The Use of Tests When Making High-Stakes Decisions for Students:

A Resource Guide for Educators and Policymakers

U.S. Department of Education
Office for Civil Rights

Richard W. Riley
Secretary U.S. Department of Education

Norma V. Cantu
Assistant Secretary
Office for Civil Rights
U.S. Department of Education

September 2000

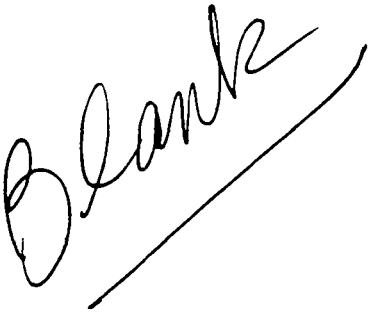
This resource guide has been developed by the U.S. Department of Education, Office for Civil Rights (OCR), in an effort to assemble the best information regarding test measurement standards, legal principles, and resources to help educators and policymakers ensure that uses of tests with high-stakes consequences for students are educationally sound and legally appropriate. The resource guide is intended to reflect existing test measurement and legal principles. The resource guide is not intended to and does not add to, or subtract from, any otherwise applicable federal requirements. This publication supercedes any earlier drafts, notes, or other preparatory versions of this document.

 The full text of the resource guide is also available at OCR's web page, www.ed.gov/offices/OCR. Individuals with disabilities may obtain this document in an alternate format (e.g., Braille, large print, audiotape, or computer diskette) on request. For -5more information, please contact OCR by telephone at 800-421-3481 or by email at OCR@ed.gov. Individuals who use a telecommunications device for the deaf (TDD) may -call OCR's TDD number at 877-521-2172.

Acknowledgements

This resource guide was developed by the U.S. Department of Education, in consultation with numerous stakeholders. The time and commitment of all those who provided comments and input is gratefully acknowledged. In particular, we want to recognize the primary drafters of this document; David Berkowitz, Barbara Walkowitz, Rebecca Fitch, and Rebecca Kopriva.

We also want to thank others from the U.S. Department of Education's Office for Civil Rights (OCR) and Office of the General Counsel who assisted in the dvelopment of the guide, including Scott Plamer, Arthur Coleman, Jeanette Lim, Susan Bowers, Cathy Lweis, Connie Butler, Doreen Dennis, Lilian Dorka, Marsha Douglas, Lisa Dyson, Ann Hoogstraten, Jerry Kravitz, Jan Pottker, Rebekah Tosado, Steven Winnick, Susan Craig, Karl Lahring, Lisa Battlia Anthony, Adina Kole, and Suzanne Sheridan. Additionally, we are grateful to the efforts of individuals, especially within OCR, who were repsonsible for developing earlier drafts of the document. Finally, we want to recognize the efforts of other persons within the U.S. Departmnet of Education, the U.S. Department of Justice's Civil Rights Division, and the National Academy of Science's Board of Testing and Assessment, who reviewed drafts of this document and provided valuable guidance.


Blank

The Use of Tests When Making High-Stakes Decisions for Students:

A Resource Guide for Educators and Policymakers

U.S. Department of Education
Office for Civil Rights

Blank

UNITED STATES DEPARTMENT OF EDUCATION

OFFICE FOR CIVIL RIGHTS

THE ASSISTANT SECRETARY

Dear Colleague:

Adherence to good test use practices in education is a shared goal of government officials, policymakers, educators, parents, and students. In an era of school reforms that place increasing emphasis on measures of accountability, such as the use of tests as part of decision-making that has high-stakes consequences for students,* the need to provide practical information about good testing practices is well documented. In January 1999, the National Research Council observed that we, in the education community, should work to better disseminate information related to good testing practices with a focus on the standards of testing professionals and the relevant legal principles that, together, "reflect many common concerns."

Sound educational policies and federal nondiscrimination laws can work together to promote educational excellence for all students and ensure that educational practices do not — intentionally or otherwise — unfairly deny educational opportunities to students based upon their race, national origin, sex, or disability. In short, federal civil rights law affirms good test use practices. Thus, an understanding of the measurement principles related to the use of tests for high-stakes purposes is an essential foundation to better understanding the federal legal standards that are significantly informed by those measurement principles.

In order to further the goal of accurate and fair judgments in high-stakes decision-making that involves the use of tests, we are pleased to provide you with this copy of *The Use of Tests When Making High-Stakes Decisions for Students: A Resource Guide for Educators and Policymakers.* This guide provides important information about the professional standards relating to the use of tests for high-stakes purposes, the relevant federal laws that apply to such practices, and references that can help shape educationally sound and legally appropriate practices.

There are few simple or definitive answers to questions about the use of tests for high-stakes purposes. Tests are a means to an end and, as such, can be understood only in

* As explained throughout the guide, the primary focus is the use of standardized tests or assessments (referred to in the guide as tests) used to make decisions with important consequences for individual students. Examples of high-stakes decisions include: student placement in gifted and talented programs or in programs serving students with limited English proficiency; determinations of disability and eligibility to receive special education services; student promotion from one grade level to another; graduation from high school and diploma awards; and admission decisions and scholarship awards. The guide does not address teacher-created tests that are used for individual classroom purposes.

the context in which they are used. The education context — in which the relationship (and attendant obligations) of the educator to the student is frequently more complex than that between employer and employee — shows time and again that any decision regarding the legality of a use of a test for high-stakes purposes under federal nondiscrimination laws cannot be made without regard to the educational interests and judgments upon which the test use is premised.

Background

Throughout the 1990s, national, state, and local education leaders focused on raising education standards and establishing strategies to promote accountability in education. In fact, the promotion of challenging learning standards for all students — coupled with assessment systems that monitor progress and hold schools accountable — has been the centerpiece of the education policy agenda of the federal government as well as of many states.

At the same time, the use of tests for making high-stakes decisions for students is on the rise. For example, the number of states using tests as a condition for high school graduation is increasing, with a majority of states projected to use tests as conditions for graduation by 2003 and several states now using tests as conditions for grade promotion.

Recently, more and more educators and policymakers have requested advice and technical assistance from the U.S. Department of Education regarding test use in the context of standard-based reforms. The Department's Office for Civil Rights (OCR) is also addressing testing issues in a more extensive array of complaints of discrimination being filed with our office, most of them in a K-12 setting with implications for high-standards learning. OCR has responsibility for enforcing Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, Section 504 of the Rehabilitation Act of 1973, and Title II of the Americans with Disabilities Act of 1990. These statutes prohibit discrimination on the basis of race, color, national origin, sex, and disability by educational institutions that receive federal funds.

In a similar vein, institutions in the post-secondary community in recent years have engaged in a thoughtful dialogue and analysis regarding merit in admissions and the appropriate use of tests as part of the process for making high-stakes admissions decisions. In some states, the use of tests in connection with admissions decisions has been an important element in public post-secondary education reform.

These trends highlight the salience of two recent conclusions of the National Research Council (NRC) Board on Testing and Assessment. The NRC observed that many policymakers and educators are unaware of the test measurement standards that should inform testing policies and practices. These standards include the *Standards for Educational and Psychological Testing* (*Joint Standards*), prepared by a joint committee of the American Psychological Association (APA), the American Educational Research

the context in which they are used. The education context — in which the relationship (and attendant obligations) of the educator to the student is frequently more complex than that between employer and employee — shows time and again that any decision regarding the legality of a use of a test for high-stakes purposes under federal nondiscrimination laws cannot be made without regard to the educational interests and judgments upon which the test use is premised.

Background

Throughout the 1990s, national, state, and local education leaders focused on raising education standards and establishing strategies to promote accountability in education. In fact, the promotion of challenging learning standards for all students — coupled with assessment systems that monitor progress and hold schools accountable — has been the centerpiece of the education policy agenda of the federal government as well as many states.

At the same time, the use of tests for making high-stakes decisions for students is on the rise. For example, the number of states using tests as a condition for high school graduation is increasing, with a majority of states projected to use tests as conditions for graduation by 2003 and several states now using tests as conditions for grade promotion.

Recently, more and more educators and policymakers have requested advice and technical assistance from the U.S. Department of Education regarding test use in the context of standard-based reforms. The Department's Office for Civil Rights (OCR) is also addressing testing issues in a more extensive array of complaints of discrimination being filed with our office, most of them in a K-12 setting with implications for high-standards learning. OCR has responsibility for enforcing Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, Section 504 of the Rehabilitation Act of 1973, and Title II of the Americans with Disabilities Act of 1990. These statutes prohibit discrimination on the basis of race, color, national origin, sex, and disability by educational institutions that receive federal funds.

In a similar vein, institutions in the post-secondary community in recent years have engaged in a thoughtful dialogue and analysis regarding merit in admissions and the appropriate use of tests as part of the process for making high-stakes admissions decisions. In some states, the use of tests in connection with admissions decisions has been an important element in public post-secondary education reform.

These trends highlight the salience of two recent conclusions of the National Research Council (NRC) Board on Testing and Assessment. The NRC observed that many policymakers and educators are unaware of the test measurement standards that should inform testing policies and practices. These standards include the *Standards for Educational and Psychological Testing* (*Joint Standards*), prepared by a joint committee of the American Psychological Association (APA), the American Educational Research

Association (AERA), and the National Council on Measurement in Education (NCME). The NRC also concluded that it "is essential that educators and policymakers alike be aware of both the letter of the laws and their implications for test takers and test users" [National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation,* (Heubert and Hauser, eds., 1999)].

The Resource Guide

Toward this end, OCR has prepared this guide in an effort to assemble the best information regarding test measurement standards, legal principles, and resources to help educators and policymakers frame strategies and programs that promote learning to high standards in ways consistent with federal non-discrimination laws. Our goal is to inform decisions related to the use of tests as part of decision-making that has high-stakes consequences for students, such as when they move from grade to grade or graduate from high school. Just as we know that good test use practices can advance high standards for learning and equal opportunity, we know that educationally inappropriate uses of tests do not. If we want this generation of test-taking students and their teachers and schools to meet high standards, then we should insist that the tests they take meet high standards. When tests are used in ways that profoundly shape the lives of students, they must also be used in ways that accurately reflect educational standards and that do not deny opportunities or benefits to students based on their race, national origin, sex, or disability.

The guide is organized to provide practical guidance related to the test measurement principles and applicable federal laws that guide the use of tests as part of decision-making that has high-stakes consequences for students. The Introduction to the guide provides a broad, conceptual overview of relevant principles so that those who are not familiar with test measurement principles or applicable federal laws can better understand the kinds of issues that relate to the use of tests in many contexts. Chapter One of the guide provides a detailed discussion of the test measurement principles that provide a foundation for making well-informed decisions related to the use of tests for high-stakes purposes. The Joint Standards that have been approved by the APA, AERA, and NCME are discussed in detail in this chapter. Adherence to relevant professional standards can help reduce the risk of legal liability when schools are using assessments for high-stakes purposes. Chapter Two provides an overview of the existing legal principles that have guided federal courts and OCR when analyzing claims of race, national origin, sex, and disability discrimination related to the use of tests for high-stakes purposes. These principles, as applied by the courts and OCR, underscore the importance of adhering to educationally sound testing practices. The Appendix includes a Glossary of Legal Terms, a Glossary of Test Measurement Terms, a Compendium of Federal Statutes and Regulations, and a Resources and References section.

Central Principles

There are several central principles reflected in the text of this guide.

First, the goals of promoting high educational standards and ensuring nondiscrimination are complementary objectives. The ultimate question regarding the use of tests for high-stakes purposes, as a matter of federal nondiscrimination law and sound educational policy, centers on educational sufficiency: Is the test appropriate for the purposes used? That is, are the inferences derived from test scores, and the high-stakes decisions based on those inferences, valid, reliable, and fair for all students? These inquiries are not an effort to dumb down academic standards or alter core education objectives integral to academic admissions or other educational decisions. Rather, they focus the educator and policymaker on ensuring that uses of tests with high-stakes consequences for students are educationally sound and legally appropriate. In applying civil rights laws to education cases, federal courts recognize the importance of providing appropriate deference to the educational judgments of educators and policymakers.

