4/99 DRAFT: For Internal Governmental Handling Only

Nondiscrimination in High-Stakes Testing: A Resource Guide

I disagree with the proposition that there are inherent racially based differences in the capacity of the American people to reach their full potential.

President Bill Clinton, October 21, 1994

An invalid test cannot measure merit.

Walls v. Mississippi State Dept. of Public Welfare, 542 F. Supp. 281, 311 (N.D. Miss. 1982), aff'd in relevant part, 730 F. 2d 306 (5th Cir. 1984).

Table of Contents

NONDISCRIMINATION IN HIGH-STAKES TESTING: AN OVERVIEW

I. Introduction

The issue of nondiscrimination in high-stakes testing is, at its core, a critical issue concerning access to education. When tests are used to make educational decisions, they should be used to measure students' abilities, knowledge, or qualifications, regardless of race, national origin, or sex. The U.S. Department of Education's Office for Civil Rights (OCR) has developed this resource guide in order to provide our staff and members of the educational community that we serve with practical guidance on testing and assessment principles that lie at the core of Title VI of the Civil Rights Act of 1964 (Title VI) and Title IX of the Education Amendments of 1972 (Title IX) case law.

This Resource Guide provides an overview of the federal standards and related educational principles that should guide the use of tests for making high-stakes educational decisions, such as those that involve: student placement in gifted and talented programs or programs serving students with limited English proficiency; referral of students for special education services; student promotion from one grade to another grade level; diploma awards; and higher education admissions decisions and scholarship awards. This Guide applies to norm-referenced and criterion-referenced tests[1] as well as to professionally designed alternative forms of assessment, which are used for making high-stakes educational decisions. The Resource Guide is not intended to apply to tests that are used to measure the performance of schools but have no high-stakes consequences for individual students nor does the Resource Guide address teacher-created classroom tests, even when such tests are being used for high-stakes educational decisions.

The issue of nondiscrimination in testing and assessment is properly viewed as consistent with standards-based reforms. Education leaders and the general public agree that there must be challenging standards for all students. In recent years, States and communities across the nation have embarked on far-reaching systemic efforts to reform their schools. Uniting their efforts has been an emphasis on high academic standards and high-quality assessments geared to those standards.

By defining what students should know and be able to do, standards keep schools focused on the desired results for students and can stimulate the development of appropriate curricula and the application of effective teaching strategies to make these results possible. Standards also indicate what assessments must measure in order to show achievement.

---

[1] Norm-referenced tests are tests used to identify an individual's performance in relation to the performance of other people in a specified group on the same test. American Psychological Association Standards for Educational and Psychological Testing (1985) (APA Standards) at p. 92. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level. APA Standards at p. 90. In other words, criterion-referenced tests are designed to measure to what degree a learner has mastered a certain skill.

High-quality assessments can make high standards meaningful by providing communities with a mechanism by which to hold schools accountable for achievement. It is critical that high standards for academic achievement be coupled with the necessary instruction and support that help students reach those standards - as determined by valid and reliable assessments.

The U.S. Department of Education is committed to the support of high standards and challenging assessments for all students. By outlining the relevant requirements of federal civil rights law, this Guide should assist states and local educational agencies instituting high-stakes assessments for all students. The Guide is intended to help states and local educational agencies avoid potential pitfalls in their implementation of high standards when using large scale assessments with educational consequences for individual students.

Federal civil rights laws ensure that all students have equal educational opportunities. Although many of the federal legal standards that should guide sound educational decisions are importable from the federal cases addressing employment discrimination, there are, nonetheless, critical differences. The educational institution's obligation to a student does not ordinarily end once a decision is reached to, for example, place the student in a particular educational program. The educational institution is responsible for ensuring that the student has appropriate educational opportunities throughout his or her educational career to improve and develop needed academic skills. Indeed, observing the differences between the employment and education settings, a federal court recognized:

> If tests predict that a person is going to be a poor employee, the employer can legitimately deny the person the job, but if tests suggest that a young child is probably going to be a poor student, a school cannot on that basis alone deny that child the opportunity to improve and develop the academic skills necessary to success in our society.

<u>Larry P. v. Riles</u>, 793 F.2d 969, 980 (9th Cir. 1984).[2]

Similarly, the question of test use cannot be examined in a vacuum. While the Resource Guide focuses specifically on the discriminatory use of tests which are used for high-stakes educational decisions, this issue must be considered in the context of the educational objectives involved and the effect of the particular testing practice in question upon students, particularly where classification of students and the provision of services is at issue. (Tab B of this Resource Guide lists policy and technical assistance documents that provide resource information and legal guidance relating to the nondiscriminatory classification of students and the provision of services to students.)

---

[2] <u>See also</u> National Research Council, <u>High Stakes Testing for Tracking, Promotion, and Graduation</u>, at pp. 61 - 62, 76 - 77, 97 (National Academy Press, Washington, D.C. 1999).

II. Scope of the Resource Guide

The Resource Guide does not apply to modifications of tests and/or testing conditions required for the purpose of accommodating individuals with disabilities under Section 504 of the Rehabilitation Act, the Individuals with Disabilities Education Act (IDEA) or the Americans with Disabilities Act (ADA).[3] Although the legal theories of discrimination discussed in the Resource Guide are generally applicable to disability issues that arise under Section 504, the IDEA and the ADA, an additional analysis regarding testing accommodations provided to individuals with disabilities is also required. See, e.g., Attachment A: Dear Colleague Letter (September 29, 1997) (addressing the inclusion of students with disabilities in statewide assessment systems). This analysis is beyond the scope of the Resource Guide.

The Resource Guide, along with the attached Compendium of Legal and Technical Resources (Appendix), should be read as an explanation of the legal and conceptual framework needed for understanding the issues raised by challenges to high-stakes testing. The model (and pragmatic) questions set out in Tab A should be viewed as a starting point for addressing questions of great complexity regarding challenges to testing and assessment practices. These model questions do not define the "floor" of what must be asked any more than they define the "ceiling" of what may be asked. Those decisions are inherently case-specific. Tab C provides a glossary of terms relating to test validity.

III. Foundations of the Resource Guide

A. Professional Standards

Generally-accepted professional standards for evaluating standardized tests provide a significant foundation for this guide. They include those described in the Standards for Educational and Psychological Tests prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education; the Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices; and the Uniform Guidelines on Employee Selection Procedures. At OCR's request, the National Academy of Sciences' Board on Testing and Assessment (BOTA) reviewed earlier drafts of this guide and provided comments, which have helped to ensure that the Resource Guide is consistent with existing professional standards.

---

[3] Section 504 of the Rehabilitation Act and Title II of the Americans with Disabilities Act are enforced by OCR; the Individuals with Disabilities Education Act is administered by the U.S. Department of Education's Office of Special Education Programs.

B. Legal Standards

This guide outlines two separate legal theories of discrimination: disparate treatment and disparate impact. Each theory is based on settled federal legal principles under Title VI of the Civil Rights Act of 1964 and Title IX of the Education Amendments of 1972 and neither breaks any new legal ground.

A disparate treatment analysis is used to determine whether a policy or practice regarding testing is being applied differently to an individual student or group of students because of their race, national origin, or gender, without legal justification for doing so, e.g., as a remedy for past de jure discrimination. This analysis would be used to determine, for example, whether black students and white students are being tested under different conditions or whether students with the same test scores are being treated differently by an educational institution.

Under a disparate impact analysis, the focus is on the "effects" of the application of a facially neutral policy or practice, regardless of whether the adverse consequences for a particular race, national origin, or gender were intended. The use of a disparate impact analysis is appropriate when the use of a test pursuant to a race-neutral policy or practice creates a significant difference in the granting or denial of benefits or opportunities on the basis of race, national origin or sex. Tests that have a disparate impact on the basis of race, national origin, or sex must be educationally necessary; otherwise, they are not permissible under Title VI or Title IX. Educational necessity involves a showing that the test is valid and reliable for the purpose for which it is being used[4]. The use of the test is still not permissible under Title VI or Title IX if the test is not the least discriminatory practical alternative that can serve the education institution's educational purpose. See Attachment B: Memorandum from the Attorney General for Heads of Departments and Agencies that Provide Federal Financial Assistance, "Use of the Disparate Impact Standard in Administrative Regulations under Title VI of the Civil Rights Act," July 14, 1994.

---

[4] Section III. B. of the Resource Guide contains a discussion of test validity and reliability. Tab C provides a glossary of terms relating to test validity.

RESOURCE GUIDE

I. Introduction

When tests are used to make educational decisions, they should be used to measure students' abilities, knowledge, or qualifications, regardless of race, national origin, or sex. Civil rights concerns arise when test uses do not satisfy federal antidiscrimination standards. This Resource Guide outlines the requirements of federal law prohibiting misuse of tests and other assessment procedures that result in discrimination based on race, national origin, or sex. It is designed to provide a general analytical framework under Title VI and Title IX for determining the proper use of tests and other assessment procedures in the educational context.

