some minimum level of knowledge – which might be met by most or all of the candidates – is of interest, while the normative approach assumes that competence represents a continuous scale, and we wish to select just some proportion of those who are most competent. Most licensure procedures for any occupation, regardless of where they set the cut score, assume an absolute, rather than a normative, standard.

Despite the generally well-founded basis for cut scores set by EPSB, there might be some reason to think that some scores could be reset. The Angoff procedure is the best-accepted method for setting cut scores in licensure testing, but it is not without limitations. One of its limitations is that it tends to set cut scores that are a bit more liberal than those of other similar procedures. The practical effect of this tendency to

liberality is that when cut scores are set below the population mean, it tends to underestimate the level of minimum competence required for the job. This would tend in general to increase the number of false positives relative to other types of procedures. Additionally, we have to consider the credibility of our cut score levels: if the public perceives that the tests are too easy, their confidence in the teaching profession will suffer.

The advantage of setting cut scores at higher levels based on the above analysis would be to reduce the number of false positives. This would increase the number of false negatives, depriving some capable individuals of the right to teach, but would assure that the largest possible number of unsuitable persons were kept from the classroom. Since individuals are given an opportunity to repeat the test, one could argue that many individuals who were in the false negative category would on retesting pass, ameliorating somewhat the greater number of false negatives. This is an especially attractive supposition in light of the success of the conditional certification program, which gives capable persons who fall into the false negative category every opportunity to demonstrate on subsequent testing that they are indeed capable of serving in their content area, without altogether denying them the opportunity to teach. Especially since districts who hire these persons are required to provide them with assistance, this measure can be seen as providing them every reasonable opportunity to succeed.

An additional advantage of raising cut scores might be to foster the impression among the public and policymakers that Kentucky is making a serious effort to assure teacher quality, and that the average quality of persons entering the teaching force is increasing. There is a hazard to this view, however: when teacher competency tests have resulted in a large number of failures in the past, public confidence in the quality of the teacherforce actually declined, due to the perception that large numbers of unqualified persons were entering teaching.

3. What consequences, positive or negative, might result from resetting cut scores?

All of the possible consequences, both positive and negative, flow from the indisputable fact that raising cut scores would eliminate some candidates who otherwise would be certified and would serve as teachers.

If it is true, as is often suggested in the press, that the average quality of teachers is less than desirable, then raising cut scores would have the effect of increasing the overall quality of the teacherforce. A number of recent studies have suggested that teachers are the single greatest contributor to children's academic success. There is some doubt whether this is true in an absolute sense but it certainly is true that teacher effects are the largest contributors that are controllable by educational policy. The research has failed to show a strong relationship between teachers' basic skills or content knowledge and student outcomes, but the idea has such intuitive appeal that almost everyone nonetheless believes that higher basic skills and content scores will improve the quality of teaching. The question then is only how much of an improvement in teacher effectiveness is likely to result from increases in measures of teacher academic skills.

We could not however increase cut scores on our tests and thereby raise the average quality of teaching without at the same time creating negative consequences. The principal problem would be the creation of critical shortages in some categories of

certification. Of the 3000 or so persons who apply every year to become first-time teachers, about a third are elementary education majors. There continues to be an oversupply of candidates in this area, and resetting the cut score at the 30th percentile instead of the 10th would probably not result in a shortage. Resetting the middle school mathematics test at the 40th instead of the 30th percentile might. Resetting any of the tests at the 50th percentile would likely result in shortages in most certification areas.

The problem with thus reducing the pool of qualified persons is that whether or not there are enough such persons, every classroom in the Commonwealth must be staffed. Barring an increase in the maximum class size, the only way to assure this under the above scenario is to allow districts to employ emergency certified staff or staff with conditional certificates.[19] This would do nothing to raise the average quality of teachers in the classroom.

If cut scores are raised, there are really just three possible outcomes relative to the number of persons available to teach:

1. We might have a reduction in the number of candidates, with no replacement from any more capable population.
2. There might be no change in the population of persons applying for teaching certificates, but the average score might increase due to better teacher preparation.
3. Persons eliminated from the pool of candidates might be replaced by better-qualified persons.

Some authors suggest that the third possibility might well occur, as persons who previously viewed education with disdain because of its perceived dominance by persons of less ability would now select education as a career. There is no research evidence to suggest that this either would or would not happen. What does seem reasonable is the idea that resetting cut scores alone, without concomitant efforts at recruitment and changes in teacher education programs, would probably not measurably change the quality of the teacher candidate pool.