Second, when tests, including large-scale standardized tests, are used in valid, reliable, and educationally appropriate ways, their use is not inconsistent with federal nondiscrimination laws. Importantly, tests can help indicate inequalities in the kinds of educational opportunities students are receiving, and in turn, may stimulate efforts to ensure that all students have equal opportunity to achieve high standards. When tests accurately indicate performance gaps, our focus should be on the quality of educational opportunities afforded to under-performing students. The key question in the context of standards-based reforms and the use of tests as measures of student accountability is: Have all students been provided quality instruction, sufficient resources, and the kind of learning environment that would foster success?

Third, a test score disparity among groups of students does not alone constitute discrimination under federal law. The guarantee under federal law is for equal opportunity, not equal results. Test results indicating that groups of students perform differently should be a cause for further inquiry and examination, with a focus upon the relevant educational programs and testing practices at issue. Differences in test scores may result from a range of factors, some of which a school may be able to influence, and others over which it has little control. Federal law recognizes this point, as it must. The legal non-discrimination standard regarding neutral practices (referred to by the courts as the "disparate impact" standard) provides that if the education decisions based upon test scores reflect significant disparities based on race, national origin, sex, or disability in the kinds of educational benefits afforded to students, then questions about the education practices at issue (including testing practices) should be thoroughly examined to ensure that they are in fact non-discriminatory and educationally sound.

In short, the goal of the federal legal standards is to help promote accurate and fair decisions that have real consequences for students, not to water down academic standards or deter educators from establishing and applying sensible and rigorous standards. In fact, properly understood, the legal standards are an aid to meaningful education reform — by helping to ensure that instruction and assessments are aligned and structured to promote the high-level skills and knowledge that rigorous standards seek for all children.

Finally, while this guide focuses on the use of tests, similar principles apply to the overall decision-making process used to make high-stakes decisions for students. In fact, the *Joint Standards*, the NRC, APA, AERA, NCME, and others caution against making high-stakes decisions based on a single test score. "Other relevant information should be taken into account if it will enhance the overall validity of the decision." [American Educational Research Association, American Psychological Association, National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, p. 146 (1999)].

Conclusion

Recognizing the responsibility that educators and policymakers must shoulder in making the promise of high-standards learning a reality, U.S. Secretary of Education Richard Riley in his commemoration of the 45th anniversary of the *Brown v. Board of Education* decision said, "A quality education must be considered a key civil right for the twenty first century." This is the driving force behind OCR's continuing effort to provide assistance to policymakers and educators as we continue to enforce federal laws that prohibit discrimination against students. Rather than creating false and polarizing "win-lose" choices on this all-important set of issues, we need to, as Secretary Riley admonishes, "search for common ground" — ground, that is, in this case, expansive.

We have worked with literally dozens of groups and individuals, including educators, parents, teachers, business leaders, policymakers, test publishers, and others, to solicit input and advice regarding the scope, framing, and kinds of resources to include in this guide, and we are grateful for their time and assistance. The first draft of the testing guide was released in April 1999 and was the subject of substantial comments leading to extensive revisions. The second draft was released in December 1999 and once again received substantial comments. This draft also was independently reviewed by the NRC's Board on Testing and Assessment, which held a hearing earlier this year to discuss the draft guide and issued a letter report in June 2000 commenting on the draft. We are grateful for the NRC's tireless efforts. The third draft was released for public comment in July 2000, this time with notice of availability in the *Federal Register.*

OCR has made numerous changes throughout the guide in response to comments seeking to clarify, make more accurate, or expand key sections. It is important to keep in mind that the guide is not designed to answer all questions related to the use of tests when making high-stakes decisions for students. However, working together with our education partners, we believe that we are providing a useful resource that will serve the education community as it addresses the very complex and important questions that stem from the institution of high standards and accountability systems designed to promote the best schools in the world.

Very truly yours,

Norma V. Cantú

Table of Contents

INTRODUCTION: An Overview of the Resource Guide

I. Introduction

When decisions are made affecting students' educational opportunities and benefits, it is important that they be made accurately and fairly. When tests are used in making educational decisions for individual students, it is important that they accurately measure students' abilities, knowledge, skills, or needs, and that they do so in ways that do not discriminate in violation of federal law on the basis of students' race, national origin, sex, or disability. The U.S. Department of Education's Office for Civil Rights (OCR)[1] has developed this resource guide in order to provide educators and policymakers with a useful, practical tool to assist in their development and implementation of policies that involve the use of tests as part of decision-making that has high-stakes consequences for students.

> When tests are used in ways that meet relevant psychometric, legal, and educational standards, students' scores provide important information that, combined with information from other sources, can lead to decisions that promote student learning and equality of opportunity .... When test use is inappropriate, especially in making high-stakes decisions about individuals, it can undermine the quality of education and equality of opportunity. .... This lends special urgency to the requirement that test use with high-stakes consequences for individual students be appropriate and fair.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 4 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Chapter One of this guide provides information about professionally recognized test measurement principles. Chapter Two provides the legal frameworks that have guided federal courts and OCR when addressing the use of tests that have high-stakes consequences for students. This document does not establish any new legal or test measurement principles. The test measurement principles described in Chapter One are not legal principles. However, the use of tests in educationally appropriate ways

[1] OCR enforces laws that prohibit discrimination on the basis of race, national origin, sex, disability, and age by educational institutions that receive federal funds. The laws enforced by OCR are: 1) Title VI of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000d *et seq.* (2000) (Title VI), which prohibits discrimination on the basis of race, color, or national origin; 2) Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq.* (1999) (Title IX), which prohibits discrimination on the basis of sex; 3) Section 504 of the Rehabilitation Act of 1973, 29 U.S.C. §§ 794 *et seq.* (1999) (Section 504), which prohibits discrimination on the basis of disability; 4) the Age Discrimination Act of 1975, 42 U.S.C. §§ 6101 *et seq.* (1995 & Supp. 1999) (as amended), which prohibits age discrimination; and 5) Title II of the Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12134 *et seq.* (1995 & Supp. 1999) (Title II), which prohibits discrimination on the basis of disability by public entities, whether or not they receive federal financial assistance.

— consistent with the principles described in chapter one — can help minimize the risk of noncompliance with the federal nondiscrimination laws discussed in chapter two.

The guide also includes a collection of resources related to the test measurement and nondiscrimination principles discussed in the guide — all in an effort to help policymakers and educators ensure that decisions that have high-stakes consequences for students are made accurately and fairly.

Recently educational stakeholders at all levels have approached OCR requesting advice and technical assistance in a variety of test-use contexts, particularly as states and districts use tests as part of their standards-based reforms. Also, OCR is increasingly addressing testing issues in a broader and more extensive array of complaints of discrimination that have been filed with OCR. These developments confirm the need to provide a useful resource that captures legal and test measurement principles and resources to assist educators and policymakers.

> High-stakes decisions in this guide refer to decisions with important consequences for students, such as placement in special programs, promotion, graduation, and admissions decisions.

As used in this resource guide, "high-stakes decisions" refer to decisions with important consequences for individual students. Education entities, including state agencies, local education agencies, and individual education institutions, make a variety of decisions affecting individual students during the course of their academic careers, beginning in elementary school and extending through the post-secondary school years. Examples of high-stakes decisions affecting students include: student placement in gifted and talented programs or in programs serving students with limited-English proficiency; determinations of disability and eligibility to receive special education services; student promotion from one grade level to another; graduation from high school and diploma awards; and admissions decisions and scholarship awards.[2]

This guide is intended to apply to standardized tests that are used as part of decision-making that has high-stakes consequences for individual students and that are addressed in the *Standards for Educational and Psychological Testing* (*Joint Standards 1999*). The *Joint Standards* is viewed as the primary technical authority on educational test measurement issues. They have been prepared by a joint committee of the American Educational Research Association, the American Psychological

[2] The purpose of this guide is to address tests that are used in making high-stakes decisions for individual students. In addition to using tests for high-stakes purposes for individual students, states and school districts are also using tests to hold schools and districts accountable for student performance. Although the use of tests for this purpose is not the focus of the guide, we have provided some useful background information about relevant principles and federal statutory requirements.

Association, and the National Council on Measurement in Education —the three leading organizations in the area of educational test measurement. The *Joint Standards* were developed and revised by these three organizations through a process that involved the participation of hundreds of testing professionals and thousands of pages of written comments from both professionals and the public. The current edition of the *Joint Standards* reflects the experience gained from many years of wide use of previous versions of the *Joint Standards* in the testing community.

The *Joint Standards*, which is discussed in more detail below, apply to standardized measures generally recognized as tests, and also may be applied usefully to a broad range of systemwide standardized assessment procedures.[3] For the sake of simplicity, this guide will refer to tests, regardless of the type of label that might otherwise be applied to them. The guide does not address teacher-created tests that are used for individual classroom purposes.

States and school districts are also using assessment systems for the purpose of promoting school and district accountability.[4] For example, under Title I of the Elementary and Secondary Education Act, states are required to develop content standards, performance standards, and assessment systems that measure the progress that schools and districts are making in educating students to the standards established by the state. The Title I statute explicitly requires that assessments be valid and reliable for their intended purpose and be consistent with relevant, nationally recognized technical and professional standards.[5] When educators and policymakers consider using the same test for school or district accountability purposes and for individual student high-stakes purposes, they need to ensure that the test score inferences are valid and reliable for each particular use for which the test is being considered.[6]

---

[3] The *Joint Standards* note that their applicability to an evaluation device or method is not altered by the label used (e.g., test, assessment scale, inventory). A more complete discussion about the instruments covered by the *Joint Standards* can be found in the introduction section of that document. *See* American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, pp. 3-4 (1999) (hereinafter *Joint Standards*).

[4] The Goals 2000: Educate America Act supports state efforts to develop clear and rigorous standards for what every child should know and be able to do, and supports comprehensive state and districtwide planning and implementation of school improvement efforts focused on improving student achievement to those standards. *See* 20 U.S.C. §§ 5801 *et seq.* (1994). Largely through state awards that are distributed on a competitive basis to local school districts, Goals 2000 promotes education reform in every state and thousands of districts and schools.

[5] *See* 20 U.S.C. § 6311(b)(3)(C).

[6] For example, if there is a major gap between the skills and knowledge being assessed and what is being taught, this does not undermine the validity of an assessment used for purposes of program evaluation and accountability – indeed the purpose of the assessment is to detect such gaps. In contrast, such a gap would raise serious concerns about the appropriateness of the use of the assessments for promotion and graduation decision-making where students are being held accountable for what they have purportedly been taught.

Although this guide focuses on the use of tests, policymakers and the education community need to ensure that the operation of the entire high-stakes decision-making process, and each element of that process, do not result in the discriminatory denial of educational opportunities or benefits to students.[7]

Indeed, the *Joint Standards* state that, in educational settings, a high-stakes decision "should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision."[8] As explained in the *Joint Standards*, "When interpreting and using scores about individuals or groups of students, considerations of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance."[9] The *Joint Standards* also note that "as the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly. The validity of individual interpretations can be enhanced by taking into account other relevant information about individual students before making important decisions. It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions."[10]

Used appropriately, tests can provide important information about a student's knowledge and help improve educational opportunity and acheivement. However, "no single test score can be a definitive measure of a student's knowledge."[11] Thus, educational institutions are in a stronger position when they use measures that enhance the validity of high-stakes decisions for students.

---

[7] *See* regulations implementing Title VI of the Civil Rights Act of 1964, 34 C.F.R. §§ 100.3(a), 100.3(b)(1)(i) and (vi), 100.3(b)(2); regulations implementing Section 504 of the Rehabilitation Act of 1973, 34 C.F.R. §§ 104.4(a), 104.4(b)(1)(i) and (iv), 104.4(b)(4); regulations implementing Title IX of the Education Amendments of 1972, 34 C.F.R. §§ 106.31(a), 106.31(b).