In evaluating a test or other assessment procedure, it is important to consider how the test is being used. In some cases, it may be used to make a certification or selection decision (e.g., admission to a school, awarding of a scholarship, or teacher certification). In other cases, it may be used to classify students (e.g., to identify students as needing special education or special language services or to identify students as gifted and talented).

When high-stakes educational decisions are made, tests may be used in conjunction with other criteria, such as teachers' recommendations. Ordinarily, if there are allegations or evidence regarding possible discrimination on the basis of race, national origin, or sex with respect to the use of a test or other criteria as part of a high-stakes decision making process, there should be an inquiry into the operation of the entire assessment process. There should be an inquiry into what criteria are being utilized as part of the entire process and the weight being given to each of the criteria in the process. This Resource Guide focuses on cases where the test or assessment contributes significantly to the high-stakes decision. However, if other criteria are contributing to a disparate impact on the basis of race, national origin, or sex, they should be evaluated to ensure that they are educationally appropriate and necessary, as well.

II. Basic Federal Standards

The requirements of Title VI and Title IX apply to all educational institutions that receive federal funds. These laws apply to all of the academic, athletic, and extracurricular programs of the institution, whether conducted in facilities of the recipient or elsewhere. Title VI prohibits race and national origin discrimination in programs and activities that receive Federal financial assistance. Title IX prohibits sex discrimination in education programs that receive Federal financial assistance. Title VI and Title IX cover the uses of property that the recipient owns and the activities that the recipient sponsors. Title VI and Title IX cover these operations, whether the individuals involved in a given activity are students, faculty, employees, applicants, or other participants. See Compendium at pp. 1 - 3.

Some federal courts have addressed challenges to the use of tests for high-stakes purposes under the equal protection and due process clauses of the Fourteenth Amendment to the U.S. Constitution. Although OCR enforces statutory rights under Title VI and Title IX rather than constitutional rights, to the extent the claim is that a school district's use of tests is discriminatory, those actions may violate both the statutes and the Constitution. OCR normally would not be involved, however, in cases in which there were no allegations of discrimination. Thus, those cases challenging the use of tests for constitutional reasons unrelated to discrimination would not fall within OCR's jurisdiction. Some federal cases in which discrimination claims have been raised have also involved equal protection challenges to a jurisdiction's use of tests in which the claim is based not on discriminatory intent but on the jurisdiction's use of tests to separate out those students who should not be allowed to graduate.[5] Under these circumstances, since there is no claim of discrimination based on membership in a suspect class, the equal protection claim is reviewed under the rational basis standard. The jurisdiction thus need show only that the use of the tests has a rational relation to a valid state interest. See Debra P. v. Turlington, 644 F.2d 397, 406 (5th Cir. 1981); Erik V. v. Causby, 977 F. Supp. 384, 389 (E.D.N.C. 1997).[6]

Due process challenges to the use of tests fall into two categories, substantive and procedural due process. Analyses under the due process clause address whether students have been denied, based on test scores, educational benefits or opportunities to which they had a legitimate claim of entitlement. Such cases typically involve a procedural due process claim that student were not given sufficient notice of the test and its requirements, or a substantive due process claim that the students were not taught the material on which the tests were based. Debra P., 664 F.2d at 404-405; Crump v. Gilmer Independent School District, 797 F. Supp. 552, 555-556 (E.D.Tex. 1992).

III. Disparate Impact Analysis

A disparate impact analysis may be applied to allegations involving discriminatory test use by educational institutions. Under this analysis, the use of any educational test which has a significant disparate impact on members of any particular race, national origin, or sex is discriminatory, and a violation of Title VI and/or Title IX, respectively, unless it is

---

[5] As a general matter, courts express reluctance to second guess a state's educational policy choices when faced with such challenges, although recognizing that a state cannot "exercise that [plenary] power without reasons and without regard to the United States' Constitution." Debra P. v. Turlington, 644 F.2d 397, 403 (5th Cir. 1981).

[6] Where, however, the use of a facially race-neutral test perpetuated the effects of the prior dual school system in which students were intentionally segregated on the basis of race, such a test could violate the equal protection clause even absent direct evidence of discriminatory intent. Debra P., 644 F.2d at 407, citing Arlington Heights v. Metropolitan Housing Development Corp., 429 U.S. 252, 266-268 (1977); see also Anderson v. Banks, 520 F. Supp. 472, 500 (S.D.Ga. 1981) (discriminatory impact of the test cannot be considered separately from the de jure discrimination that preceded it).

educationally necessary and there is no practicable alternative form of assessment which meets the educational institution's educational needs and would have less of a disparate impact on the basis of race, national origin, or sex.

In applying a disparate impact analysis, the following questions should be addressed:

A. Does the educational institution's use of an educational test result in the significantly disproportionate denial of an educational benefit or opportunity to members of a particular race, national origin, or sex?

B. If so, is the use of the test educationally necessary?

C. If so, do there exist practicable alternative forms of assessment which would substantially serve the school's stated purpose and are valid and reliable for that purpose, but which have less of a disparate impact on the basis of race, national origin, or sex?

Each question is discussed in more detail below. Where, based on evidence, there is a finding that the use of a test or assessment procedure caused or contributed to a disparate impact on members of a particular race, national origin, or sex (the first question), and the test or procedure does not meet the legal standard of educational necessity (the second question) or there is a practicable alternative form of assessment which would meet the educational institution's educational needs and would have less of a disparate impact on the basis of race, national origin, or sex (the third question), there is a violation of Title VI or Title IX under this disparate impact analysis.

A. Establishing Disparate Impact

Under a disparate impact analysis, a school's use of an educational test that causes or contributes to a disproportionate denial of an educational benefit or opportunity to members of a particular race, national origin, or sex is sufficient information to indicate a possible failure of compliance with Title VI or Title IX which should be investigated further. It is important to note that disparate impact by itself does not necessarily mean that discrimination has taken place. Disparate impact may lead to a finding of discrimination only when the use of the test in question is not educationally necessary or when there is no practicable alternative form of assessment which would meet the educational institution's educational needs and have less of a disparate impact on the basis of race, national origin, or sex.

B. Establishing Educational Necessity

Once it has been determined that a disparate impact exists, it must then be determined whether the use of the test or assessment procedure is educationally necessary.[7] To meet

[7] Where a test is being used as the sole or principal criterion for making educational decisions and where it was clearly not designed to be used as such, there is no basis upon which to conclude that the test is educationally necessary.

the educational necessity standard, the test or assessment procedure must be valid and reliable for the purpose for which it is being used.

In evaluating the validity and reliability of a test or assessment procedure, generally accepted professional standards should be the foundation for such decision making. These standards include the Standards for Educational and Psychological Testing prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education; the Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices; and the Uniform Guidelines on Employee Selection Procedures[8]. All decisions as to whether a test or procedure has met professionally accepted standards should be made in consultation with experts.

As discussed below, in determining whether a test or assessment procedure is educationally necessary, it must be shown that the test or procedure is valid and reliable for the purpose for which it is being used.

1. Technical Considerations

Validity

Establishing validity is the process of evaluating the degree to which a test measures what it claims to measure and leads to legitimate inferences that are appropriate or meaningful.[9] The demonstration of validity is multifaceted and depends on the type of assessment and the purposes for which the test was designed to be used.

Often, validity demonstrations will require careful analysis of data according to existing professional standards. This is a complex and specialized endeavor, and professionally accepted validation standards and techniques are evolving (for example, the 1985 Standards for Educational and Psychological Testing are currently being revised). Tab C contains a glossary of terms related to test validity.

---

[8] Although there are many principles in the Uniform Guidelines that apply to educational testing in general terms, the Uniform Guidelines do not address educational testing issues. There are critical, contextual differences between employment and educational testing that should not be overlooked when using the Uniform Guidelines as a resource in the educational setting. The Uniform Guidelines were adopted by and are currently used by the U.S. Equal Employment Opportunity Commission, the U.S. Department of Labor, and the U.S. Department of Justice.

[9] Indeed, it may not be technically correct to refer to a test or assessment procedure as being valid. Rather, it is the inferences and interpretation drawn from the responses to the test or procedure that must be valid. However, for simplicity§s sake, this guidance will often use the more common approach of referring to the test or procedure as being valid for the purpose for which it is being use.

In some cases, a test will clearly not be valid for the purpose for which it is being used. For instance, where a test manufacturer states that a test is not valid for use as a sole criterion in educational decision making, it is a clear misuse of that test if the school, in fact, uses only the test results in making a high-stakes educational decision.

Construct validity is relevant when an assessment is used to measure a particular characteristic, property, skill, ability, capacity, academic achievement, or behavior. The construct validation of a test usually involves a series of studies, using a variety of research methodologies.

The validation of constructs of academic content are relevant when a recipient is using a test to measure the acquisition of specific knowledge or academic skills. For example, a statewide proficiency test designed to measure whether students have learned specific skills or gained specific knowledge in order to determine whether they should receive a diploma would be subject to an assessment of the validity of the constructs of its content.