Another problem that would be created, as noted above, would be a serious disproportionate reduction in the number of minority candidates who would pass the tests. This was a serious concern when some cut scores were adjusted upward a few years ago, and is likely to be controversial at any time in the future. As should be clear from analysis of the legal issues, disparate impact is not in itself a reason to refrain from raising cut scores, but because of the sensitivity of this issue, it is important to assure that any decision to do so does not appear capricious.

It is not given that minorities will inevitably score less on tests than will whites. Candidates from some traditionally minority institutions, such as Grambling University, have a proven track record of scoring as well as candidates from any institution.[20] It is possible that the problem of disparate impact could in fact be ameliorated by application

[19] A third possibility would be to recruit many more persons via alternative certification programs. It is unclear at present whether enough such persons exist.

[20] Many of the African-American subjects who failed various Kentucky tests in recent years came from two particular programs. Substantial efforts have been made to improve these programs, and the pass rates of their graduates have subsequently come up to acceptable levels.

of program changes at institutions that serve many minority candidates. As with the case of the overall reduction of teacher candidates, this would require that any change in cut scores be accompanied by program changes.

Finally, it is important to note that cut score setting decisions may affect our efforts to achieve other policy goals. Consider that many believe that low cut scores are evidence of poor teacher quality and wish to raise cut scores in order to improve teacher quality. At the same time, federal law requires that teacher training programs be judged in terms of their success rates, determined in part by the proportion of their graduates who pass certification tests. At any given cut score level, a high failure rate cannot be both a measure of good teacher quality and poor program quality. If cut scores are adjusted upwards, then program success requirements must be adjusted downwards, or institutions must be given adequate time to adjust their program to the new levels.

**Summary**

The process of setting cut scores for teacher certification tests is a complex matter involving legal, technical, political, and public relations considerations. Cut scores divide candidates who take the tests into two groups, those who pass and those who fail. The use of such tests is subject to challenge on legal and technical grounds, and it is essential to use technically defensible methods, and to carefully document the chain of inference and rationale used to justify the selected level. Decisions that appear capricious are likely to subject the test administrator to the risk of challenge.

Regardless of the technical methods used, cut scores inevitably will present the administrator with consequences that may not be acceptable. Most importantly, setting cut scores too high may restrict the pool of available candidates, causing shortages; may have a disproportionately severe and legally significant impact on the pass rates of minorities; and may place teacher training institutions in a bad light. These consequences must be carefully considered when changes in cut score levels are contemplated.

Kentucky's use of tests has been technically sound in general, and both test validation and cut score setting procedures have been appropriate. The relatively low level of some of the cut scores is a consequence of the basic philosophy of the procedures used to set cut scores, selection of individuals with a minimum level of competence. Adjustment upward of some cut scores might be indicated and might not result in severe negative consequences, but resetting the cut score for any test near the population median would probably result in severe consequences. Although most Kentucky tests exceed the disparate impact criterion at the current cut score levels, the technical quality of the development procedures probably provides considerable protection from liability. Substantial increases in cut scores would leave the state open to challenge on disparate impact grounds, and any decision to do so must take the requirement for careful attention to the reasonableness of the chain of inference and requirement for noncapriciousness into account.

Some of the academic achievement measures selected for teacher training admissions are more problematic. Case law may suggest that use of these tests for this purpose is inappropriate, and it is doubtful that EPSB could justify the decision to use them on technical grounds. In the case of the Praxis I and other institution-selected tests,

teacher training programs have a responsibility to justify their cut score setting procedures, just as does EPSB.