[8] Standard 13.7 states, "In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision." *Joint Standards*, *supra* note 3, at p. 146.

[9] *Joint Standards*, *supra* note 3, at p. 141.

[10] *Joint Standards*, *supra* note 3, at p. 141. Many test developers also caution against using their tests as the sole criterion in making a decision with high-stakes consequences for students. Discussion of this issue can be found in interpretive guides from test publishers, such as Riverside Publishing, Harcourt Brace, CTB McGraw Hill, and the Educational Testing Service, regarding the use of tests.

[11] National Research Council, *High Stakes: Testing for Tracking, Promotion, and* Graduation, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999) (hereinafter *High Stakes*).

Applicable standards for technical quality set forth in the *Joint Standards* are important principles to consider when other criteria affect high-stakes decisions. Educators should carefully monitor inputs into the high-stakes decision-making process and outcomes over time so that potential discrimination arising from the use of any of the criteria can be identified and eliminated.

> Standardized tests ... offer important benefits that should not be overlooked. ... Both the SAT [I] and ACT cover relatively broad domains that most observers would likely agree are relevant to the ability to do college work. Neither, however, measures the full range of abilities that are needed to succeed in college; important attributes not measured include, for example, persistence, intellectual curiosity, and writing ability. Moreover, these tests are neither complete nor precise measures of 'merit'—even academic merit.
>
> National Research Council, *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions*, pp. 21-22 (Alexandra Beatty, M.R.C. Greenwood & Robert L. Linn eds., 1999).

Finally, this guide focuses primarily on tests used in making high-stakes decisions at the elementary and secondary education level. However, it is important to recognize that the general principles of sound educational measurement apply equally to tests used at the post-secondary education level, including admissions and other types of tests.[12] For example, post-secondary admissions policies and practices should be derived from and clearly linked to an institution's overarching educational goals, and the use of tests in the admissions process should serve those institutional goals.[13]

[12] For additional information regarding testing at the post-secondary level, *see Joint Standards*, *supra* note 3, at pp. 142-143; National Research Council, *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions* (Alexandra Beatty, M.R.C. Greenwood & Robert L. Linn eds., 1999) (hereinafter *Myths and Tradeoffs*); Samuel Messick, *Validity*, in *Educational Measurement*, pp. 13-103 (Robert L. Linn ed., 3rd ed. 1989); *Ability Testing: Uses, Consequences, and Controversies*, Chapter 5 (Alexandra K. Wigdor & Wendell R. Garner eds., 1982).

[13] *See Myths and Tradeoffs*, *supra* note 12, at p. 1.

II. Foundations of the Resource Guide

A. Professional Standards of Sound Testing Practices

> The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education ... The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions.
>
> American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards of Educational and Psychological Testing*, Introduction, p. 1 (1999).

Chapter One summarizes the leading professionally recognized standards of sound testing practices within the educational measurement field. They include those described in the *Joint Standards*, which represent the primary statement of professional consensus regarding educational testing. Other leading professionally recognized standards of sound testing practices within the educational measurement field include the *Code of Fair Testing Practices in Education* (1988), and the *Code of Professional Responsibilities in Educational Measurement* (1995). The guide also cites recent reports from the National Research Council's Board on Testing and Assessment, including *High Stakes: Testing for Tracking, Promotion and Graduation* (*High Stakes,1999*); *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions* (*Myths and Tradeoffs, 1999*);*Testing, Teaching, and Learning: A Guide for States and School Districts* (*Testing, Teaching, and Learning, 1999*); *Improving Schooling for Language-Minority Children: A Research Agenda* (*Improving Schooling for Language-Minority Children, 1997*); and *Educating One & All: Students with Disabilities and Standards-Based Reform* (*Educating One & All, 1997*).[14] These reports help explain or elaborate on principles that are stated in the *Joint Standards*.

Designed to provide criteria for the evaluation of tests, testing practices, and the effects of test use, the *Joint Standards* recommend that all professional test developers, sponsors, publishers, and users make efforts to observe the *Joint Standards* and encourage others to do so.[15] The *Joint Standards* include chapters on the test development process (with a focus primarily on the responsibilities of test developers), the specific uses and applications of tests (with a focus primarily on the responsibilities of test users), and the rights and responsibilities of test takers. Because the *Joint*

[14] The National Academy of Sciences, which is an independent, private, nonprofit entity, established the Board on Testing and Assessment in 1993 to help policymakers evaluate the use of tests, alternative assessments, and other indicators commonly used as tools of public policy. The Board provides guidance for judging the quality of testing or assessment technologies and the intended and unintended consequences of particular uses of these technologies. The Board concentrates on topics and conducts activities that serve the general public interest.

[15] *See, e.g., Joint Standards, supra* note 3, at Introduction, p. 2.

Standards are the most widely accepted professional standards that are relied upon in developing testing instruments, this guide includes a discussion of specific standards that are contained within the *Joint Standards*, where relevant. Numbered standards that are referenced throughout this guide refer to specific standards that are contained within the *Joint Standards*.

To ensure that information presented in this guide is readable and accessible to educators and policymakers, we have paraphrased language from relevant standards. Our goal in paraphrasing is to be concise and accurate. Where we have paraphrased in the text, we have also provided the full text of the relevant standards in the footnotes. Because the *Joint Standards* provide additional relevant discussion, we always encourage readers also to review the full document.

Professional test measurement standards provide important information that is relevant to making determinations about appropriate test use. The *Joint Standards* provide a frame of reference to assist in the evaluation of tests, testing practices, and the effects of test use. The *Joint Standards* caution that the acceptability of a test or test application does not rest on the literal satisfaction of every standard in the *Joint Standards* and cannot be determined by using a checklist.[16] The exercise of professional judgment is a critical element in the interpretation and application of the standards,[17] and the interpretation of individual standards should be considered in the overall context of the use of the test in question. Finally, while the Joint Standards and federal nondiscrimination laws are closely aligned and mutually reinforcing, the failure to meet a particular professional test measurement standard does not necessarily constitute a lack of compliance with federal civil rights laws. Conversely, compliance with professional test measurement standards does not necessarily constitute compliance with all applicable federal civil rights laws.

B. Legal Principles

Chapter Two of the guide discusses the federal constitutional, statutory, and regulatory nondiscrimination principles that apply to the use of tests for high-stakes purposes. This guide is intended to reflect existing legal principles and does not establish new federal legal requirements. The primary legal focus of the resource guide is an explanation of principles that are clearly embedded in four nondiscrimination laws that have been enacted by Congress: Title VI of the Civil Rights Act of 1964 (Title VI), Title IX of the Education Amendments of 1972 (Title IX), Section 504 of the Rehabilitation Act of 1973 (Section 504), and Title II of the Americans with Disabilities Act of 1990

[16] *See Joint Standards, supra* note 3, at Introduction, p. 4.

[17] *See Joint Standards, supra* note 3, at Introduction, p. 4.

(Title II).[18] Within the U.S. Department of Education, the Office for Civil Rights has responsibility for enforcing the requirements of these four statutes and their implementing regulations. The due process and equal protection requirements of the Fifth and Fourteenth Amendments to the U.S. Constitution have also been applied by courts to issues regarding the use of tests in making high-stakes educational decisions. Although the Office for Civil Rights does not enforce federal constitutional provisions, a brief overview of these fundamental constitutional principles has been included to provide educators with a more complete picture of relevant legal standards.

III. Basic Principles

The brief overview of the test measurement and legal principles that follows establishes the framework for more detailed discussions of test quality in Chapter One and federal legal standards in Chapter Two.

A. Test Use Principles

1. Educational Objectives and Context

> Decisions about tracking, promotion, and graduation differ from one another in important ways. They differ most importantly in the role that mastery of past material and readiness for new material play.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 4 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Tests that are used in educationally appropriate ways and that are valid for the purposes used can serve as important instruments to help educators do their job. Before any state, school district, or educational institution administers a test, the objectives for using the test should be clear: What are the intended goals for and uses of the test in question? As an educational matter, the answer to this question will guide all other relevant inquiries about whether the test use is educationally appropriate. The context in which a test is to be administered, the population of test takers, the intended purpose for which the test will be used, and the consequences of

[18] Title VI prohibits discrimination on the basis of race, color and national origin in the programs and activities of recipients that receive federal financial assistance. The U.S. Department of Education's regulation implementing Title VI is found at 34 C.F.R. Part 100. Title IX prohibits discrimination on the basis of sex in educational programs and activities of recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title IX is found at 34 C.F.R. Part 106. Section 504 prohibits discrimination on the basis of disability in the programs and activities of recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Section 504 is found at 34 C.F.R. Part 104. Title II prohibits discrimination on the basis of disability by public entities, regardless of whether they receive federal funding. The U.S. Department of Justice's regulation implementing Title II is found at 28 C.F.R. Part 35.

such use are important considerations in determining whether the test would be appropriate for a specific type of decision, including placement, promotion or graduation decisions.

a. Placement Decisions

Placement decisions are by their very nature used to make a decision about the future. Tests used in placement decisions generally determine what kinds of programs, services, or interventions will be most appropriate for particular students. Decisions concerning the appropriate educational program for a student with a disability, placement in gifted and talented programs, and access to language services are examples of placement decisions. The *Joint Standards* state that there should be adequate evidence documenting the relationship among test scores, appropriate instructional programs, and desired student outcomes.[19] When evidence about the relationship is limited, the test results should usually be considered in light of other relevant student information.[20]

> [At the elementary and secondary education level,] appropriate test use for ... all students requires that their scores not lead to decisions or placements that are educationally detrimental.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, (Jay P. Heubert & Robert M. Hauser eds., 1999).

[19] Standard 13.9 states, "When test scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical information is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student." *Joint Standards*, *supra* note 3, at p. 147.

[20] *See Standard 13.9 (n.19) in Joint Standards*, *supra* note 3, at p. 147.

b. Promotion Decisions

Student promotion decisions are generally viewed as decisions incorporating a determination about whether a student has mastered the subject matter or content of instruction provided to the student and a determination regarding whether the student will be able to master the content at the next grade level (a placement decision).[21] When a test given for promotion purposes is being used to certify mastery, the use of the test should adhere to professional standards for certifying knowledge and skills for all students.[22] As indicated in the *Joint Standards*, it is important that there "be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn." [23] Educational institutions should have information indicating an alignment among the curriculum, instruction, and material covered on such a test used

> Neither a test score or any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional support services. In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.
>
> National Research Council, *High-Stakes: Testing for Tracking, Promotion, and Graduation*, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999).

[21] *See High Stakes, supra* note 11, at p. 123.

[22] *See Standard 13.5 and 13.6 in Joint Standards, supra* note 3, at p. 146; *High Stakes, supra* note 11, at p. 123.

Standard 13.5 states, "When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn." *Joint Standards, supra* note 3, at p. 146.

Standard 13.6 states, "Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences." *Joint Standards, supra* note 3, at p. 146.

[23] *See Standard 13.5 (n.22) in Joint Standards, supra* note 3, at p. 146; *High Stakes, supra* note 11 at pp. 124-125.

for high-stakes purposes. To the extent that a test for promotion purposes is being used as a placement device, it should also adhere, as appropriate, to professional standards regarding tests used for placement purposes.[24]

c. Graduation Decisions

Graduation decisions are generally certification decisions: The diploma certifies that the student has reached an acceptable level of mastery of knowledge and skills.[25] When large-scale standardized tests are used in making graduation decisions, as indicated in the *Joint Standards*, there should "be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn."[26] Therefore, all students should be provided a meaningful opportunity to acquire the knowledge and skills that are being tested, and information should indicate an alignment among the curriculum, instruction, and material covered on the test used as a condition for graduation.

2. Overarching Principles

In the elementary and secondary education context, regardless of whether tests are being used to make placement, promotion, or graduation decisions, the National Academy of Sciences' Board on Testing and Assessment has identified three principal criteria, which are based on established professional standards, that can help inform and guide conclusions regarding the appropriateness of a particular test use.[27]

> Is it ever appropriate to test [elementary or secondary] students on material they have not been taught? Yes, if the test is used to find out whether the schools are doing their job. But if that same test is used to hold students "accountable" for the failure of the schools, most testing professionals would find such use inappropriate. It is not the test itself that is the culprit in the latter case; results from a test that is valid for one purpose can be used improperly for other purposes.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, p. 21 (Jay P. Heubert & Robert M. Hauser eds., 1999).