Criterion-related validity is relevant when scores on a test or assessment procedure are related to the examinee's performance on some other measure, which is known as a criterion. For example, when a recipient is using test scores to accept or reject applicants to a particular program, school, or curriculum, it should have evidence that the test scores correlate significantly with success in the program, school, or curriculum.

Reliability

Along with evidence of a test's validity, evidence of a test's reliability over time and over students should be considered and must conform to accepted professional standards.[10] Reliability is the degree to which test scores are consistent, dependable, or repeatable. For a test to be considered reliable, there should be evidence that the same students, taking the test multiple times with no change in preparation, receive corresponding scores. No test is perfectly reliable and differing amounts of error or unreliability are tolerated, depending upon the purposes for which the test or procedure is designed to be used. Reliability may be affected by the type of assessment procedure at issue, e.g., a standardized test versus a performance-based assessment.[11]

---

[10] The 1985 Standards for Educational and Psychological Testing discuss reliability. See APA Standards at pp. 19 - 23.

[11] Performance-based assessment requires students to generate rather than choose a response. Students are required to actively accomplish complex and significant tasks, while bringing to bear prior knowledge, recent learning, and relevant skills to solve problems. Demonstrations, written or oral responses, journals and portfolios are examples of performance-based assessment. Herman, J.L., Aschbacher, P.R., & Winters, L. (1992). A Practical Guide to Alternative Assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Fairness

Within the constraints of the defined purposes of a test or procedure, it is expected that the assessment will be valid and reliable for all students taking the assessment.[12] That is, there must be adequate evidence that the test is measuring the same academic constructs for all students, and that the results are sufficiently precise for all students.

Use

Assessment results can be used appropriately or inappropriately. Misuse can stem from two test-related considerations, as well as other problems in the decision-making process. That is, users may suggest a test or procedure is measuring what it is not, thereby producing invalid inferences. They may attempt to use results in making decisions which require a higher level of precision or reliability than the assessment is designed to produce. An example of this type of misuse is a school district using results from a test as a sole criterion in making a high-stakes decision when the test publisher has stated that the test is not to be used as a sole criterion.[13] The processes which users engage in to make decisions about individuals or groups may themselves be flawed, so that the results of tests with reasonably valid and reliable inferences are used inappropriately.

Invalid inferences can stem from misalignment between what is described as being assessed and what is actually being measured. It might also stem from a misalignment between curriculum goals or standards and what the high-stakes test or procedure is measuring, or between what is being assessed and what is being taught in classrooms. In each situation, the source of the misalignment must be established so that it can be determined where changes are needed. For instance, in determining whether a high-stakes test is being used appropriately, it may be appropriate to determine the degree to which schools provide instruction in the knowledge and

---

[12] The 1985 Standards for Educational and Psychological Testing address technical issues of fairness in testing. See e.g., APA Standards at standard 1.2, 1.5, 1.8, 1.10, 1.13, 3.5, and 3.10. See also Paul W. Holland & Howard Wainer, Differential Item Functioning (Lawrence Erlbaum Associates, Publishers 1993); National Research Council, High Stakes Testing for Tracking, Promotion, and Graduation, at pp. 78 - 82.

[13] See also APA Standards at standard 8.12 ("[i]n elementary and secondary education, a decision ... that will have a major impact on a test taker should not automatically be made on the basis of a single test score. Other relevant information for the decision should also be taken into account by the professionals making the decision."); National Research Council, High Stakes Testing for Tracking, Promotion, and Graduation, at p. 3.

skills measured by the test.[14] A statewide proficiency test could be subject to an analysis of the degree to which the schools in the State provide adequate instruction in the content areas measured by the test. Often, it is necessary to determine whether curriculum goals or standards have been clearly identified. If they have been clearly identified, then the alignment of instruction and assessment should flow from these standards. Misalignment would occur if either the instruction or assessment is not consistent with the standards.

2. Establishing Technical Merit

Tab A includes guidance on the types of questions to ask and information to obtain regarding the technical merit of assessments. These sample questions should be considered as starting points for appropriate inquiry. In most cases, these questions should be refined, modified, and supplemented based on the facts of the case and the advice of testing and/or other education experts.

The following guidelines should be considered when evaluating evidence of technical merit:

a. **No assumption of technical merit.** The general reputation of a test, its author, or its publisher, or casual reports of its validity are not evidence of a test's technical merit. A test is not considered technically viable under federal law based on a test's name or descriptive labels; promotional literature about the test; data regarding the frequency of a test's use; or testimonial statements and credentials of test publishers, consultants, or schools which have previously used the test. A publisher's test manual may provide technical evidence; this alone is not sufficient to determine technical merit.

b. **Acceptable types of evidence.** The use of a test should be supported by studies of the same test conducted by test publishers or professional researchers which demonstrate adequate validity and reliability for the particular use. Such studies must show that the use of the test by the school is the professionally accepted equivalent to the use for which the test was validated. The use of the test by the school should be within the technical parameters defined by the publisher and demonstrated by the evidence.

---

[14] Several federal court decisions have addressed the degree to which schools have provided adequate instruction in the knowledge and skills measured by a test. See Debra P. v. Turlington, 644 F. 2d 397, 405 (5th Cir. 1981); Crump v. Gilmer Independent School District, 797 F. Supp. 552, 555-6 (E.D. Tex. 1992). The inquiry regarding whether there is an alignment between knowledge and skills that are being tested and the curriculum and instruction that are being provided to students is critical when the test use in question involves an assessment of learning or achievement in school.

As one part of the process of showing that a test or procedure is technically sound, it may be appropriate to assess the degree of relationship between test scores and performance criteria. This may be done by researchers using professionally accepted research and statistical procedures.

3. Cutoff Scores

In determining whether a test or procedure with a disparate impact is educationally necessary, it is necessary to look to how the test or procedure is actually used by the recipient. In some cases, a test or assessment procedure may be used without a specific passing or cutoff score. In other cases, a score may be set, either by the test developer or the test user. Standard 6.9 of the 1985 Standards for Educational and Psychological Tests states that "[w]hen a specific cut score is used to select, classify or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report." This information must be considered in determining whether the cutoff score used by a recipient was set by some systematic process that reflects the good faith exercise of professional judgment.

C. Alternatives With Less Disparate Impact

Even if a school can show that a test or assessment procedure is valid and reliable, the school's continued use of the test or procedure may be in violation of federal law if one or more instruments, criteria, or procedures are available as a practicable alternative to the challenged test or procedure, <u>and</u> if any such alternative 1) substantially serves the educational purposes for which the test or procedure is used, 2) is valid and reliable for those purposes, and 3) would have a lesser disparate impact.

It is a good educational practice for school administrators to review the results of any testing program. If school administrators find that a test results in a disparate impact based on race, national origin or sex, the best educational practice is to inquire about other testing instruments or measures that would serve the school's educational purpose, be valid and reliable for that purpose, and have a less negative impact on students of a particular race, national origin, or sex.

IV. Different Treatment Analysis

If warranted by the nature and scope of the allegations or evidence, a different treatment analysis may be utilized, as described below, to determine whether the educational institution administered a test or assessment procedure differently or used scores differently because of the students' race, national origin, or sex, without a legitimate, nondiscriminatory reason. If the reason for the different treatment was, <u>e.g.</u>,: 1) the provision of testing accommodations or auxiliary aids to qualified individuals with disabilities as required by Section 504 of the Rehabilitation Act of 1973 or Title II of the Americans with Disabilities Act of 1991; or 2) voluntary or remedial affirmative action undertaken in accordance with federal law, the educational institution may have a legal justification for the different treatment and there may be no violation of federal law.

Otherwise, tests and assessment procedures must be administered and scores used in the same manner regardless of race, national origin, or sex. Even if a test or procedure is supported by sufficient evidence of educational necessity, an educational institution may still be in violation of Title VI and/or Title IX if the test or procedure is administered differently or the scores are used differently for students because of their race, national origin, or sex.

V. Equal Opportunity for Limited-English Proficient Students

The requirements of Title VI discussed above regarding the use of tests for making high-stakes educational decisions are applicable when tests are being used to make high-stakes educational decisions concerning students with limited English proficiency. Under Title VI and other federal laws, State educational agencies and school districts are required to ensure that students are not denied equal educational opportunities because a student has limited proficiency in English.

Limited English proficient (LEP) students must ordinarily be included in assessment programs. When LEP students are included in assessment programs, the inferences and interpretations drawn from the students' responses to the test or assessment procedure must be valid and reliable. Further, the students must be provided appropriate accommodations in order to ensure valid and reliable results. Accommodations might occur in the test format (including editing accommodations) and/or in the administration, response or scoring conditions. Depending upon the nature and purpose of the test and the particular needs of a LEP student, if students are literate in their native language, and if the instruction has been in that language, providing a valid and reliable version of the test in the student's native language might be an appropriate accommodation. Other accommodations may include extended time or the use of bilingual dictionaries. If students are excluded from assessment programs, based on legitimate educational or psychometric justifications for their exclusion, comparable information about their academic progress must be collected for these students.