Table 1
Mean Scores by Ethnicity on Selected Certification Tests

| Test Name | Mean of all subjects | Standard deviation all subjects | African-American mean | African-American Standard deviation | White mean | White standard deviation |
|---|---|---|---|---|---|---|
| PPST Writing | 180.2 | 30.8 | 175.2 | 23.1 | 180.6 | 31.3 |
| PPST Mathematics | 179.8 | 31.4 | 171.6 | 21.5 | 180.5 | 31.9 |
| PPST Reading | 182.5 | 33.9 | 175.4 | 22.6 | 183.1 | 34.6 |
| Principles of Learning and Teaching: Grades K-6 | 172.3 | 12 | 162 | 13.4 | 172.9 | 11.6 |
| Interdisciplinary Early Childhood Education Test | 165.7 | 12.3 | 157.5 | 7.3 | 166.7 | 12.6 |
| Elementary Education: Curriculum, Instruction & Assessment | 171.9 | 16.3 | 156.1 | 18.1 | 172.9 | 15.6 |
| English Language, Literature and Composition: Content Knowledge | 173.1 | 15.2 | 157.1 | 14.3 | 174.2 | 14.6 |
| English Language, Literature, & Composition: Essays | 156.8 | 11.4 | 148 | 12.2 | 157.3 | 11.1 |
| Middle School English Language Arts | 166.8 | 15.5 | 153.7 | 14.5 | 167.7 | 15.2 |
| Mathematics: Content Knowledge | 139.6 | 19.7 | 123.5 | 16.5 | 140.6 | 19.3 |
| Mathematics: Proofs, Models, & Problems, Part 1 | 156.3 | 20.4 | 141.3 | 17 | 157.4 | 19.9 |
| Middle School Mathematics | 162.1 | 17.8 | 148.1 | 13 | 162.9 | 17.7 |
| Social Studies: Content Knowledge | 161.7 | 16.6 | 151.7 | 14.9 | 161.9 | 16.6 |
| Social Studies: Interpretation of Materials | 164.8 | 11.6 | 154.5 | 11.3 | 165.3 | 11.4 |
| Middle School Social Studies | 162 | 16.9 | 151.2 | 16.9 | 162.5 | 16.7 |
| Physical Education: Content Knowledge | 151.2 | 10 | 145.2 | 9.7 | 151.6 | 9.9 |
| Physical Education: Movement Forms-Analysis, Design | 155.7 | 9 | 149.4 | 11.9 | 156.1 | 8.7 |
| Business Education | 638 | 54.3 | 602.1 | 53.1 | 638.4 | 53.2 |
| Music: Concepts and Processes | 151.2 | 15.3 | 142.8 | 12.5 | 151.7 | 15.1 |
| Music: Content Knowledge | 162.3 | 12.8 | 147.6 | 11.4 | 163.5 | 12.1 |
| Biology: Content Knowledge Part 1 | 163 | 16.7 | 144.2 | 14.7 | 164.4 | 16 |
| Health Education | 665.2 | 64.1 | 626.5 | 63.6 | 666 | 63.3 |
| Special Education: Application of Core Principles Across Categories of Disability | 149.7 | 12.5 | 139.7 | 13.6 | 150.2 | 12.2 |
| Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances | 161.9 | 13.8 | 150.6 | 14 | 162.6 | 13.4 |
| Principles of Learning and Teaching: Grades 5-9 | 168.4 | 12.7 | 158.3 | 12.4 | 169.1 | 12.5 |
| Principles of Learning and | 170.9 | 11.3 | 163 | 13.6 | 171.3 | 10.9 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Teaching: Grades 7-12 | | | | | | |
| Kentucky Principal Test (KYPT) | 93.4 | 28.2 | 88.2 | 10.6 | 93.1 | 23 |
| School Leadership Licensure Assessment | 175.9 | 8.4 | 171.8 | 9.4 | 176.2 | 8.3 |
| Communication Skills | 661.9 | 10 | 651.6 | 11.1 | 661.6 | 9.9 |

Table 2
Pass Rates for African-American and White Subjects
At current cut score levels