[24] *See Standard 13.2 and 13.9 (n.19) in Joint Standards, supra* note 3, at pp. 145, 147; *High Stakes, supra* note 11, at p. 123.

Standard 13.2 states, "In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose." *Joint Standards, supra* note 3, at p. 145.

[25] *See High Stakes, supra* note 11, at p. 166.

[26] *See Standard 13.5 (n.22) in Joint Standards, supra* note 3, at p. 146.

[27] *See High Stakes, supra* note 11, at p. 23 (citing National Research Council, *Placing Children in Special Education: A Strategy for Equity* (1982)).

(1) *Measurement validity: Is a test valid for a particular purpose, and does it accurately measure the test taker's knowledge in the content area being tested?*

State and local education agencies and educational institutions should ensure that a test actually measures what it is intended to measure for all students. The inferences derived from the test scores for a given use — for a specific purpose, in a specific type of situation, and with specific types of students — are validated, rather than the test itself. It is important for educators who use the test to obtain adequate evidence of test quality (including validity and reliability evidence), evaluate the evidence, and ensure that the test is used appropriately in a manner that is consistent with information provided by the developers or through supplemental validation studies.

(2) *Attribution of cause: Does a student's performance on a test reflect knowledge and skills based on appropriate instruction, or is it attributable to poor instruction or to such factors as language barriers unrelated to the skills being tested?*

In some contexts, whether a particular test use is appropriate depends on whether test scores are an accurate reflection of a student's knowledge or skills or whether they are influenced by extraneous factors unrelated to the specific skills being tested. For example, when tests are used in making student promotion or graduation decisions, state and local education agencies should ensure that all students have an equal opportunity to acquire the knowledge and skills that are being tested.[28] In some situations, it may be necessary to provide appropriate accommodations for limited English proficient students and students with disabilities to accurately and effectively measure students' knowledge and skills in the particular content area being assessed.[29]

[28] Standard 7.10 states, "When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance. While initially, the responsibility of the test developer, the test user bears responsibility for uses with groups other than those specified by the developer." *Joint Standards, supra* note 3, at p. 83.

[29] *See Joint Standards, supra* note 3, at p. 143.

(3) *Effectiveness of treatment: Do test scores lead to placements and other consequences that are educationally beneficial?*

The most basic obligation of educators at the elementary and secondary school levels is to meet the needs of students as they find them, with their different backgrounds, and to teach knowledge and skills to allow them to grow to maturity with meaningful expectations of a productive life in the workforce and elsewhere.[30] This obligation regarding elementary and secondary education is no less present when educators administer tests and evaluate and act on students' test results than it is during classroom instruction. Recognizing that tests used in the education setting should be integral to the learning and achievement of students, one federal court distinguished between testing in the employment and education settings:

> 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.[31]

Tests, in short, should be instruments used by elementary and secondary educators to help students achieve their full potential. Test scores should lead to consequences that are educationally beneficial for students. When making high-stakes decisions that involve the use of tests, it is important for policymakers and educators to consider the intended and unintended consequences that may result from the use of the test scores.[32]

---

[30] *See Brown v. Board of Educ.*, 347 U.S. 483, 493 (1954) (stating that "[education] is required in the performance of our most basic public responsibilities, . . . is the very foundation of good citizenship. . . . [and] is [a] principal instrument . . . in preparing [the child] for later professional training . . . .").

[31] *Larry P. v. Riles*, 793 F.2d 969, 980 (9th Cir. 1984) (quoting *Larry P. v. Riles*, 495 F. Supp. 926, 969 (N.D. Cal. 1979)).

[32] Research indicates that students in low-track classes often do not have the opportunity to acquire knowledge and skills strongly associated with future success that is offered to students in other tracks. The National Research Council recommends that neither test scores nor other information should be used to place students in such classes. *See High Stakes*, *supra* note 11, at p. 282.

These criteria [measurement validity, attribution of cause, and effectiveness of treatment], based on established professional standards, lead to the following basic principles of appropriate test use for educational decisions:

- The important thing about a test is not its validity in general, but its validity when used for a specific purpose. Thus, tests that are valid for influencing classroom practice, "leading" the curriculum, or holding schools accountable are not appropriate for making high-stakes decisions about individual student mastery unless the curriculum, the teaching, and the test(s) are aligned.
- Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills. A student's score can be expected to vary across different versions of a test – within a margin of error determined by the reliability of the test – as a function of the particular sample of questions asked and/or transitory factors, such as the student's health on the day of the test. Thus, no single test score can be considered a definitive measure of a student's knowledge.
- An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about the student's knowledge and skills should also be taken into account.
- Neither a test score nor any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional supports. In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.

National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999).

B. Legal Principles

Federal constitutional, statutory, and regulatory principles form the federal legal nondiscrimination framework applicable to the use of tests for high-stakes purposes. Title VI, Title IX, Section 504, and Title II, as well as the equal protection clause of the Fourteenth Amendment to the United States Constitution, prohibit intentional discrimination based on race, national origin, sex, or disability.[33] In addition, the regulations that implement Title VI, Title IX, Section 504, and Title II prohibit intentional discrimination and policies or practices that have a discriminatory disparate

[33] The United States Supreme Court has held that "Title VI itself directly reached only instances of intentional discrimination . . . [but that] actions having an unjustifiable disparate impact on minorities could be redressed through agency regulations designed to implement the purposes of Title VI." *Alexander v. Choate*, 469 U.S. 287, 293 (1985) (discussing *Guardians Ass'n v. City Service Comm'n of N.Y.*, 463 U.S. 582 (1983)). The United States Supreme Court has never expressly ruled on whether Section 504, Title II and/or Title IX directly reach disparate impact discrimination. *See, e.g., Alexander*, 469 U.S. at 294 & n. 11 (noting the possibly different congressional purpose with respect to Section 504 as compared with Title VI). Section 504 and Title II require reasonable modifications where necessary to enable persons with disabilities to participate in or enjoy the benefits of public services. Nonetheless, the regulations implementing Section 504, Title II, and Title IX, like the Title VI regulations, explicitly prohibit actions having discriminatory effects as well as actions that are intentionally discriminatory.

impact on students based on their race, national origin, sex, or disability.[34] The Section 504 regulation and the Individuals with Disabilities Education Act (IDEA) contain specific provisions relative to the use of high-stakes tests for individuals with disabilities.[35]

These sources of legal authority should be considered in conjunction with the test measurement principles discussed in this guide to ensure that standardized tests are used in a manner that supports sound educational decisions, regardless of the race, national origin (including limited English proficiency), sex, or disability of the students affected. Some of the issues that have been considered by federal courts in assessing the legality of specific testing practices for making high-stakes decisions include:[36]

- The use of an educational test for a purpose for which the test was not designed or validated;
- The use of a test score as the sole criterion for the educational decision;
- The nature and quality of the opportunity provided to students to master required content, including whether classroom instruction includes the material covered by a test administered to determine student achievement;
- The significance of any fairness problems identified, including evidence of differential prediction of a criterion and possible cultural biases in the test or in test items; and
- The educational basis for establishing passing or cutoff scores.

[34] *See* 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. §§ 106.21(b)(2), 106.36(b), 106.52 (Title IX); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3) (Title II).

The authority of federal agencies to issue regulations with an "effects" standard has been consistently acknowledged by U.S. Supreme Court decisions and applied by lower federal courts addressing claims of discrimination in education. *See, e.g., Alexander*, 469 U.S. at 289-300; *Guardians Ass'n*, 463 U.S. at 584-93; *Lau v. Nichols*, 414 U.S. 563, 568 (1974); *see also* 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 of 1964* (July 14, 1994).

[35] The IDEA establishes rights and protections for students with disabilities and their families. It also provides federal funds to local school districts and state agencies to assist in educating students with disabilities. *See* Individuals with Disabilities Education Act, 20 U.S.C. §§ 1400(1)(c) *et seq*. The specific sections of the regulations implementing Section 504 and the IDEA bearing on testing are 20 U.S.C. §§ 1412(a)(17), 1414(b); 34 C.F.R. §§ 104.4(b)(4), 104.33, 104.35, 104.42(b), 104.44, 300.138 - .139, 300.530 - .536.

[36] For specific court decisions examining these issues, *see* discussion *infra* Chapter 2 (Legal Principles) & nn. 161-165.

1. Frameworks for Analysis

a. Different Treatment

Under federal law, policies and practices generally must be applied consistently to similarly situated individuals or groups, regardless of their race, national origin, sex, or disability. For example, a court concluded that a school district had intentionally treated students differently on the basis of race where minority students whose test scores qualified them for two or more ability levels were more likely to be assigned to the lower level class than similarly situated white students, and no explanatory reason was evident.[37]

In addition, educational systems that were previously segregated by race in violation of the Fourteenth Amendment and have not achieved unitary status have an obligation to dismantle their prior *de jure* segregation. In such instances, school districts are under "a 'heavy burden' of showing that actions that [have] increased or continued the effects of the dual system serve important and legitimate ends."[38] When a school district or other educational system uses a test or assessment procedure for a high-stakes purpose that has a racially disparate effects, the school district can justify the test use only by showing that the test results are not due to the present effects of prior segregation or that the practice or procedure remedies the adverse effects of such segregation by offering better educational opportunities.[39]

[37] *See People Who Care v. Rockford Bd. of Educ.*, 851 F. Supp. 905, 958-1001 (N.D. Ill. 1994), *remedial order rev'd, in part*, 111 F.3d 528 (7th Cir. 1997). On appeal, the Seventh Circuit Court of Appeals stated that the appropriate remedy based on the facts in this case was to require the district to use objective, non-racial criteria to assign students to classes, rather than abolishing the district's tracking system. *See People Who Care*, 111 F.3d at 536.

[38] *Dayton Bd. of Educ. v. Brinkman*, 443 U.S. 526, 538 (1979) (quoting *Green v. County School Bd.*, 391 U.S. 430, 439 (1968)).

[39] *See Debra P. v. Turlington*, 644 F.2d 397, 407 (5th Cir. 1981) ("[Defendants] failed to demonstrate either that the disproportionate failure [rate] of blacks was not due to the present effects of past intentional segregation or, that as presently used, the diploma section was necessary [in order] to remedy those effects."); *McNeal v. Tate County Sch. Dist.*, 508 F.2d 1017, 1020 (5th Cir. 1975) (ability grouping method that causes segregation may nonetheless be used "if the school district can demonstrate that its assignment method is not based on the present results of past segregation or that the method of assignment will remedy such effects through better educational opportunities"); *see also 3GI Forum v. Texas Educ. Agency*, 87 F. Supp. 2d 667, 674 (W.D. Tex. 2000) (inequalities in education did not cause disproportionate failure rate since all students have an equal opportunity to learn the items on the test and testing program, along with school accountability and remedial follow up, helps to address the effects of any prior discrimination and remaining disparities); *United States v. Fordice*, 505 U.S. 717, 731 (1992) ("If the State [university system] perpetuates policies and practices traceable to its prior system that continue to have segregative effects . . . and such policies are without sound educational justification and can be practically eliminated, the State has not satisfied its burden of proving that it has dismantled its prior system.").

b. Disparate Impact

The federal nondiscrimination regulations also provide that a recipient of federal funds may not "utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination."[40] Thus, discrimination under federal law may occur where the application of neutral criteria has disparate effects and those criteria are not educationally justified or there are alternative practices available that are equally effective in serving the educational institution's goals that have a less disparate impact.

> It is ... important to note that group differences in test performance do not necessarily indicate problems in a test, because test scores may reflect real differences in achievement. These, in turn, may be due to a lack of access to a high quality curriculum and instruction. Thus, a finding of group differences calls for a careful effort to determine their cause.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 5 (Jay P. Heubert & Robert M. Hauser eds., 1999).