VI. Analysis Where Prior Dual System

School districts that have operated dual systems and have not been declared unitary have an obligation to dismantle their prior de jure segregated systems. The use of any educational test or assessment procedure may be a violation of Title VI if it had been used to achieve the segregation or if it perpetuates the segregation. Where such tests or assessment procedures are being used, school districts have an obligation to identify, consider and implement less discriminatory criteria consistent with sound educational policy, to the extent practicable.

VII. Remedies

Where an educational institution is in violation of Title VI or Title IX, there are a range of remedies that may be used to come into compliance with federal law. Remedies should be designed to ensure that educational institutions comply with civil rights statutes when meeting their educational goals. Depending on the facts of a given case, there are many permissible responses to correcting a violation. If the administration or design of a test is discriminatory on the basis of race, national origin, or sex, appropriate remedies might include: supplementing the use of the test with other assessment measures; revising the test instrument within a reasonable period of time to address compliance concerns; or substituting the test with another available instrument that more appropriately measures what is intended to be measured. If the test or assessment procedure reflects discriminatory educational practices with respect to the adequacy of instruction provided to students to prepare them to take the test, an appropriate remedy might include enhancing learning opportunities for students to perform well on the test.

APPENDIX: COMPENDIUM OF LEGAL AND TECHNICAL RESOURCES

This compendium provides an outline of key legal and technical resources to serve as a reference for inquiries regarding potential discrimination in the use of an educational test or assessment procedures.

The investigation and analysis of disparate impact cases under Title VI of the Civil Rights Act of 1964 (Title VI), 42 U.S.C. § 2000d, and Title IX of the Education Amendments of 1972 (Title IX), 20 U.S.C. § 1681, rely, to a large extent, on case law developed under Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e, which prohibits discrimination on the basis of race, color, national origin, sex, and religion in employment. See United States v. LULAC, 793 F.2d 636, 648-49 (5th Cir. 1986); Georgia State Conference of Branches of NAACP v. Georgia, 775 F.2d 1403, 1417 (11th Cir. 1985); NAACP v. Medical Center, Inc., 657 F.2d 1322 (3rd Cir. 1981); Dillon County District No. 1 and South Carolina State Department of Education, No. 84-VI-16 (Civil Rights Reviewing Authority 1987).

I. Basic Federal Standards

A. Title VI and Title IX Prohibit Discrimination in Federally Funded Programs and Activities

Title VI prohibits race and national origin discrimination in programs and activities that receive Federal financial assistance. Title IX prohibits sex discrimination in education programs that receive Federal financial assistance. See also 34 C.F.R. Part 100 (regulations implementing provisions of Title VI) and 34 C.F.R. Part 106 (regulations implementing provisions of Title IX). Under the Civil Rights Restoration Act of 1987, OCR generally has institution-wide jurisdiction over a recipient of Federal funds. See 42 U.S.C. § 2000d-4 (1989).

B. Specific Discriminatory Actions Prohibited By Title VI and Title IX[15]

The regulations implementing Title VI do not specifically address the use of tests and assessment procedures, but do include a general provision prohibiting discrimination based on race or national origin. 34 C.F.R. § 100.3(b)(5).

The regulations implementing Title IX specifically prohibit the discriminatory use of tests or assessment procedures in admissions, 34 C.F.R. § 106.21, employment, 34 C.F.R. § 106.52, and counseling, 34 C.F.R. § 106.36. Title IX further prohibits discrimination in areas in which test or assessment procedure results are often used to allocate benefits and opportunities. See 34 C.F.R. § 106.37(a) (prohibition against discrimination in financial aid awards and against assisting any entity which provides financial aid to students in a manner which discriminates based on sex); 34 C.F.R. § 106.31(b)(6) (prohibition against providing "significant assistance" to entities which discriminate on the basis of sex in providing any aid, benefit or service to students or employees).

See also 34 C.F.R. § 100, Appendix B, part K (Guidelines for Eliminating Discrimination and Denial of Services on the Basis of Race, Color, National Origin, Sex, and Handicap in Vocational Education Programs) ("if a recipient can demonstrate that . . . criteria [that disproportionately exclude persons of a particular race, color, national origin, sex, or disability] have been validated as essential to participation in a given program and that

---

[15] Some federal courts have addressed challenges to the use of tests for high-stakes purposes under the equal protection and due process clauses of the Fourteenth Amendment to the U.S. Constitution. Although OCR enforces statutory rights under Title VI and Title IX rather than constitutional rights, to the extent the claim is that a school district's use of tests is discriminatory, those actions may violate both the statutes and the Constitution. Some federal cases in which discrimination claims have been raised have also involved equal protection challenges to a jurisdiction's use of tests in which the claim is based not on race or sex discrimination, but on the jurisdiction's use of tests to determine, for example, those students who should be allowed to graduate. Under these circumstances, the claim is reviewed under the rational basis standard and the jurisdiction need show only that the use of the tests has a rational relation to a valid state interest. See Debra P. v. Turlington, 644 F.2d 397, 406 (5th Cir. 1981); Erik V. v. Causby, 977 F. Supp. 384, 389 (E.D.N.C. 1997).

Due process challenges to the use of tests fall into two categories, substantive and procedural due process. Analyses under the due process clause address whether students have been denied, based on test scores, educational benefits or opportunities to which they had a legitimate claim of entitlement. Such cases typically involve a procedural due process claim that students were not given sufficient notice of the test and its requirements, or a substantive due process claim that the students were not taught the material on which the tests were based. Debra P., 644 F.2d at 404-405; Crump v. Gilmer Independent School District, 797 F. Supp. 552, 555-556 (E.D.Tex. 1992); cf. Williams v. Austin Independent School District, 796 F. Supp. 251 (W.D.Tex. 1992).

alternative equally valid criteria that do not have such a disproportionate adverse effect are unavailable, the criteria will be judged nondiscriminatory. Examples of admission criteria that must meet this test or assessment procedure are ... interest inventories ... and standardized test or assessment procedures").

II. Disparate Impact Analysis

Because the regulations that implement Title VI and Title IX incorporate an effects standard, a recipient's use of facially neutral policies that have a disparate impact on the basis of race, national origin, or sex may constitute a violation of Title VI or Title IX. See 34 C.F.R. § 100.3(b)(2); 34 C.F.R. § 106.21(b)(2); 34 C.F.R. § 106.36(b); 34 C.F.R. § 106.52. See also Guardians Assn. v. City Service Commission of City of N.Y., 463 U.S. 582 (1983) Lau v. Nichols, 414 U.S. 563 (1974). See also Attachment B: Memorandum from the Attorney General for Heads of Department and Agencies that Provide Federal Financial Assistance, "Use of the Disparate Impact Standard in Administrative Regulations under Title VI of the Civil Rights Act," July 14, 1994.

A. Establishing Disparate Impact

There is no rigid mathematical threshold that must be met to demonstrate a disparate impact. Watson v. Fort Worth Bank and Trust, 487 U.S. 977, 994-95 (1988)("statistical disparities must be sufficiently substantial to raise ... an inference of causation," i.e., "show that the practice in question caused the exclusion of applicants for jobs or promotions because of their membership in a protected group").

Groves v. Alabama State Board of Education, 776 F.Supp. 1518, 1523-1529 (M.D. Ala. 1991)(discussion on establishing a statistical prima facie case of disparate impact).

Georgia State Conference of Branches of NAACP, supra at 1421 ("Generally, to establish a prima facie case of disparate impact based on race the plaintiffs must show that the defendants' racially neutral practice detrimentally affects persons of a particular race to a greater extent than other races. . . ").

B. Establishing Educational Necessity

The use of an educational test or assessment procedure which has a disparate impact on members of any race, national origin, or sex group is discriminatory, and a violation of Title VI or Title IX, unless the recipient justifies the use as educationally necessary. See Board of Education v. Harris, 444 U.S. 130, 151 (1979)(in disparate impact cases in the education context defendants are required to show an educational necessity instead of a business necessity); Griggs v. Duke Power Company, 401 U.S. 424 (1971)(sets similar standard for disparate impact of an employment test or assessment procedure); Branches of NAACP v. State of Georgia, 775 F.2d 1403, 1417 (11th Cir.

1985); and Sharif v. New York State Education Department, 709 F.Supp. 345 (S.D. N.Y. 1989)(standard for disparate impact of an educational test or assessment procedure is educational necessity).

See also Memorandum from then Acting Assistant Secretary for Civil Rights, Alicia Coro, to then Acting Regional Civil Rights Director, Region V, Linda A. McGovern (PCD # 70 October 22, 1986)(standard in case involving alleged disparate impact of LSAT scores as an admissions criterion at DePaul University College of Law and Illinois Institute of Technology/Chicago/Kent College of Law, is educational necessity).