| | African-American Pass Rate | White Pass Rate | African-American/White Pass Ratio |
|---|---|---|---|
| PPST Writing | 0.5 | 0.74 | 0.73 |
| PPST Mathematics | 0.3 | 0.59 | 0.54 |
| PPST Reading | 0.5 | 0.69 | 0.77 |
| Principles of Learning and Teaching: Grades K-6 | 0.6 | 0.86 | 0.66 |
| Interdisciplinary Early Childhood Education Test | 0.9 | 0.97 | 0.94 |
| Elementary Education: Curriculum, Instruction & Assessment | 0.4 | 0.76 | 0.49 |
| English Language, Literature and Composition: Content Knowledge | 0.4 | 0.82 | 0.52 |
| English Language, Literature, & Composition: Essays | 0.4 | 0.68 | 0.54 |
| Middle School English Language Arts | 0.6 | 0.83 | 0.71 |
| Mathematics: Content Knowledge | 0.5 | 0.8 | 0.59 |
| Mathematics: Proofs, Models, & Problems, Part 1 | 0.4 | 0.8 | 0.55 |
| Middle School Mathematics | 0.6 | 0.87 | 0.73 |
| Social Studies: Content Knowledge | 0.5 | 0.73 | 0.7 |
| Social Studies: Interpretation of Materials | 0.5 | 0.85 | 0.62 |
| Middle School Social Studies | 0.6 | 0.86 | 0.72 |
| Physical Education: Content Knowledge | 0.5 | 0.71 | 0.67 |
| Physical Education: Movement Forms-Analysis, Design | 0.6 | 0.81 | 0.72 |
| Business Education | 0.7 | 0.89 | 0.78 |
| Music: Concepts and Processes | 0.5 | 0.73 | 0.65 |
| Music: Content Knowledge | 0.4 | 0.87 | 0.49 |
| Biology: Content Knowledge Part 1 | 0.2 | 0.7 | 0.3 |
| Health Education | 0.6 | 0.75 | 0.77 |
| Special Education: Application of Core Principles Across | 0.3 | 0.66 | 0.49 |

| | | | |
|---|---|---|---|
| Categories of Disability | | | |
| Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances | 0.3 | 0.67 | 0.45 |
| Principles of Learning and Teaching: Grades 5-9 | 0.4 | 0.77 | 0.54 |
| Principles of Learning and Teaching: Grades 7-12 | 0.6 | 0.85 | 0.7 |
| Kentucky Principal Test (KYPT) | 0.7 | 0.86 | 0.81 |
| School Leadership Licensure Assessment | 1 | 0.99 | 0.97 |
| Communication Skills | 0.7 | 0.94 | 0.75 |

Table 3
Pass Rates by Ethnicity assuming a cut score at the median

| Test Name | Population Median | African-American Pass Rate | White Pass Rate | African-American/White Pass Ratio |
|---|---|---|---|---|
| Interdisciplinary Early Childhood Education Test | 165 | 0.16 | 0.57 | 0.28 |
| English Language, Literature and Composition: Content Knowledge | 174 | 0.14 | 0.54 | 0.26 |
| English Language, Literature, & Composition: Essays | 155 | 0.37 | 0.68 | 0.54 |
| Middle School English Language Arts | 167 | 0.17 | 0.54 | 0.31 |
| Mathematics: Content Knowledge | 138 | 0.21 | 0.54 | 0.39 |
| Mathematics: Proofs, Models, & Problems, Part 1 | 156 | 0.2 | 0.53 | 0.38 |
| Middle School Mathematics | 163 | 0.15 | 0.53 | 0.28 |
| Social Studies: Content Knowledge | 161 | 0.28 | 0.53 | 0.53 |
| Social Studies: Interpretation of Materials | 165 | 0.23 | 0.53 | 0.43 |
| Middle School Social Studies | 162 | 0.36 | 0.52 | 0.69 |
| Physical Education: Content Knowledge | 151 | 0.31 | 0.55 | 0.56 |
| Physical Education: Movement Forms-Analysis, Design | 157 | 0.23 | 0.52 | 0.44 |
| Business Education | 640 | 0.3 | 0.52 | 0.58 |
| Music: Concepts and Processes | 150 | 0.4 | 0.6 | 0.67 |
| Music: Content Knowledge | 162 | 0.09 | 0.57 | 0.16 |
| Biology: Content Knowledge Part 1 | 163 | 0.09 | 0.55 | 0.16 |
| Health Education | 670 | 0.29 | 0.55 | 0.53 |
| Special Education: Application of Core Principles Across Categories of Disability | 150 | 0.25 | 0.55 | 0.45 |
| Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances | 161 | 0.23 | 0.59 | 0.39 |
| Principles of Learning and Teaching: Grades K-6 | 174 | 0.2 | 0.52 | 0.38 |
| Principles of Learning and Teaching: Grades 5-9 | 169 | 0.23 | 0.54 | 0.43 |
| Principles of Learning and Teaching: Grades 7-12 | 171 | 0.3 | 0.55 | 0.55 |
| Kentucky Principal Test (KYPT) | 94 | 0.33 | 0.52 | 0.63 |
| PPST Writing | 174 | 0.33 | 0.49 | 0.67 |
| PPST Mathematics | 174 | 0.25 | 0.5 | 0.5 |
| PPST Reading | 175 | 0.41 | 0.55 | 0.75 |
| School Leadership Licensure Assessment | 176 | 0.43 | 0.56 | 0.77 |
| General Knowledge | 657 | 0.18 | 0.56 | 0.32 |
| Communication Skills | 663 | 0.18 | 0.5 | 0.36 |

## References


American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards of Educational and Psychological Testing* (1999).