The disparate impact analysis has been frequently misunderstood to indicate a violation of law based merely on disparities in student performance and to obligate educational institutions to change their policies and procedures to guarantee equal results. Under federal law, a statistically significant difference in outcomes creates the need for further examination of the educational practices in question that have caused the disparities in order to ensure accurate and nondiscriminatory decision-making, but disparate impact alone is not sufficient to prove a violation of federal civil rights laws.

Courts applying the disparate impact test have generally examined three questions to determine if the practices at issue are discriminatory: (1) Does the practice or procedure in question result in substantial differences in the award of benefits or services based on race, national origin, or sex? (2) Is the practice or procedure educationally justified? (3) Is there an equally effective alternative that can

[40] 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3)(i) (Title II); *see also* 34 C.F.R. §§ 106.21, 106.31, 106.36(b), 106.52 (Title IX). In *Guardians Association*, the United States Supreme Court upheld the use of the effects test, stating that the Title VI regulation forbids the use of federal funds "not only in programs that intentionally discriminate, but also in those endeavors that have a [racially disproportionate] impact on racial minorities." 463 U.S. at 589.

accomplish the institution's educational goal with less disparity?[41] (For a discussion of disability discrimination, including disparate impact discrimination, *see* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing Students with Disabilities).[42])

Under the disparate impact analysis, the party challenging the test has the burden of establishing disparate impact, generally through evidence of a statistically significant difference in the awards of benefits or services. If disparate impact is established, the educational institution must demonstrate the "educational necessity" for the practice in question.[43] If sufficient evidence of an educational justification has been provided, the party challenging the test must then demonstrate, in order to prevail, that an alternative practice with less disparate impact is equally effective in meeting the institution's educational goals or needs.[44]

2. Principles Relating to Inclusion and Accommodations

a. Limited English Proficient Students

The obligations of states and school districts with regard to testing of limited English proficient students for high-stakes purposes in elementary and secondary schools must be examined within the overall context of the Title VI obligation to provide equal educational opportunities to limited English proficient students. Under Title VI, school districts have an obligation to identify limited English proficient students and to provide them with an instructional program that enables them to acquire English-language proficiency as well as the knowledge and skills that all students are required to master.[45] This program of instruction should provide students with a meaningful opportunity to acquire the academic knowledge and skills covered by tests required for graduation or other educational benefits.

[41] Courts use a variety of terms when discussing whether an alternative offered by the party challenging the practice is feasible and would also effectively meet the institution's goals. *See, e.g., Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1417 (11th Cir. 1985) (party challenging the practice "may ultimately prevail by proffering an equally effective alternative practice which results in less racial disproportionality"); *Elston v. Talladega*, 997 F.2d 1394, 1407 (11th Cir. 1993) (party challenging the practice "will still prevail if able to show that there exists a comparably effective alternative practice which would result in less disproportionality"). These terms ("equally effective" and "comparably effective") appear to be used synonymously.

[42] Disparate impact disability discrimination may take forms that are not always amenable to analysis through the three-part approach used in race and sex discrimination cases. For example, statistical proof may not be necessary when evaluating the effects of architectural barriers. *See Alexander*, 469 U.S. at 297-300. For this reason, disability discrimination is discussed separately in this guide. *See* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).

[43] *See Elston*, 997 F.2d at 1412.

[44] *See Georgia State Conf.*, 775 F.2d at 1417; *see also* Department of Justice, *Title VI Legal Manual*, p. 2.

[45] *See* Equal Educational Opportunities Act of 1974, 20 U.S.C. §§ 1701-1720; *Lau*, 414 U.S. at 568-69; *Castaneda v. Pickard*, 648 F.2d 989, 1011 (5th Cir. 1981); Michael L. Williams, Former Assistant Secretary for Civil Rights, *Memorandum to OCR Senior Staff* (September 27, 1991) (hereinafter *Williams Memorandum*).

In addition, states or school districts using tests for high-stakes purposes must ensure that, as with all students, the tests effectively measure limited English proficient students' knowledge and skills in the particular content area being assessed. For limited English proficient elementary and secondary school students in particular, it may be necessary in some situations to provide accommodations so that the tests provide accurate information about the knowledge and skills intended to be measured.[46]

b. Students with Disabilities

Under Section 504, Title II, and the IDEA,[47] school districts have a responsibility to provide elementary and secondary school students with disabilities with a free appropriate public education. Providing effective instruction in the general curriculum for students with disabilities is an important aspect of providing a free appropriate public education. Under federal law, students with disabilities must be included in statewide or districtwide assessment programs and provided with appropriate accommodations, if necessary.[48] There must be an individualized determination of whether a student with a disability will participate in a particular test and the appropriate accommodations, if any, that a student with a disability will need. This individualized determination must be addressed through the individualized education

[46] States and school districts are also required to provide LEP students with "reasonable adaptations and accommodations" in certain situations when using assessments for the purpose of holding schools and districts accountable for student performance under Title I. *See* Title I of the Elementary and Secondary Education Act, 20 U.S.C. § 6311(b)(3)(F)(ii). Moreover, Title I requires States, to the extent practicable, to provide native-language assessments to LEP students for Title I accountability purposes if that is the language and form of assessment most likely to yield accurate and reliable information about what students know and can do. *See* 20 U.S.C. § 6311(b)(3)(F)(iii). For a discussion of comparability issues arising in the testing of LEP students, *see* discussion *infra* Chapter 2 (Legal Principles) Part II (Testing of Students with Limited English Proficiency).

[47] The Section 504 regulation is found at 34 C.F.R. Part 104. The Title II regulation is found at 28 C.F.R. Part 35. The IDEA regulation is found at 34 C.F.R. Part 300.

[48] States and school districts are also required to provide students with disabilities with "reasonable adaptations and accommodations" in certain situations when using assessments for the purpose of holding schools and districts accountable for student performance under Title I. *See* 20 U.S.C. § 6311(b)(3)(F)(ii).

[49] Under the IDEA, students with disabilities must be included in state and districtwide assessment programs. *See* 34 C.F.R. § 300.138(a). However, if the IEP team determines that a student should not participate in a particular statewide or districtwide assessment of student achievement (or part of such an assessment), the student's IEP must include statements of why that test is not appropriate for the student and how the student will be assessed. *See* 34 C.F.R. § 300.347(a)(5). The IDEA also requires state or local education agencies to develop guidelines for students with disabilities who cannot take part in state and districtwide assessments to participate in alternate assessments; these alternate assessments must be developed and conducted beginning not later than July 1, 2000. *See* 34 C.F.R. § 300.138(b).

program (IEP) process or other applicable evaluation procedures and included in either the student's IEP or Section 504 plan.[49] The IDEA also requires state or local education agencies to develop guidelines for the relatively small number of students with disabilities who cannot take part in statewide or districtwide tests to participate in alternate assessments.[50]

Finally, under Section 504, post-secondary education institutions may not make use of any test or criterion for admission that has a disproportionate adverse impact on individuals with disabilities unless (1) the test or criterion, as used by the institution, has been validated as a predictor of success in the education program or activity and (2) alternate tests or criteria that have a less disproportionate adverse impact are not shown to be available by the party asserting that the test or criterion is discriminatory.[51] Admissions tests must be selected and administered so as best to ensure that, when a test is administered to an applicant with a disability, the test results accurately reflect the applicant's aptitude or achievement level, rather than reflecting the effect of the disability (except where the functions impaired by the disability are the factors the test purports to measure).[52] A student requesting an accommodation must provide documentation of the disability and the type of accommodation needed. Admissions tests designed for persons with impaired sensory, manual, or speaking skills must be offered as often and in as timely a manner as are other admissions tests. Admissions tests also must be offered in facilities that, on the whole, are accessible to individuals with disabilities.

3. Federal constitutional Questions Related to Testing of Elementary and Secondary Students for High-Stakes Purposes

The equal protection and due process requirements of the Fifth and Fourteenth Amendments to the U.S. Constitution also apply to ensure that high-stakes decisions by public schools or states involving the use of tests are made appropriately.[53] The equal protection principles involved in discrimination cases are, generally speaking, the same

[50] *See* 34 C.F.R. § 300.138(b).

[51] *See* 34 C.F.R. § 104.42(b)(2).

[52] *See* 34 C.F.R. § 104.42(b)(3).

[53] The requirements of Title VI, Title IX and Section 504 apply only to recipients of federal financial assistance. The protections afforded by the Fifth and Fourteenth Amendments to the U.S. Constitution apply to actions by "state actors" and are not dependent on their receipt of federal financial assistance.

[54] Federal cases may also involve equal protection challenges to a jurisdiction's use of tests in which the claim is not based on race or sex discrimination, but, instead, on assertions that the classifications made by the jurisdiction on the basis of test scores are unreasonable, regardless of the race or sex of the students affected. *See GI Forum*, 87 F. Supp. 2d at 682. As a general matter, courts express reluctance to second guess a state's educational policy choices when faced with such challenges, although they recognize that a state cannot "exercise that [plenary] power without reason and without regard to the United States Constitution." *Debra P.*, 644 F.2d at 403. When there is no claim of discrimination based on membership in a suspect class, the equal protection claim is reviewed under the rational basis standard. In these cases, the jurisdiction need show only that the use of the

as the standards applied to intentional discrimination (or different treatment) claims under the applicable federal nondiscrimination statutes.[54] Courts addressing due process claims have examined three questions related to the use of tests as bases for promotion or graduation decisions:

- Is the testing program related to a legitimate educational purpose?
- Have students received adequate notice of the test and its consequences?
- Have students actually been taught the knowledge and skills measured by the test?

Federal courts have typically deferred to educators' authority to formulate appropriate educational goals.[55] For example, improving the quality of education, ensuring that students can compete on a national and international level, and encouraging educational achievement through the establishment of academic standards have been found to be legitimate goals for testing programs.[56] The constitutional inquiry then proceeds to examine whether the challenged testing program is reasonably related to the educators' legitimate goals or whether the program is arbitrary and capricious or fundamentally unfair.[57]

In due process cases, courts have generally required advance notice of test requirements in order to give students a reasonable chance to understand the standards against which they will be evaluated and to learn the material for which they are to be accountable.[58] A reasonable transition period is required between the development of a new academic requirement and the attachment of high-stakes consequences to tests used to measure academic achievement. That time period varies, however, depending upon the precise context in which the high-stakes decision is to be made. Relevant inquiries affecting determinations about the constitutionality of notice and timing have included questions about the alignment of curriculum and instruction with material tested, the number of test taking opportunities provided to students, tutorial or remedial opportunities provided to students, and whether factors in

tests has a rational relationship to a valid state interest. *See Debra P.*, 644 F.2d at 406; *Erik V. v. Causby*, 977 F. Supp. 384, 389 (E.D. N.C. 1997).

[55] *See Regents of the Univ. of Mich. v. Ewing*, 474 U.S. 214, 226-27 (1985); *Debra P.*, 644 F.2d at 406; *Anderson v. Banks*, 520 F. Supp. 472, 506 (S.D. Ga. 1981).

[56] *See Ewing*, 474 U.S. at 226-27; *Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 506.

[57] *See Ewing*, 474 U.S. at 222, 226-27; *Debra P.*, 644 F.2d at 406; *GI Forum*, 87 F. Supp. 2d at 682; *Anderson*, 520 F. Supp. at 506.

[58] *See Brookhart v. Illinois Bd. Of Educ.*, 697 F.2d 179, 185 (7th Cir. 1983); *Debra P.*, 644 F.2d at 404; *Erik V.*, 977 F. Supp. at 389-90; *Anderson*, 520 F. Supp. at 1410-12.

addition to test scores can affect high-stakes decisions.