Whether a test or assessment procedure is educationally necessary depends on whether the test or assessment procedure is valid for the purpose for which it is being used. Sharif, supra; State of Georgia, supra; cf. Final Order of the Civil Rights Reviewing Authority, Dillon County School District No. 1, Docket No.84-IV-16.

The Guidance is consistent with professional standards - See APA Standards and the Code of Fair Testing Practices in Education.

1. Validity

Memorandum from Harry M. Singleton to John E. Palomino (PCD # 57 April 4, 1985)(valid test or assessment procedures "successfully measure what they claim to measure; are used only for the specific purpose(s) for which they were developed; and, are administered in conformance with the instructions provided by the publisher").

See APA Standards at p. 11 (defining criterion-related evidence) and standard 1.11, 1.12, 1.18 (describing criterion-related validation studies).

See APA Standards at p. 10 (defining content-related evidence) and standard 1.6 and 1.7 (describing content-related validation studies).

See APA Standards at p. 9 (defining construct-related evidence) and standard 1.8, 1.9, and 1.10 (describing construct-related validation studies).

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979), aff'd, 793 F.2d 969 (9th Cir. 1984)(State of California, requesting approval from the court to use standardized IQ test or assessment procedures for the purpose of placing black children in EMR classes, required to, among other things, provide statistics showing the mean scores of blacks and whites on the test or assessment procedure and information supporting the validity of the test or assessment procedure for the purpose of identifying and placing students in EMR classes).

Debra P. v. Turlington, 730 F.2d 1405 (11th Cir. 1984)(court approved validity study which consisted of a number of surveys and site visits that analyzed whether the students had received the instruction necessary for them to have mastered the skills that were being tested).

See also Code of Fair Testing Practices in Education at A-1 and A-7 (encouraging test developers to describe the population for which the test is appropriate and encouraging test users to select tests appropriate for the testing purpose and population of test takers).

American Psychological Association Standards for Educational and Psychological Testing (1985) (APA Standards) at standard 1.16 (permitting criterion-related validation inferences to be drawn from a set of prior studies, where "local validation evidence" is not available, depending on the degree of similarity between the test or assessment procedure use and validation sample); and at pp. 12-13 (describing the concept of differential prediction).

2. Reliability

APA Standards at pp. 19-20 (discussing reliability and error of measurement).

3. Cutoff scores

Evans v. City of Evanston, 881 F.2d 382 (7th Cir. 1989) (while test was valid for the job, cutoff score was set one standard deviation above the mean; the court rejected this because there was no attempt to connect the score to level of performance: "...the ability to perform firefighting tasks adequately depends not on relative but on absolute test performance.").

Richardson v. Lamar County Bd. of Education, 729 F. Supp. 806 (M.D. Ala. 1989) (passing score rejected because of the lack of any relationship to actually measuring competence; instead, based on what was "politically acceptable").

APA Standards at standard 6.9 (when a specific cut-off score is used to select, classify or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report).

C. Alternatives With Less Disparate Impact

Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975)(in Title VII case challenging use of employment test that had a disparate impact, court stated that employee can still prevail, even if test is valid, if other tests or selection devices with less disparate impact would serve the employer's interests).

NAACP v. State of Georgia, 775 F.2d. 1403 (11th Cir. 1985) (considering less discriminatory alternatives in Title VI education context).

Sharif v. New York State Education Department, 709 F.Supp. 345 (S.D.N.Y. 1989) (where use of the SAT had a disparate impact based on sex in awarding state merit scholarships, court approved awards being based on a combination system - using both grade point averages and SAT scores - as a legally sufficient alternative to sole reliance on the SAT; court found that, compared with sole reliance on the SAT, combination system would better advance the state's goal of awarding high school performance and would better provide all students with an equal opportunity to compete for prestigious state scholarships; court found that feasibility argument about the combination system advanced by the state education agency lacked merit).

Bridgeport Guardians, Inc. v. City of Bridgeport, 735 F. Supp. 1126, 1136-1137 (D. Conn. 1990), aff'd, 933 F.2d 1140, 1148 (2nd Cir. 1991), cert. denied, 502 U.S. 924 (1991) (where making promotion decisions for police department on a strict rank-order basis based on examination scores had a disparate impact on racial minority candidates, district court rejected the use of video simulations as an alternate selection criteria on the basis that they have generally not increased the relative standing of minority candidates and because it substantially adds to the cost of the promotion process; district court approved use of banding as an alternate selection criteria as it found that there is no evidence that any added burdens that a banding analysis would impose are more than minimal; appeals court upheld the use of banding noting, based on testimony from the city's industrial psychologist, that small variances in the examination scores did not indicate that there were real differences in the qualifications of the candidates).

Brunet v. City of Columbus, 1 F.3d 390, 411-412 (6th Cir. 1993), cert. denied, 510 U.S. 1164 (1994) (where making hiring decisions for fire fighter positions on a strict rank-order basis based on the results of a physical capability test (PCT) and a cognitive ability test (CAT) had a disparate impact on the basis of sex, although the appeals court rejected both alternatives presented by the plaintiffs, the appeals court found error because there was no indication in the record that the district court required the city, pursuant to the Uniform Guidelines on Employee Selection Procedures, to conduct its own investigation of viable alternatives with lesser or no impact on female applicants before implementing the strict rank-order process; the appeals court also found nothing in the record that requires the CAT and the PCT to be weighted equally; the appeals court also indicated that the city should be required to demonstrate why the CAT, which arguably is more predictive than the PCT, should not be weighted more than the PCT, noting that the change might result in a lesser disparate impact on women).

FairTest v. College Entrance Examination Board and Educational Testing Service, OCR Case No. 02-94-2048 (where use of the PSAT had disparate impact based on sex in selecting National Merit Scholarship semi-finalists, recipients agreed to modify the test to include a writing skills component and to study whether academic records could also be considered).

III. Different Treatment Analysis

As with other claims of race, national origin, or sex discrimination under Title VI and Title IX, a different treatment analysis may apply when a policy or practice regarding testing or assessment is being applied differently by an educational institution to different groups of students because of their race, national origin, or sex. This is the touchstone of what is a classic violation of Title VI and Title IX and their implementing regulations.

Where there is direct evidence that an educational decision was made based on race, national origin, or sex, a prima facie violation of Title VI or Title IX has been established. The recipient then has the burden of establishing a legitimate reason (an affirmative defense) for the different treatment, such as showing that the disparate treatment was the result of a valid affirmative action plan. See Regents of the University of California v. Bakke, 438 U.S. 265, 320 (1978) (Although the U.S. Supreme Court found that an applicant to medical school had been discriminated against on the basis of race under an unlawful admissions process, five justices agreed that the portion of the lower court decision that enjoined the university from ever considering the race of any applicant should be reversed. Justice Powell recognized that the "State has a substantial interest that legitimately may be served by a properly devised admissions program involving the competitive consideration of race or ethnic origin.") But see Hopwood v. Texas, 78 F.3d 932 (5th Cir. 1996), cert. denied, 518 U.S. 1033 (1996) (The Fifth Circuit held that a university's interest in achieving a diverse student body can never constitute a compelling governmental interest justifying the use of race in university admissions selections.)

Note that there need not be direct proof of intentional discrimination in order to make a disparate treatment case. "In most disparate treatment cases, intent to discriminate is established inferentially, through circumstantial evidence." Equal Employment Opportunity Commission's "Revised Enforcement Guidance on Recent Developments in Disparate Treatment Theory," July 14, 1992, Number N 915.002. at 2. The basic elements of a different treatment case in which there is no direct evidence of discrimination were set out by the U.S. Supreme Court in McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973), a Title VII employment case. See also United States Postal Service Board of Governors v. Aikens, 460 U.S. 711 (1983); Texas Department of Community Affairs v. Burdine, 450 U.S. 248 (1981).

IV. Analysis Where Prior Dual System

United States v. Fordice, 112 S.Ct. 2727 (1992) (Mississippi's admission policy which required higher ACT scores for historically white public universities than for historically black public universities was constitutionally suspect as it was originally enacted by historically white universities to maintain prior dual system of higher education. States and schools districts that have operated a dual system have an obligation to dismantle the prior de jure segregated system.).

Knight v. Alabama, 14 F.3d 1534, 1540-42 (11th Cir.1994) (A recipient's "burden of proving that [less discriminatory] alternatives are impractical or educationally unsound is a heavy one.").

Debra P. v. Turlington, 644 F.2d 397, 407 (5th Cir. 1981) (In an equal protection challenge to the requirement that students pass a functional literacy test or assessment procedure to receive a high school diploma, trial court instructed to consider whether disproportionate failures of black students could be attributed, in part, to unequal education received during period of dual school system).