Armstead v. Starkville Municipal Separate School District, 461 F.2d 276 (5th Cir. 1972)

Block, J.H. (1978) Standards and criteria: a response. *Journal of Educational Measurement*, 15(4): 291 -295.

*Board of Regents v. Roth*, 408 U.S. 564, 33 L.Ed.2d 548, 92 S.Ct. 2701 (1972)

Bowker, L. (2004) Personnel communication.

Bybdahl, C.S., Shaw, D.G., and Edwards, D. (1997) Teacher testing: reason or rhetoric. *Journal of Research and Development in Education*, 30(4): 248 – 254

Cahan, S., and Gamliel, E. (2001) Prediction bias and selection bias: an empirical analysis. *Applied Measurement in Education*, 14(2), 109-123.

Chinn, R.N., and Herts, N.R. (2002) Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education*, 15(1), 1-14.

Congress of the United States (1964). Civil Rights Act of 1964. P.L. 88-352.

Constitution of the United States of America (1787), as Amended

Debra P. v. Turlington, 474 F.Supp. 244 (M.D. FL 1979)

Dent v. West Virginia 129 US 114 (1889)

Educational Testing Service (1997, 1998, 1999, 2001, 2003, 2004). Standard Setting and Validation in Kentucky for Selected Tests in the Praxis Series. (Various study dates) Princeton, N.J.: Author

Educational Testing Service (2003). Validation and Standard-Setting Procedures for Tests in the Praxis Series. Princeton, N.J.: Author

Emanuel, S.L. (1995) Constitutional Law. Larchmont, N.Y.: Emanuel Law Outlines.

Flippo, R.F. (2002) Repeating history: Teacher licensure testing in Massachusetts. *Journal of Personnel Evaluation in Education*, 16(3): 211-229.

Gipps, C. (1999) Socio-Cultural aspects of assessment. In (A. Iran-Nejad and P.D. Pearson, Eds.) *Review of Research in Education*, 24: 355-392. American Educational Research Association

Glass, G.V. (1978) Standards and criteria. *Journal of Educational Measurement*, 15(4): 237 – 261.

Goss, S.J. (1986) Professional licensure and quality: the evidence. *Cato Policy Analysis no 79*. Retrieved from: www.cato.org/pubs/pas/pa079.html

Griggs v. Duke Power Co., 401 U.S. 424, 91 S.Ct. 849, 28 L.Ed.2d 158 (1971)

Groves v. Alabama State Bd. of Education, 776 F. Supp. 1518, 1532 (M.D. Ala. 1991)

Guyton, E., and Farokhi, E. (1987) Relationships among academic performance, basic skills, subject matter knowledge, and teaching skills of teacher education graduates. *Journal of Teacher Education*, September-October 1987: 37 – 42.

Hambleton, R.K., and Slater, S.C. (1997) Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education*, 10(1): 19 – 38.

Heller, D.E. and Shapiro (2001, April) Legal Challenges to High-Stakes Testing: A Case of Disparate Impact in Michigan? Paper Presented at the Annual Meeting of the American Educational Research Association, Seattle, Washington April, 2001. Retrieved from: http://www.personal.psu.edu/faculty/d/e/deh29/papers/aera01.pdf

Hurtz, G.M., and Auerbach, M.A. (2003) A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63(4): 584 – 601.

Hurtz, G.M. and Hertz, N.R. (1999) How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59(6): 885-897

Impara, J.C., and Plake, B.S. A comparison of cut scores using multiple standard setting methods. Buros Institute for Assessment Consultation and Outreach, University of Nebraska – Lincoln. Retrieved from:
www.unl.edu/BIACO/coop/lsac/aeramillardsympfinal.pdf

Jensen, A. (1980) Bias in Mental Testing. New York: Free Press.

Kane, M.T. (1994) Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3): 425-461.