Ultimately, in due process cases, federal courts have required, as a matter of "fundamental fairness," that students have a reasonable opportunity to learn the material covered by the test where passing the test is a condition of receipt of a high school diploma or a condition for grade-to-grade promotion.[59] For the test to meaningfully measure student achievement, the test, the curriculum, and classroom instruction should be aligned.[60]

[59] *See Brookhart*, 697 F.2d at 184-87; *Debra P.*, 644 F.2d at 406; *GI Forum*, 87 F. Supp. 2d at 682; *Anderson*, 520 F. Supp. at 509.

[60] *See Brookhart*, 697 F.2d at 184-87; *Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 509. Insofar as due process cases may involve additional questions regarding the validity, reliability, and fairness of the test used to address the educational institution's stated purposes, these issues are discussed in the portions of the guide addressing discrimination under federal civil rights laws.

CHAPTER 1. Test Measurement Principles

This chapter explains basic test measurement standards and related educational principles for determining whether tests used as part of making high-stakes decisions for students provide accurate and fair information. As explained in chapter two below, federal court decisions have been informed and guided by professional test measurement standards and principles. Professional test measurement standards, products of the test measurement community, can provide a basis for compliance with federal nondiscrimination laws.[61] This chapter is intended as a helpful discussion of how to understand test measurement concepts and their use. These are not specific legal requirements, but rather are foundations for understanding appropriate test use.

Educational institutions use tests to accomplish specific purposes based on their educational goals, including making placement, promotion, graduation, admissions, and other decisions. It is only after they have determined the underlying goal they want to accomplish that they can identify the types of information that will best inform their decision-making. Information may include test results, as well as other relevant measures, that will be able to effectively, accurately, and fairly address the purposes and goals specified by the institutions.[62] As stated in the *Joint Standards*, "When interpreting and using scores about individuals or groups of students, consideration of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance.... As the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly."[63]

[61] *See, e.g.*, National Research Council. *High Stakes: Testing for Tracking, Promotion, and Graduation*, pp. 59-60 (Jay P. Heubert & Robert M. Hauser eds., 1999) (hereinafter *High Stakes*).

[62] Among other considerations, institutions will determine if they want test score interpretations that are norm-referenced or criterion-referenced, or both. Norm-referenced means that the performances of students are compared to the performances of other students in a specified reference population; criterion-referenced indicates the extent to which students have mastered specific knowledge and skills.

[63] American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, p. 141 (1999) (hereinafter *Joint Standards*); *See also Standard 13.7 (n.8) in Joint Standards, supra*, at p. 146.

Although this guide focuses on the use of tests, policymakers and educators need to consider the soundness and relevance of the entire high stakes decision-making process including other information used in conjunction with test results.[64]

In using tests to make high-stakes decisions, educational institutions should ensure that the test will provide accurate results that are valid, reliable, and fair for all test takers. This includes obtaining adequate evidence of test quality about the current test being proposed, evaluating the evidence, and ensuring that appropriate test use is based on adequate evidence.[65] When test results are used to make high-stakes decisions about student promotion or graduation, educational institutions should ensure that evidence is available which documents that students have had an adequate opportunity to learn the material being tested.[66]

I. Key Considerations in Test Use

This section addresses the fundamental concepts of test validity and reliability. It will also discuss issues associated with ensuring fairness in the meaning of test scores, and issues related to using appropriate cutscores in high-stakes tests. Test developers and users as appropriate determine adequate validity and reliability, ensure fairness, and determine where to set and how to use cutscores appropriately for all students by accumulating evidence of test quality from all relevant groups of test takers.

A. Validity

Test validity refers to a determination of how well a test actually measures what it says it measures. The *Joint Standards* defines validity as "[t]he degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed

[64] "It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions." *Joint Standards*, *supra* note 63, at p. 141.

[65] In order to provide educational institutions with tests that are accurate and fair, test developers should develop tests in accordance with professionally recognized standards, and provide educational institutions with adequate evidence of test quality.

Standard 1.4 states, "If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary." *Joint Standards*, *supra* note 63, at p. 18.

Standard 11.2 states, "When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test's validity and reliability for this purpose." *Joint Standards*, *supra* note 63, at p. 113.

[66] *See Standard 7.5 and 13.5 (n.22) in Joint Standards*, *supra* note 63, at pp. 82, 146.

Standard 7.5 states, "In testing applications involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test taker's performance on that test at that time." *Joint Standards*, *supra* note 63, at p. 82.

uses of a test."[67] The demonstration of validity is multifaceted and must always be determined within the context of the specific use of a test. In order to promote readability, the discussion on validity presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. The *Joint Standards* identify and discuss in detail principles related to determining the validity of test results within the context of their use, and readers are encouraged to review the *Joint Standards*, Chapter 1, Validity, for additional, relevant discussion.[68]

There are three central points to keep in mind:

- The focus of validity is not really on the test itself, but on the validity of the inferences drawn from the test results for a given use.
- All validity is really a form of "construct validity."
- In validating the inferences of the test results, one must also consider the consequences of the test's interpretation and use. Consequences may be the result of improper test procedures or they may involve factors external to the test.

1. Validity of the Inferences Drawn from the Scores

It is not the test that is validated per se, but the inferences or meaning derived from the test scores for a given use—that is, for a specific type of purpose, in a specific type of situation, and with specific groups of students. The meaning of test scores will differ based on such factors as how the test is designed, the types of questions that are asked, and the documentation that supports how all groups of students are interpreting what the test is asking and how effectively their performance can be generalized beyond the test.

For instance, in one case, the educational institution may want to evaluate how well students can analyze complex issues and evaluate implications in history. For a given amount of test time, they would want to use a test that measures the ability of students to think deeply about a few selected history topics. The meaning of the scores should reflect this purpose and the limits of the range of topics being measured on the test. In another case, the institution may want to assess how well students know a range of facts about a wide variety of historical events. The institution would want to use a test that measures a broad range of knowledge about many different occurrences in history. The inferences drawn from the scores should be validated to determine how well they measure students' knowledge of a broad range of historical facts.

[67] *Joint Standards*, *supra* note 63, at pp. 9, 184.

[68] *See Joint Standards*, *supra* note 63, at pp. 9-24.

2. Construct Validity

Construct validity refers to the degree to which the scores of test takers accurately reflect the constructs a test is attempting to measure. The *Joint Standards* defines a construct as "the concept or the characteristic that a test is designed to measure."[69] Test scores and their inferences are validated to measure one or more constructs, which together comprise a particular content domain.[70] In K-12 education, these domains are often codified in state or district content standards covering various subject areas. For instance, the domain of mathematics as described in the state's elementary mathematics content standards, may involve the constructs of mathematical problem-solving and knowledge of number systems. Items may be selected for a test that sample from this domain, and should be properly representative of the constructs identified within it. In that way, the meaning of the test scores should accurately reflect the knowledge and skills defined in the mathematics content standards domain.

Validity should be viewed as the overarching, integrative evaluation of the degree to which all accumulated evidence supports the intended interpretation of the test scores for a proposed purpose.[71] This unitary and comprehensive concept of validity is referred to as "construct validity." Different sources of validity evidence may illuminate different aspects of validity, but they do not represent distinct types of validity.[72]

Therefore, "construct validity" is not just one of the many types of validity—it *is* validity. Demonstrating construct validity then means gathering a variety of types of evidence to support the intended interpretations and uses of test scores. All validity evidence and the interpretation of the evidence are focused on the basic question: Is the test measuring the concept, skill, or trait in question? Is it, for example, really measuring mathematical reasoning or reading comprehension for the types of students that are being tested? A variety of types of evidence can be used to answer this question—none of which provides a simple yes or no answer. The exact nature of the types of evidence that need to be accumulated is directly related to the intended use of

[69] *Joint Standards*, *supra* note 63, at p. 173.

[70] The *Joint Standards* defines a content domain as "the set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified." *Joint Standards*, *supra* note 63, at p. 174. A domain, then, represents a definition of a content area for the purposes of a particular test. Other tests will likely have a different definition of what knowledge and skills a particular content area entails.

[71] *See Joint Standards*, *supra* note 63, at pp. 9-11, 184.

[72] Therefore, construct validity can be seen as an umbrella that encompasses what has previously been described as predictive validity, content validity, criterion validity, discriminant validity, etc. Rather, these terms refer to types or sources of evidence that can be accumulated to support the validity argument. Definitions of these terms can be found in Appendix B, Measurement Glossary.

the test, which includes evidence regarding the skills and knowledge being measured, evidence documenting validity for the stated purpose, and evidence of validity for all groups of students taking the test.[73]

For instance, an educational institution may want to use a test to help make promotion decisions. It may also want to use a test to place students in the appropriate sequence of courses. In each situation, the types of validity evidence an institution would expect to see would depend on how the test is being used.

In making promotion decisions, the test should reflect content the student has learned. Appropriate validation would include adequate evidence that the test is measuring the constructs identified in the curriculum, and that the inferences of the scores accurately reflect the intended constructs for all test takers. Validation of the decision process involving the use of the test would include adequate evidence that low scores reflect lack of knowledge of students after they have been taught the material, rather than lack of exposure to the curriculum in the first place.

In making placement decisions, on the other hand, the test may not need to measure content that the student has already learned. Rather, at least in part, the educational institution may want the test to measure aptitude for the future learning of knowledge or skills that have been identified as necessary to complete a course sequence. Appropriate validation would include documentation of the relationship between what constructs are being measured in the test, and what skills and knowledge are actually needed in the future placements. Evidence should also provide documentation that scores are not significantly confounded by other factors irrelevant to the knowledge and skills the test is intending to measure.

Institutions often think about using the same test for two or more purposes. This is appropriate as long as the validity evidence properly supports the use for the test for each purpose, and properly supports that the inferences of the results accurately reflect what the test is measuring for all students taking the test.[74]

The empirical evidence related to the various aspects of construct validity is collected throughout test development, during test construction, and after the test is completed. It is important for educators and policymakers to understand and expect that the accumulated evidence spans the range of test development and implementation. There is not just one set of documentation collected at one point in time.[75]

---

[73] Rather than follow the traditional nomenclature (e.g. predictive validity, content validity, criterion validity, discriminant validity, etc.), the *Joint Standards* defines sources of validity evidence as evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. *See Joint Standards, supra* note 63, at pp. 11-17.

[74] *See Joint Standards, supra* note 63, at pp. 9-24 (Chapter 1, Validity).

When the empirical database is large and includes results from a number of studies related to a given purpose, situation, and type of test takers, it may be appropriate to generalize validity findings beyond validity data gathered for one particular test use. That is, it may be appropriate to use evidence collected in one setting when determining the validity of the meaning of the test scores for a similar use. If the accumulated validity evidence for a particular purpose, situation, or subgroup is small, or features of the proposed use of the test differ markedly from an adequate amount of validity evidence already collected, evidence from this particular type of test use will generally need to be compiled.[76] Regardless of where the evidence is collected, educational institutions should expect adequate documentation of construct validity based on needs defined by the particular purposes and populations for which a test is being used.

3. Sources of Validity Error

When considering the types of construct validity evidence to collect, the *Joint Standards* emphasize that it is important to guard against the two major sources of validity error. This error can distort the intended meaning of scores for particular groups of students, situations, or purposes.[77]

One potential source of error omits some important aspects of the intended construct being tested. This is called construct underrepresentation.[78] An example would be a test that is being used to measure English language proficiency. When the institution has defined English language proficiency as including specific skills in listening, speaking, reading, and writing the English language, and wants to use a test which measures these aspects, construct underrepresentation would occur if the test only measured the reading skills.

---

[75] Standard 3.6 states "The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented." *Joint Standards*, *supra* note 63, at p. 44

[76] As indicated in the *Joint Standards*, "The extent to which predictive or concurrent evidence of validity generalization can be used in new situations is in large measure a function of accumulated research. Although evidence of generalization can often help to support a claim of validity in a new situation, the extent of available data limits the extent to which the claim can be sustained." *Joint Standards*, *supra* note 63, at pp. 15-16.

[77] *See Joint Standards*, *supra* note 63, at p. 10.