TAB A

SAMPLE QUESTIONS FOR EVALUATING EVIDENCE OF EDUCATIONAL NECESSITY AND DETERMINING WHETHER THERE ARE PRACTICABLE ALTERNATIVES WITH LESS DISPARATE IMPACT

Once it has been determined that a test or assessment procedure has a disparate impact on students of a particular race, national origin, or sex, there should be an inquiry into the test or assessment procedure's educational necessity. Educational necessity means that the challenged test or assessment procedure is valid and reliable for the purpose for which it is being used. There should then be an inquiry to determine whether there are any practicable alternatives to the test or assessment procedure that are available, which would (i) substantially serve the educational purpose identified by the educational institution, (ii) be valid and reliable for that purpose, and (iii) have a lesser disparate impact.

As stated in the resource guide, evidence of the general reputation of a test or assessment procedure is not sufficient to establish validity. Also, a manual developed by the test or assessment procedure publisher is not presumptive evidence of validity. Rather, the use of a test or assessment procedure may be supported by validity studies of the same test conducted by the school, other schools, test publishers or distributors, or professional researchers.

The following questions are designed to assist OCR staff and members of the educational community in evaluating evidence of educational necessity and in determining whether there are any practicable alternatives to the test or assessment procedure which would meet the educational institution's educational needs and have a lesser disparate impact:

1. What test or assessment procedure is the educational institution administering?

2. For what purpose is the test or assessment procedure being administered (purpose should be stated in specific terms, i.e., predicting grades in algebra, rather than in general terms, i.e., measuring intelligence of seventh-graders)?

3. What is the educational institution's justification for the purpose for which the test or assessment procedure is being used? If the school cannot show that the test is educationally necessary, the use of the test or assessment procedure will be in violation of Title VI or Title IX, as appropriate.

4. For what purpose was the test or assessment procedure developed (if it is clear from preliminary evidence)? Is the school using the test or assessment procedure for this purpose? If not, the use of the test or assessment procedure cannot be justified as educationally necessary and will be in violation of Title VI or Title IX, as appropriate.

5. Is the test or assessment procedure being used as the sole criterion for making an educational decision? If so, was the test or assessment procedure designed to be used this way? If not, the test or assessment procedure used cannot be justified as educationally necessary and there is a violation of federal law.

6. What is the racial, ethnic, and/or sex composition of the test or assessment procedure-taking population?

7. Does the school have evidence that it has developed or that has been developed by the test or assessment procedure publisher that the test or assessment procedure is valid and reliable for the purpose for which it is being used?

a. What is the form of the evidence, i.e., study or report? Raw data or very general information is not acceptable evidence of the validity of a test or assessment procedure.

b. Who conducted the study, i.e., the school, another school, the test or assessment procedure developer, an independent researcher?

c. How long ago was the study conducted?

d. Is there evidence under professionally accepted standards that the test or assessment procedure is valid and reliable?

e. For what use was the test or assessment procedure validated? Is this use identical to, or the professionally accepted equivalent of, the purpose for which the school is using the test or assessment procedure?

f. Was the test or assessment procedure administered and scored properly?

i. Were all the students treated the same way as to how the test or procedure was administered or scored?

ii. Did school officials clearly articulate to school personnel who administered, scored, or interpreted the results the construct(s) or variable(s) the procedure was designed to measure?

iii. Did those who administered, scored, or interpreted the results have the appropriate skills to perform these functions adequately? For example, are those who interpret the scores able to understand and interpret commonly reported scores, such as percentile ranks, standard scores, stanines, normed curve equivalents, and grade equivalents (as appropriate to the particular test)? Are they able to understand and interpret commonly reported summary indexes such as central tendency measures, estimates of reliabilities, and standard errors of measurement (again, as appropriate to the particular test)?

8. What type of statistical test was used in the validation study upon which the school relies?

9. If the school has empirical evidence that the test or assessment procedure is valid and reliable for the purpose for which it is being used, has the school looked at whether there are practicable alternatives to the test or assessment procedure with less impact that would substantially serve the school's stated purpose and that are valid and reliable for that purpose?

a. How great was the disparate impact caused by the challenged test or assessment procedure?

b. What alternatives has the school looked at?

c. Did the criteria or criterion have less impact than the challenged test or assessment procedure?

d. If the school did not look at alternatives, what is the reason for this decision?

e. If the school did look at alternatives, but chose not to use them, what is the reason for this decision?

10. Are there one or more criteria which, either alone, or in combination with other criteria, would have less impact, serve the school's educational purpose, and be valid and reliable for that purpose?

11. Does the school use a "cutoff score" on the test or assessment procedure which determines whether a student receives an educational benefit or opportunity? If so:

a. Does the cutoff score have a disparate impact on students of a particular race, national origin, or sex?

b. Is the cutoff score being used as the sole criterion for making an educational decision? If so, was the cutoff score designed to be used in this way? If not, the use of the cutoff score cannot be justified as educationally necessary and there is a violation of federal law.

c. Does the school have evidence regarding the method and rationale for setting the cutoff score?

d. Does the evidence provided under c. reflect a systematic process that evidences the good faith exercise of professional judgment?

TAB B

LIST OF POLICY AND TECHNICAL ASSISTANCE DOCUMENTS ADDRESSING THE CLASSIFICATION OF STUDENTS AND THE NONDISCRIMINATORY PROVISION OF SERVICES[16]

I. Provision of Educational Services to Limited-English Proficient Students

U.S. Department of Education, Office for Civil Rights. (1991). "Policy Update on Schools' Obligations Toward National Origin Minority Students With Limited-English Proficiency."

> This policy update is primarily designed for use by OCR staff to determine whether schools are complying with their obligation under Title VI to provide any alternative language programs necessary to ensure that national origin minority students with limited-English proficiency (LEP) have meaningful access to schools' programs. The policy update provides additional guidance for applying the May 1970 and December 1985 memoranda, which are described below.

U.S. Department of Education, Office for Civil Rights. (1985). "The Office for Civil Rights' Title VI Language Minority Compliance Procedures."

> This memorandum provides a description of the procedures followed by OCR in making determinations of compliance with Title VI as regards the treatment of LEP students in educational programs that receive federal financial assistance from the Department of Education.

U.S. Department of Education, Office for Civil Rights. (1970). "Identification of Discrimination and Denial of Services on the Basis of National Origin," 35 Fed. Reg. 11595 (May 1970 Memorandum)

> This memorandum was designed to clarify the policy of the Department of Health, Education and Welfare, now the Department of Education, on issues concerning the responsibility of school districts to provide equal educational opportunity to LEP students. This memorandum was a foundation for the

[16] The policy and technical assistance documents that are listed were either developed by the U.S. Department of Education's Office for Civil Rights or were developed by other organizations, where the U.S. Department of Education was a contributing party.

U.S. Supreme Court's decision in Lau v. Nichols, 414 U.S. 653 (1974). The memorandum was affirmed by the U.S. Supreme Court in the Lau decision.

August D. & Hakuta, K. (Eds.). (1997). Improving America's Schools for Language-Minority Children: A Research Agenda. Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students, National Research Council. Washington, D.C.: National Academy Press.

The National Research Council (NRC) completed an extensive study of LEP students, which is summarized in this report. The report provides a review of the state of knowledge regarding the education of LEP students and identifies a research agenda that will address key gaps in present knowledge on the topic. Among the topics covered in the report are student assessment and program evaluation.

U.S. Department of Education, Office for Civil Rights. (1991). "Technical Assistance Resource Package on the Provision of Equal Educational Opportunity to National Origin Minority and Native American Students Who Are Limited-English-Proficient."

This resource package focuses on the provision of equal educational opportunities to national origin minority and native American students who are limited-English-proficient. The package contains: 1) information about the history and importance of issues pertaining to the education of LEP students; 2) technical definitions and explanations of prominent educational approaches to teaching LEP students; 3) a summary of case law regarding civil rights requirements for educating LEP students; 4) a summary of compliance review letters of findings; 5) summaries and listings of major research publications, studies, and reports addressing LEP issues; and 6) a listing of major professional and beneficiary organizations involved in gathering information about LEP issues.

II. Minority Students and Special Education

U.S. Department of Education, Office for Civil Rights. (1995). "Minority Students and Special Education - Legal Approaches for Investigations."

This memorandum provides an overview of the legal theories and approaches to be employed in OCR investigations regarding the disproportionate representation of minority students in special education.

Markowitz, J., García, S., & Eichelberger, J.H. (1997). Addressing the Disproportionate Representation of Students from Racial and Ethnic Groups in Special Education: A Resource Document. Alexandria, VA: Project FORUM, National Association of State Directors of Special Education.

This document is intended to enhance the knowledge base of technical assistance providers to enable them to provide more effective technical assistance and guidance to state and local education personnel who are addressing the problem of disproportionate representation. The document includes a compilation of approaches that have the potential for effectively preventing and correcting disproportionate representation, an annotated bibliography of print resources, and a list of individuals who are knowledgeable about one or more of the main topic areas presented in this document.

Morison, P., White, S.H., & Feuer, M.J. (Eds.) (1996). The Use of IQ Tests in Special Education Decision Making and Planning: Summary of Two Workshops. Board on Testing and Assessment, National Research Council. Washington, D.C.: National Academy Press.