Kane, M.T. (2002) Conducting Examinee-Centered Standard-Setting Studies Based on Standards of Practice. *The Bar Examiner*, 71(4)

Kentucky Administrative Law, 1st Edition (1999) Lexington: University of Kentucky

Kentucky Legislative Research Commission (2003a). 16 KAR 5:020. Standards for admission to teacher education.

Kentucky Legislative Research Commission (2003b). 16 KAR 6:010. Written examination prerequisites for teacher certification

Kentucky Legislative Research Commission (2003c). 16 KAR 7:010. Kentucky Teacher Internship Program.

Kentucky Legislative Research Commission (2003d). 16 KAR 2:180. One (1) year conditional certificate.

Klein, S.P. (1998) Standards for teacher tests. *Journal of Personnel Evaluation in Education*, 12(2): 123-138.

Linn, R.L. (2003) Performance Standards: Utility for Different Uses of Assessments. *Education Policy Analysis Archives,* 11(31). Retrieved from: http://epaa.asu.edu/epaa/v11n31/

McDonald v. Board of Elections, 394 U.S. 802 (1969)

McNeal v. Tate, 508 F.2d 1017 (5th Cir. 1975)

Mehrens, W.A. (1991) Social issues in teacher testing. *Journal of Personnel Evaluation in Education*. 4: 317 – 339

Mehrens, W.A. (1997) Validating licensing and certification test score interpretations and decisions: a response. *Applied Measurement in Education*, 10(1): 97 – 104.

Memory, D.M., Antes, R.L., Corey, N.R., and Chaney, D.E. (2001) Should tougher basic skills requirements be viewed as a means of strengthening the teaching force? *Journal of Personnel Evaluation in Education*, 15(3): 181-191.

Meyer v. State of Nebraska, 262 U.S. 390 (1923)

Miller v. State Board of Pharmacy, 262 So. 2d 188, 189 (Miss. 1972)

Mills, C.N. (1995) Comments on methods of setting standards for complex performance tests. *Applied Measurement in Education*, 8(1): 93-97.

Minnesota v. Clover Leaf Creamery Co., 449 U.S. 456 (1981)

Montalvo v. Mississippi State Board of Medical Licensure NO. 92-CC-01338-SCT, 1996

National research Council (2001). Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality. 2001, National Academy of Sciences.

Nguyen v. Medical Quality Assurance Commission, No. 68994-6, (Slip Op., August 23, 2001).

Norcini, J.J. and Shea, J.A. (1997) The credibility and comparability of standards. *Applied Measurement in Education*, 10(1): 39 – 59.

Parkes, J., and Stevens, J.J. (2003) Legal Issues in School Accountability Systems. *Applied Measurement in Education*, 16(2), 141-158.

Personnel Administrator of Massachusetts v. Feeney, 442 U.S. 256, 99 S.Ct. 2282, 60 L.Ed.2d 870 (1979)

Ravitch, D. (2004) A Brief History of Teacher Professionalism. White House Conference on Preparing Tomorrow's Teachers. Retrieved from http://www.ed.gov/admins/tchrqual/learn/preparingteachersconference/ravitch.html

Rice, J.B. (2004). Personal communication.

*Richardson v. Town of Eastover*, 922 F.2d 1152, 1156-1157 (4th Cir. 1991):

Rogers, P.S. (2004). Personal communication.

Sharif v. New York State Education Department, 709 F.Supp. 345 (1989)

Shepherd, L. (1980) Technical issues in minimum competency testing. In D.C. Berliner (Ed) *Review of Research in Education*, 8: 30 – 80. American Educational Research Association

Smith, G.P. (1984) The Critical Issues of Excellence and Equity in Competency Testing. *Journal of Teacher Education*, 35(2), 6-9.

Stone, G.R., Seidman, L.M., Sunstein, C.R., and Tushnet, M.V. (1996) Constitutional Law. New York: Little-Brown.

Uniform Guidelines on Employee Selection Procedures (1978). Code of Federal Regulations, Title 29, Volume 4, part 1607, Revised as of July 1, 2000.

United States v. Texas Education Agency, 564 F.2d 162 (1977)

Washington v. Davis, 426 U.S. 229, 96 S.Ct. 2040, 48 L.Ed.2d 597(1977)

Wilkerson, J.R., and Lang, W.S. (2003) The portfolio: panacea or Pandora's box? *Education Policy Analysis Archives*, 11(45).