[78] *See* Samuel Messick, *Validity*, in *Educational Measurement*, pp. 13-103 (Robert L. Linn ed., 3rd ed. 1989) (hereinafter Messick, *Validity*); Samuel Messick, *Validity of Psychological Assessment: Validations of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*, American Psychologist 50(9), pp. 741-749 (September 1995) (hereinafter Messick, *Validity of Psychological Assessment*).

The other potential source of error occurs when a test measures material that is extraneous to the intended construct, confounding the ability of the test to measure the construct that it intends to measure. This source of error is called construct irrelevance.[79] For instance, how well a student reads a mathematics test may influence the student's subtest score in mathematics computation. In this case, the student's reading skills may be irrelevant when the skill of mathematics computation is what is being measured by the subtest.[80]

An essential part of the accumulated validity information is collecting evidence not only about what a test measures in particular types of situations or for particular groups of students, but also evidence that seeks to document that the intended meaning of the test scores is not unduly influenced by either of the two sources of validity error.

4. Considering the Consequences of Test Use

Evidence about the intended and unintended consequences of test use can provide important information about the validity of the inferences to be drawn from the test results, or it can raise concerns about an inappropriate use of a test where the inferences may be valid for other uses.

For instance, significant differences in placement test scores based on race, gender, or national origin may trigger a further inquiry about the test and how it is being used to make placement decisions.[81] The validity of the test scores would be called into question if the test scores are substantially affected by irrelevant factors that are not related to the academic knowledge and skills that the test is supposed to measure.[82]
On the other hand, a test may accurately measure differences in the level of students' academic achievement. That is, low scores may accurately reflect that some students

[79] *See* Messick, *Validity*, *supra* note 78; Messick, *Validity of Psychological Assessment*, *supra* note 78.

[80] On the other hand, if an item is measuring the student's ability to apply mathematical skills in a written format (for instance when an item requires students to fill out an order form), then writing skills may not be extraneous to the construct being measured in this item.

[81] *See* Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (1988).

[82] *See Standard 1.24, 7.5 (n.66) and 7.6 in Joint Standards*, *supra* note 63, at pp. 23-24, 82.

Standard 7.6 states, "When empirical studies of differential prediction of a criterion for members of different subgroups are conducted, they should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables." *Joint Standards*, *supra* note 63, at p. 82.

Standard 1.24 states, "When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure fully to represent the intended construct." *Joint Standards*, *supra* note 63, at p. 23.

do not know the content. However, test users should ensure that they interpret those scores correctly in the context of their high-stakes decisions.[83] For instance, test users could incorrectly conclude that the scores reflect lack of ability to master the content for some students when, in fact, the low test scores reflect the limited educational opportunities that the students have received. In this case, it would be inappropriate to use the test scores to place low-performing students in a special services program for students who have trouble learning and processing academic content.[84] It would be appropriate to use the test to evaluate program effectiveness, however.[85]

> Standard 13.1
>
> When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

[83] *See Standard 1.22, 1.23, 7.5, (n.66), 7.10 (n.28) and 13.9 (n.19) in Joint Standards, supra* note 63, at pp. 23, 82, 83, 147.

Standard 1.22 states, "When it is clearly stated or implied that recommended test use will results in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence." *Joint Standards, supra* note 63, at p. 23.

Standard 1.23 states, "When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted." *Joint Standards, supra* note 63, at p. 23.

[84] The Comment under Standard 13.1 states, "Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they result directly in important decisions for individuals or institutions. Frequent concerns include narrowing the curriculum to focus only on the objectives tested, increasing the number of dropouts among students who do not pass the test, or encouraging other instructional or administrative practices simply designed to raise test scores rather than to affect the quality of education." *Joint Standards, supra* note 63, at p. 145.

[85] *See High Stakes, supra* note 61, at pp. 89-113.

B. Reliability

Reliability refers to the degree of consistency of test results, over test administrations, forms, items, scorers, and/or other facets of testing.[86] All indices of reliability are estimates of consistency, and all the estimates contain some error, since no test or other source of information is ever an "error-free" measure of student performance.[87] An example of reliability of test results on different occasions is when the same students, taking the test multiple times, receive similar scores. Consistency over parallel forms of a test occurs when forms are developed to be equivalent in content and technical characteristics. Reliability can also include estimates of a high degree of relationship across similar items within a single test or subtest that are intended to measure the same knowledge or skill. For judgmentally scored tests, such as essays, another widely used index of reliability addresses stability across raters or scorers. In each case, reliability can be estimated in different ways, using one of several statistical procedures.[88] Different kinds of reliability estimates vary in degree and nature of generalization.

In order to promote readability, the discussion on reliability presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. Readers are encouraged to review Chapter 2, Reliability and Errors of Measurement, in the *Joint Standards* for additional, relevant information.[89]

---

[86] Evaluating the reliability of test results includes identifying the major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected, or the generalizability of results across items, forms, raters, sampling, administrations, and other measurement facets.

[87] All sources of assessment information, including test results, include some degree of error. There are two types of error. The first is random error that affects scores in such a way that sometimes students will score lower and sometimes higher than their "true" score (the actual mastery level of the students' knowledge and skills). This type of error, also known as measurement error, particularly affects reliability of scores. Therefore, test scores are considered reliable when evidence demonstrates that there is a minimum amount of random measurement error in the test scores for a given group.

The second type of error that affects test results is systematic error. Systematic error consistently affects scores in one direction; that is, this type of error causes some students to consistently score lower or consistently score higher than their "true" (or actual) level of mastery. For instance, visually impaired students will consistently score lower than they should on a test which has not been administered for them in Braille or large print, because their difficulty in reading the items on the page will negatively impact their score. This type of error generally affects the validity of the interpretation of the test results and is discussed in the validity section above. Systematic error should also be minimized in a test for all test takers.

When educators and policymakers are evaluating the adequacy of a test for their local population of students, it is important to consider evidence concerning both types of error.

[88] These types of reliability estimates are known as test-retest, alternate forms, internal consistency, and inter-rater estimates, respectively. *See Joint Standards*, *supra* note 63, at pp. 25-36 (examples of different procedures).

[89] *See Joint Standards*, *supra* note 63, pp. 25-36.

C. Fairness

> Fairness, like validity, cannot be properly addressed as an afterthought.... It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation, and use.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, pp. 80-81 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Tests are fair when they yield score interpretations that are valid and reliable for all groups of students who take the tests. That is, the tests must measure the same academic constructs (knowledge and skills) for all groups of students who take them, regardless of race, national origin, gender, or disability. Similarly, the scores must not substantially and systematically underestimate or overestimate the knowledge or skills of members of a particular group. The *Joint Standards* discuss fairness in testing in terms of lack of bias, equitable treatment in the testing process, equal scores for students who have equal standing on the tested constructs, and, depending on the purpose, equity in opportunity to learn the material being tested.[90] In order to promote readability, the discussion on fairness presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. Readers are encouraged to review Chapter 7, Fairness in Testing and Test Use, in the *Joint Standards* for additional, relevant information.[91]

1. Fairness in Validity

Demonstrating fairness in the validation of test score inferences focuses primarily on making sure that the scores reflect the same intended knowledge and skills for all students taking the test. For the most part this means that the test should minimize the measurement of material that is extraneous to the intended constructs and which confounds the ability of the test to accurately measure the constructs that it intends to

---

[90] *See Joint Standards, supra* note 63, at pp. 74-80. In test measurement, the term fairness has a specific set of technical interpretations. Four of these interpretations are discussed in the *Joint Standards*. For instance, bias is discussed in relation to fairness and is defined in the *Joint Standards* in two ways: "In a statistical context, (bias refers to) a systematic error in a test score. In discussing test fairness, bias (also) may refer to construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers." *Joint Standards, supra* note 63, at p. 172. Fairness as equitable treatment in the testing process "requires consideration not only of the test itself, but also the context and purpose of testing, and the manner for which test scores are used." *Joint Standards, supra* note 63, at p. 74. Equal scores for students of equal standing reflects that "examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership." *Joint Standards, supra* note 63, at p. 74. For purposes such as promotion and graduation, "[w]hen some test takers have not had the opportunity to learn the subject matter covered by the test content, they are likely to get low scores . . . low scores may have resulted in part from not having had the opportunity to learn the material tested as well as from having had the opportunity and failed to learn." *Joint Standards, supra* note 63, at p. 76.

[91] *See Joint Standards, supra* note 63, at pp. 73-84.

measure. Rather, a test score should accurately reflect how well each student has mastered the intended constructs. The score should not be significantly impacted by construct irrelevant influences.

The *Joint Standards* identify a number of standards that outline important elements related to validly measuring the intended constructs for all students.[92] The elements span considerations of test development, test implementation, and the proper use of reported test results.

Documenting fairness during test development involves gathering adequate evidence that items and test scores are constructed so that the inferences validly reflect what is intended. For all groups of test takers, evidence should support that valid inferences can be drawn from the scores.[93] When credible research reports that item and test results differ in meaning across examinee subgroups, then to the extent feasible, separate validity evidence for each relevant subgroup should be collected.[94] When items function differently across relevant subgroups, appropriate studies should be conducted, when feasible, so that bias in items due to test design, content, and format is detected and eliminated.[95] Developers should strive to identify and eliminate language, form, and content in tests that have a different meaning in one subgroup than in others, or that generally have sensitive connotations, except when judged to be necessary for adequate representation of the intended constructs.[96] Adequate

---

[92] *See Joint Standards, supra* note 63, at pp. 80-84

[93] Standard 7.2 states, "When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance of some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores." *Joint Standards, supra* note 63, at p. 81.

[94] Standard 7.1 states, "When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions." *Joint Standards, supra* note 63, at p. 80.

Standard 7.3 states, "When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups." *Joint Standards, supra* note 63, at p. 81.

[95] *See Standard 7.3 (n.94) in Joint Standards, supra* note 63, at p. 81.

[96] *See Standard 7.3 (n.94) and 7.4 in Joint Standards, supra* note 63, at pp. 81-82).

Standard 7.4 states, "Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain." *Joint Standards, supra* note 63, at p. 82.

subgroup analyses should be conducted when evaluating the validity of scores for prediction purposes.[97]

Adequate evidence should document the fair implementation of tests for all test takers. The testing process should reflect equitable treatment for all examinees.[98] The Joint Standards state, "In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct."[99]

Documentation of appropriate reporting and test use should be available. Reported data should be clear and accurate, especially when there are high-stakes consequences for students.[100] When tests are used in decisions that have high-stakes consequences for students, evidence of mean score differences between relevant subgroups should be examined, where feasible. When mean differences are found between subgroups, investigations should be undertaken to determine that such differences are not attributable to construct underrepresentation or construct irrelevant error.[101] Evidence about differences in mean scores and the significance of the validity errors should also be considered when deciding which test to use.[102] In using test results for purposes other than selection, a test taker's score should not be accepted as a reflection of

---

The Comment to Standard 7.4 states, "Two issues are involved. The first deals with the inadvertent use of language that, unknown to the test developer, has a different meaning or connotation in one subgroup than in others. Test publishers often conduct sensitivity reviews of all test material to detect and remove sensitive material from the test. The second deals with settings in which sensitive material is essential for validity. For example, history tests may appropriately include material on slavery or Nazis. Tests on subjects from life sciences may appropriately include material on evolution. A test of understanding of an organization's sexual harassment policy may require employees to evaluate examples of potentially offensive behavior." *Joint Standards, supra* note 63, at p. 82.

[97] *See Standard 7.6 (n.82) in Joint Standards, supra* note 63, at p. 82.

[98] Standard 7.12 states, "The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process." *Joint Standards, supra* note 63, at p. 84.

[99] *See Standard 7.7 in Joint Standards, supra* note 63, at p. 82 (Standard 7.7).

[100] *See Standard 1.24 (n.82), 7.8, 7.9, and 7.10 (n.28) in Joint Standards, supra* note 63, at pp. 23, 83.

Standard 7.8 states, "When scores are disaggregated and publicly reported for groups identified by characteristics such as gender, ethnicity, age, language proficiency, or disability, cautionary statements should be included whenever credible research reports that test scores may not have comparable meaning across these different groups." *Joint Standards, supra* note 63, at p. 83.