To assist the U.S. Department of Education, the Board on Testing and Assessment convened two workshops to facilitate an examination and discussion of research evidence regarding the uses of intelligence tests (IQ) tests in special education placement decisions, with particular focus on mental retardation and learning disabilities. The workshops had the following objectives: 1) to provide an overview of legal, policy, and measurement issues regarding the use of IQ tests in special education; 2) to examine issues related to the validity and fairness of IQ testing for classification and placement of students in special programs, with emphasis on potential adverse effects on minority students; and 3) to explore some possible alternative assessment methods that could be used in combination - or as substitutes for - traditional IQ tests. The summary report provides a synthesis of the key themes and ideas discussed at the workshops.

TAB C

VALIDITY, A GLOSSARY

Introduction

"Validation is the most important consideration in test evaluation. ...Test validation is the process of accumulating evidence to support ... inferences (made from test scores)." (American Psychological Association Standards for Educational and Psychological Testing (1985). (APA Standards at p. 9).

In general the inference that a test or assessment procedure is valid is justified when the research evidence indicates the following is true:

1. the test or assessment procedure measures the construct (characteristic, property, skill, ability, capacity, or behavior) it was intended to measure;

2. the test or assessment procedure is used in a correct and appropriate manner, with regard to testing setting, testing procedure (including the qualifications of the test-giver and the manner in which the test is given), tested sample of people (e.g., using a test validated for adults to assess children would be improper); and

3. the inferences drawn from the resulting test or assessment procedure data are appropriate and correct.

"Traditionally, the various means of accumulating validity evidence have been grouped into categories called *content-related, criterion-related, and construct-related evidence of validity*. ... These categories are convenient ... but the use of the category labels does not imply there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use ... An ideal validation includes several types of evidence, which span all three of the traditional categories ... Professional judgment should guide the decisions regarding the forms of evidence that are most necessary and feasible in light of the intended uses of the test and any likely alternatives to testing." (APA Standards at p. 9).

Evidence can be gathered by use of such particular statistical techniques as correlation and regression analyses with test items or scores and other test or non-test variables, factor analysis, item response theory (IRT) and other level of difficulty techniques, and differential item functioning (DIF) analyses. It might also be gathered by the systematic judgment evaluation of individual responses, or a formal evaluation of one or a number of test construction, implementation, or data analytic processes. (Aiken, 1994; Holland and Wainer, 1993; Wainer and Braun, 1988).

Regardless of technique, evidence is obtained which demonstrates that information derived from the assessment is accurately reflective of what is supposed to be measured. This is done by designing investigations which focus on convergence, or high positive relationships, and discriminatory analyses, which seek to demonstrate divergence between the assessment information and related, but distinct, variables.

Construct-Related Validity[17]

Construct-related evidence of validity is "evidence that supports a proposed construct interpretation of scores on a test based on theoretical implications associated with the construct label." (APA Standards at p. 90). The construct-related validity of a test or assessment procedure is the extent to which the assessment may be said to measure a theoretical construct or trait. (Aiken, 1994; Anastasi, 1988; Groth-Marnat, 1990). "Reasoning ability, spatial visualization, and reading comprehension are constructs ...The construct of interest for a particular test should be embedded in a conceptual framework ... The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables." (APA Standards at pp. 9 - 10).

Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations represents appropriate evidence for this validation. (Anastasi, 1988). AThe 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. "[V]alidating inferences about a construct also requires paying careful attention to aspects of measurement such as test format, administration conditions, or language level, that may affect test meaning and interpretation materially." (APA Standards at p. 10).

Construct validity, which is the most general type of validity, is not determined by a single way or by one investigation. Rather it involves a network of investigations and other procedures designed to determine whether an assessment instrument that purportedly measures a certain variable is actually doing its job. (Aiken, 1994; Groth- Marnat, 1990).

---

[17] As indicated throughout the Resource Guide, OCR relies upon generally accepted existing professional standards when evaluating the validity and reliability of a test or assessment procedure. However, it should be noted that there is a trend among measurement theorists to consider construct validity to be the fundamental, unifying framework for conceptualizing validity evaluations (see, e.g., Shepard, 1993, and Wainer and Braun, 1988). Under this framework, since all validation is subsumed under construct validation, there are not different types of validity. Also, as part of this framework, various sources of evidence, including, but not necessarily limited to, content-related evidence, criterion-related evidence, and prediction-related evidence, can be, and usually are, used to evaluate the degree to which score-based inferences and actions are supported. Some testing and assessment experts include such additional evidence as the consequences of test use on individuals and groups in society as part of the construct validity framework (Messick, 1989).

Content-Related Evidence of Validity

Content-related evidence of validity is "evidence that shows the extent to which the content domain of a test is appropriate relative to its intended purpose." (APA Standards at p. 90). "In general, content-related evidence demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content. ... For some educational decisions, it is important to determine the agreement between the test and the curricular or instructional domains it is meant to cover. ... [I]nferences about content are linked to test construction as well as to establishing evidence of validity after a test has been developed and chosen for use." (APA Standards at pp. 9 - 10).

Content validity is concerned with whether the content of the test or assessment procedure elicit the range of responses representing the entire domain or universe of skills, understandings, or other behaviors that the test or assessment procedure was supposed to measure (Gregory, 1992; Aiken, 1994; Anastasi, 1988). "Methods classed in the content-related category should often be concerned with the ... construct underlying the test as well as the character of test content. There is often no sharp distinction between test content and test construct." (APA Standards at p. 11.).

Criterion Validity[18]

Criterion-related evidence of validity "demonstrates that test scores are systematically related to one or more outcome criteria." (APA Standards at p. 11). This type of validity evidence is produced by relating scores on the test or assessment procedure to performance criterion measures, standards, or variables (Aiken, 1994). According to Anastasi (1988), a criterion is a direct and independent measure of that which the test is designed to predict. Criterion-related procedures indicate the effectiveness of the assessment where performance on the test is checked against a criterion. "The choice of the criterion and the measurement procedures used to obtain criterion scores are of central importance. Logically, the value of a criterion-related study depends on the relevance of the criterion measure that is used." (APA Standards at p. 11).

Two types of criterion-related evidence are those obtained from investigations which focus on prediction and those which focus on concurrent relationships. "Two designs for obtaining criterion-related evidence - predictive and concurrent - can be distinguished. A predictive study obtains information about the accuracy with which early test data can be used to estimate criterion scores that will be obtained in the future. A concurrent study serves the same purpose, but it obtains prediction and criterion information simultaneously. Predictive

[18] Correlations between a test and a criterion are validity coefficients. A study of predictive or concurrent validity is nearly always reported in terms of a correlation coefficient. Cronbach, L.J. (1990). Essentials of Psychological Testing (5th ed.). New York:Harper Collins Publishers, Inc.

studies are frequently, but not always, preferable to concurrent studies of selection tests for education or employment, whereas concurrent evidence is usually preferable for achievement tests, tests used for certification, diagnostic clinical tests, or for tests used as measures of a specified construct." (APA Standards at p. 11).

TAB D

BIBLIOGRAPHY

Selected Validity References

Aiken, L.R. (1994). Psychological Testing and Assessment, (8th ed.). Needham Heights, MA: Allyn and Bacon.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing.

Anastasi, A. (1988). Psychological Testing (6th ed.). New York: Macmillan Publishing Company.

Anastasi, A. (1986). "Evolving Concepts of Test Validation." Annual Review of Psychology, 37, 1-15.

Carver, C.S., & Schier, M.F. (1996). Perspectives on Personality (3rd ed.). Needham Heights, MA: Akkyn & Bacon.

Cattell, R.B. (1964). "Validity and reliability: A proposed more basic set of concepts." Journal of Educational Psychology, 55(1), 1-22.

Corsini, R.J, & Marsella, A.J. (1983). Personality Theories, Research, and Assessment. Itasca, IL: F.E. Peacock Publishers.

Cronbach, L.J. (1990). Essentials of Psychological Testing, (5th ed.). New York: Harper Collins Publishers, Inc.

Cureton, E.E. (1968). International Encyclopedia of the Social Sciences. U.S.A.: Crowell, Collier and Macmillan.

Ebel, R.L., & Frisbie, D.A. (1991). Essentials of Educational Measurement (5th ed.). Englewood, NJ: Prentice Hall.

English, H.B. & English, A.C. (1962). A Comprehensive Dictionary of Psychological and Psychoanalytic Terms. U.S.A.: David McKay Company, Inc.

Funder, D.C. (1997). The Personality Puzzle. New York: W.W. Norton & Company.

Gatewood, R., & Field, H. (1994). Human Resource Selection. Fort Worth: Dryden Press.

Gregory, R.J. (1992). Psychological Testing: History, Principles, and Applications. Needham Heights, MA: Allyn and Bacon.

Groth-Marnat, G. (1990). Handbook of Psychological Assessment (2nd ed.). New York: John Wiley and Sons.

Hartigan, J., & Wigdor, A. (1989). Fairness in Employment Testing. Committee on the General Aptitude Test Battery, National Research Council. Washington, DC: National Academy Press.

Holland, P.W. & Wainer, H. (Eds.). (1993). Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

McCloskey, D., & Ziliak, S. (1996). "The standard error of regressions." Journal of Economic Literature, 37, 97-114.

Meir, R. (1994). The Chronic Crisis in Psychological Measurement and Assessment: A Historical Survey. New York: Academic Press.

Messick, S. (1989). "Validity." In R.E. Linn (Ed.), Educational Measurement (3rd ed.). New York: Macmillan.

National Research Council. (1998). High Stakes Testing for Tracking, Promotion, and Graduation. Washington, D.C.: National Academy Press.

Office of Technology Assessment. (1992). Testing in American Schools: Asking the Right Questions. OTA-SET-519. Washington, DC: U.S. Government Printing Office.

Reilly, R., & Warech, M. (1994). "The validity and fairness of alternative to cognitive tests." In H. Wing & B. Gifford (Eds.), Policy Issues in Employment Testing. Boston: Kluwer.

Rust, J., & Golombok, S. (1989). Modern Psychometrics: The Science of Psychological Assessment. New York: Routledge.

Shepard, L.A. (1993). "Evaluating Test Validity." Review of Research in Education, 19, 405-450.

Travers, B.E. (Ed.). (1996). The Gale Encyclopedia of Science. Detroit: Gale Research.

Wainer, H., & Braun, H. (Eds.). (1988). Test Validity. Hillsdale, NJ: Hove and London.

Selected References Regarding Appropriate Test Use In The Context of Standards-Based Reform

Coleman, A.L., "Excellence and Equity in Education: High Standards for High-Stakes Tests." 6 Virginia Journal of Social Policy & the Law 81(1998).

Federation for Children with Special Needs. (1998). All Kids Count.

Hansche, L.N. (1998). Handbook for the Development of Performance Standards: Meeting the Requirements of Title I. Washington, D.C.: Council of Chief State School Officers and the U.S. Department of Education.

Jaeger, R.M. & Tucker, C.G. (1998). Analyzing, Disaggregating, Reporting, and Interpreting Students' Achievement Test Results: A Guide to Practice for Title I and Beyond. Washington, D.C.: Council of Chief State School Officers.

Kopriva, R.J. & Saez, S.M. (1997). Guide for Scoring Open-Ended Mathematics Responses of LEP Students. Washington, D.C.: Council of Chief State School Officers.

Kopriva, R.J. (in press). Ensuring the Accurate Assessment of LEP Students: A Practical Guide for Test Development. Washington D.C.: Council of Chief State School Officers.

Linn, R.L. & Herman, J.L. (1997). A Policymaker's Guide to Standards-Led Assessment. Denver, CO: Education Commission of the States.

National Research Council. (1999). High Stakes Testing for Tracking, Promotion, and Graduation. Washington, D.C.: National Academy Press.

Tucker, M. & Codding, J. (1998). The State of Standards: Powerful Tool or Symbolic Gesture? Washington, D.C.: National Center on Education and the Economy.
U.S. Department of Education, Office of Educational Research and Improvement. (1997). Assessment of Student Performance.

U.S. Department of Education, Office of Elementary and Secondary Education. (in progress). Toolkit for the Assessment of Language Minority Students.

U.S. Department of Education, Standards, Assessments, and Accountability. (1997). http://www.ed.gov/pubs/legsregs.html#policy. http://www.ed.gov/offices/OESE/policy.html

Attachment - A

UNITED STATES DEPARTMENT OF EDUCATION
WASHINGTON, D.C. 20202

SEP 29 1997

Dear Colleague:

We are writing to you today to highlight the importance of including students with disabilities in all educational reform activities and, in particular, in statewide assessment systems. As you know, President Clinton has announced a bold, national education initiative which includes the goal of learning to challenging and clear standards of achievement for all students, including students with disabilities. In his 1997 State of the Union address, the President announced a ten-point call to action including rigorous, voluntary national tests in reading and math embodying national standards, teaching every student to read independently by the end of the third grade, and increased accountability in public education.

Assessment is an integral aspect of accountability. Assessment systems have varied purposes. Whatever the focus of the particular assessment system - program evaluation, school and staff accountability or measuring student progress - assessments provide valuable information which benefits individual students, either directly, such as in the measurement of individual progress against standards, or indirectly, such as in evaluating programs. Given the emphasis on assessment in recent educational reform efforts, including State and Federal legislation linking assessment and school accountability, it is of utmost importance that students with disabilities be included in the development and implementation of assessment activities. Too often, in the past, students with disabilities have not fully participated in State and district assessments only to be short-changed by the low expectations and less challenging curriculum that may result from exclusion.

Given the benefits that accrue as a result of assessment, exclusion from assessments based on disability generally would not only undermine the value of the assessment but also violate Section 504 of the Rehabilitation Act of 1973 (Section 504), which prohibits exclusion from participation of, denial of benefits to, or discrimination against, individuals with disabilities on the basis of their disability in Federally-assisted programs or activities. 29 U.S.C. 794. Similarly Title II of the Americans with Disabilities Act (ADA) of 1990 provides that no qualified individual with a disability shall, by reason of such disability, be excluded from participation in or be denied the benefits of the services, programs, or activities of a public entity, or be subjected to discrimination by such entity. 42 U.S.C. 12132.

alternate assessments, if it can be reported in a statistically sound manner and would not result in disclosure of performance results identifiable to individual children. The reports must be provided with the same frequency and in the same level of detail as the State's reports on the assessment of nondisabled children. For assessments conducted after July 1, 1998, data relating to the performance of children with disabilities in regular assessments is required to be disaggregated. For those assessments conducted prior to July 1, 1998, the data for children with disabilities participating in regular assessments, is only required by IDEA to be disaggregated if the State requires disaggregation. Section 612(a)(17)(B); 111 Stat. 67-68.

The Office of Special Education Programs within OSERS has a cooperative agreement with the National Center on Educational Outcomes (NCEO) at the University of Minnesota to study and provide information on including students with disabilities in statewide and other assessments. We have enclosed a brochure on the NCEO, which may be contacted for more information.

As we work together to reform our educational system, we must ensure that all children, including students with disabilities, are part of that reform. Including students with disabilities in the development and implementation of assessments is a vital step towards providing access to the general curriculum and learning to challenging standards.

Sincerely,

Judith Heumann

Judith E. Heumann
Assistant Secretary for
Special Education and
Rehabilitative Services

Norma V. Cantú

Norma V. Cantú
Assistant Secretary for
Civil Rights

Attachment - B

Office of the Attorney General
Washington, D.C. 20530

July 14, 1994

MEMORANDUM FOR HEADS OF DEPARTMENTS AND AGENCIES
THAT PROVIDE FEDERAL FINANCIAL ASSISTANCE

FROM: THE ATTORNEY GENERAL

SUBJECT: Use of the Disparate Impact Standard in Administrative Regulations Under Title VI of the Civil Rights Act of 1964

This month marks the 30th anniversary of the passage of Title VI of the Civil Rights Act of 1964 (42 U.S.C. §§2000d to 2000d-6), which prohibits discrimination on the basis of race, color, or national origin in programs and activities that receive Federal financial assistance. The anniversary of this landmark legislation is a fitting time to remind agencies that administrative regulations implementing Title VI apply not only to intentional discrimination but also to policies and practices that have a discriminatory effect. In Guardians Association v. Civil Service Commission, 463 U.S. 582 (1983), the Supreme Court held that while Title VI itself requires proof of discriminatory intent, agencies may validly adopt regulations implementing Title VI that also prohibit discriminatory effects. Nearly all agencies have adopted such regulations. In Alexander v. Choate, 469 U.S. 287 (1985) (construing Section 504 of the Rehabilitation Act of 1973), a unanimous Supreme Court restated the holding in Guardians that disparate impact violations could be addressed through regulations implementing Title VI.

This Administration will vigorously enforce Title VI. As part of this effort, and to make certain that Title VI is not violated, each of you should ensure that the disparate impact provisions in your regulations are fully utilized so that all persons may enjoy equally the benefits of federally financed programs.

Enforcement of the disparate impact provisions is an essential component of an effective civil rights compliance program. Individuals continue to be denied, on the basis of their race, color, or national origin, the full and equal opportunity to participate in or receive the benefits of programs assisted by Federal funds. Frequently discrimination results from policies and practices that are neutral on their face but have the <u>effect</u> of discriminating. Those policies and practices must be eliminated unless they are shown to be necessary to the program's operation and there is no less discriminatory alternative.

Under Executive Order 12250, the Department of Justice is responsible for ensuring that funding agencies meet their responsibilities under Title VI. This Department is committed to productive and effective enforcement of the civil rights laws by each agency that extends Federal financial assistance. Facially neutral policies and practices that act as arbitrary and unnecessary barriers to equal opportunity must end. This was the goal of Title VI when it became law and it remains one of the highest priorities of this Administration.