Standard 7.9 states, "When tests or assessments are proposed for use as instruments of social, educational, or public policy, the test developers or users proposing the test should fully and accurately inform policymakers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use." *Joint Standards, supra* note 63, at p. 83.

[101] *See Standard 7.10 (n.28) in Joint Standards, supra* note 63, at p. 83.

standing on the intended constructs without consideration of alternative explanations for the test taker's performance.[103] Explanations might reflect limitations of the test, for instance construct irrelevant factors may have significantly impacted the student's score. Explanations may also reflect schooling factors external to the test, for instance lack of instructional opportunities.

The issue of feasibility is discussed in a few of the standards summarized above. In the comments associated with these standards, feasibility is generally addressed in terms of adequate sample size, with continued operational use of a test as a way of accumulating adequate numbers of subgroup results over administrations. When credible research reports that results differ in meaning across subgroups, collecting separate and parallel validity data verifies that the same knowledge and skills are being measured for all groups of test takers. Particularly in high-stakes situations, it is important that all feasibility decisions include the potential costs to students of using information where the validity of the scores has not been verified.[104]

2. Fairness in Reliability

Fairness in reliability focuses on making sure that scores are stable and consistently accurate for all groups of students. Two key standards address this issue. First, when there are reasons for expecting that test reliability analyses might differ substantially for different subpopulations, reliability data should be presented as soon as feasible for each major population for whom the test is recommended.[105] Second, "[w]hen significant variations are permitted in test administration procedures, separate reliability analyses should be provided for

[102] Standard 7.11 states, "When a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use." *Joint Standards*, *supra* note 63, at p. 83.

[103] *See Standard 7.5 (n.66) in Joint Standards*, *supra* note 63, at p. 82.

[104] The Comment to Standard 10.7 states, "In addition to modifying tests and test administration procedures for people who have disabilities, evidence of validity for inferences drawn from these tests is needed. *Validation is the only way to amass knowledge about the usefulness of modified tests for people with disabilities. The costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities.* This standard is feasible in the limited circumstances where a sufficient number of individuals with the same level or degree of a given disability is available." *Joint Standards*, *supra* note 63, at p. 107 (italics added).

[105] Standard 2.11 states, "If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended." *Joint Standards*, *supra* note 63, at p. 34. It should be noted that reliability estimates may differ simply because of limited variance within a group. This is not a flaw in the test leading to unfairness, but rather a function of the statistical methodologies used in calculating the estimates.

scores produced under each major variation if adequate sample sizes are available."[106] Often, continued operational use of a test is a way to accumulate an adequate sample size over administrations.

D. Cutscores

The same principles regarding fairness, validity, and reliability apply generally to the establishment and use of cutscores for the purpose of making high-stakes educational decisions. Cutscores, also known as cut points or cutoff scores, are specific points on the test or scale where test results are used to divide levels of knowledge, skill, or ability. A cutscore may divide the demonstration of acceptable and unacceptable skills, as in placement in gifted and talented programs where students are accepted or rejected. There may be multiple cutscores that identify qualitatively distinct levels of performance. Cutscores are used in a variety of contexts, including decisions for placement purposes or for other specific outcomes, such as graduation, promotion, or admissions.[107]

Many of the concepts regarding test validity apply to cutscores—that is, the cut points themselves, like all scores, must be accurate representations of the knowledge and skills of students.[108] Further, "[w]hen feasible, cutscores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria."[109] Validity evidence should generally be able to demonstrate that students above the cut point represent or demonstrate a qualitatively greater degree or different type of skills and knowledge than those below the cut point, whenever these types of inferences are made. In high-stakes situations, it is important to examine the validity of

> Where the results of the [cutscore] setting process have highly significant consequences,...those responsible for establishing cutscores should be concerned that the process...[is] clearly documented and defensible.
>
> *Joint Standards*, Introduction to Chapter 4, p. 54.

[106] *See Standard 2.18 in Joint Standards*, *supra* note 63, at p. 36.

[107] In order to promote readability, the discussion on cutscores presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. Readers are encouraged to review Chapter 4, Scales, Norms, and Score Comparability, in the *Joint Standards* for additional, relevant information about cutscores. *See Joint Standards*, *supra* note 63, at pp. 53-54; *see also Standard 1.19 and 13.9 (n.19) in Joint Standards*, *supra* note 63, at pp. 22, 147.

Standard 1.19 states, "If a test is recommended for use in assigning persons to alternative treatments or is likely to be so used, and if outcomes from those treatments can reasonably be compared on a common criterion, then, whenever feasible, supporting evidence of differential outcomes should be provided." *Joint Standards*, *supra* note 63, at p. 22.

[108] *See Joint Standards*, *supra* note 63, pp. 9-16 (Chapter 1, Validity, discusses that the interpretation of *all* scores should be an accurate representation of what is being measured).

[109] *See Standard 4.20 in Joint Standards*, *supra* note 63, at p. 60.

the inferences that underlie the specific decisions being made on the basis of the cutscores. In other words, what must be validated is the specific use of the test based on how the scores of students above and below the cutscore are being interpreted.

Reliability of the cutscores is also important. The *Joint Standards* state that where cutscores are specified for selection or placement, the degree of measurement error around each cutscore should be reported.[110] Evidence should also indicate the misclassification rates, or percentage of error in classifying students, that are likely to occur among students with comparable knowledge and skills.[111] This information should be available by group as soon as feasible if there is a prior probability that the misclassification rates may differ substantially by group.[112] Misclassification of students above or below the cutpoints can result in both false positive and false negative classifications, respectively. As an example of false negative misclassifaction one might ask, what percentage of students who should be allowed to graduate would not be allowed to do so because of error due to the test rather than differences in their actual knowledge and skills?[113] The *Joint Standards* state, "Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories."[114]

There is no single right answer to the questions of when, where and how cutscores should be set on a test with high-stakes consequences for students.[115] Some experts suggest, however, that multiple methods of determining cutscores should be used when determining a final cutscore.[116] Further, the reasonableness of the standard setting process and the consequences for students should be clearly and specifically documented for a given use.[117] Both the *Joint Standards* and *High Stakes* repeatedly

---

[110] Standard 2.14 states, "Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cutscores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score." *Joint Standards*, *supra* note 63, at p. 35.

[111] "Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mismeasured without serious consequences. Mismeasurment of examinees whose true scores are close to the cut score is a more serious concern. . . . The term *classification consistency* or *inter-rater agreement*, rather than *reliability*, would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee's score to the cut score." *Joint Standards*, *supra* note 63, at p. 30.

[112] *See Standard 2.11 (n.105) in Joint Standards*, *supra* note 63, at p. 34.

[113] *See Joint Standards at p. 30.*

[114] *Joint Standards*, *supra* note 63, at p. 59.

[115] *See High Stakes*, *supra* note 61, at p. 168.

[116] *See High Stakes*, *supra* note 61, at p.169.

[117] *See Standard 4.19 and 4.21 and their comments in Joint Standards*, *supra* note 63, at pp. 59-60; *see also High Stakes*, *supra* note 61, at pp. 89-187 (Chapters 5, 6, and 7).

state that decisions should not be made solely or automatically on the basis of a single test score, and that other relevant information should be taken into account if it will enhance the overall validity of the decision.[118]

Standard 4.19 states, "When proposed score interpretations involve one or more cutscores, the rationale and procedures used for establishing cutscores should be clearly documented." *Joint Standards*, *supra* note 63, at p. 59.

Standard 4.21 states, "When cutscores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way." *Joint Standards*, *supra* note 63, at p. 60.

[118] *See High Stakes*, *supra* note 61, at pp. 89-187 (Chapters 5, 6, and 7); *Standard 13.7 (n.8) in Joint Standards*, *supra* note 63, at p. 146.

Test Measurement Principles: Questions about Appropriate Test Use

In order to determine if a test is being used appropriately to make high-stakes decisions about students, considerations about the context of the test use need to be addressed, as well as the validity, reliability, and fairness of the score interpretations from the current test being proposed.

1. What is the purpose for which the test is being used?
2. What information, besides the test, is being collected to inform this purpose?
3. What are the particular inferences that, if true, would support the use of the test to accomplish this purpose?
4. Based on how the test results are to be used, is there adequate evidence to document the validity of the inferences for students taking the test? That is,
   - ⇑ Does the evidence support the inferences that the test accurately reflects the specific knowledge and skills the test says it measures?
   - ⇑ Does the evidence support the inferences that the test scores are valid for the stated purpose, and in the particular type of situation where the test is to be administered?
   - ⇑ Does the evidence support the inferences that the test scores are valid for the specific groups of students who are taking the test?
5. Is there adequate evidence of reliability of the test scores for the proposed use?
6. Is there adequate evidence of fairness in validity and reliability to document that the test score inferences are accurate and meaningful for all groups of students taking the test? That is,
   - ⇑ Does the evidence support the inference that the test is measuring the same constructs for all groups of students?
   - ⇑ Does the evidence support that the scores do not systematically underestimate or overestimate the knowledge or skills of members of any particular group?
7. Is there adequate evidence that cutscores have been properly established and that they will be used in ways that will provide accurate and meaningful information for all test takers?

II. The Testing of All Students: Issues of Intervention and Inclusion

All aspects of validity, reliability, fairness, and cutscores discussed above are applicable to the measurement of knowledge and skills of all students, including limited English proficient students[119] and students with disabilities. This section addresses additional issues related to accurately measuring the knowledge and skills of these two distinct populations in selected situations.

Whenever tests are intended to evaluate the knowledge of skills of different groups of students, ensuring that test score inferences accurately reflect the intended constructs for all students is a complex task. It involves several aspects of test construction, pilot testing, implementation, analysis, and reporting. For limited English proficient students and students with disabilities, the appropriate inclusion of students from these groups in validation and other technical samples, and the meaningful inclusion of relevant limited English proficient and disability experts throughout the test development process, are necessary to ensure suitable test quality for these groups of test takers.

The proper inclusion of diverse groups of students in the same academic achievement testing program helps to ensure that high-stakes decisions are made on the basis of tests results that are as comparable as possible across all groups of test takers, rather than on the basis of results from different tests developed to measure different content domains.[120] The appropriate inclusion of students can also help to ensure that educational benefits attributable to the high-stakes decisions will be available to all. In some cases, it is appropriate to test limited English proficient students and students with disabilities under standardized conditions, as long as the evidence supports the validity of the results in a given situation for these students. In other cases, the conditions may have to be accommodated to assure that the inferences of the scores validly reflect the students' mastery of the intended constructs.[121] The use of multiple measures generally enhances the accuracy of the educational decisions, and these measures can be used to confirm the validity of the test results. The use of multiple measures is particularly relevant for limited English proficient students and students with disabilities, in cases where technical data are being collected on the proper use of accommodations and the proper interpretation of test results when testing conditions are accommodated.

[119] These are students who are learning English as a second language. Other documents sometimes refer to these students as English language learners.

[120] *See High Stakes*, *supra* note 61, at pp. 7, 80.

[121] *See Joint Standards*, *supra* note 63 at pp. 71-80, 91-97, 101-106 (Chapters 7, 9, and 10).

A. General Considerations about Accommodations

Making similar inferences about scores from academic achievement tests for all test takers, and making appropriate decisions when using these scores, requires measuring the same academic constructs (knowledge and skills in specific subject areas) across groups and contexts. In measuring the knowledge and skills of limited English proficient students and students with disabilities, it is particularly important that the tests actually measure the intended knowledge and skills and not factors that are extraneous to the intended construct.[122] For instance, impaired visual capacity may influence a student's test score in science when the student must sight read a typical paper and pencil science test. In measuring science skills, the student's sight is likely not relevant to her knowledge of science. Similarly, how well a limited English proficient student reads English may influence the student's test score in mathematics when the student must read the test. In this case, the student's reading skills are not relevant when the skills of mathematics computation are to be measured. When accommodations are necessary, determining proper accommodations that will produce valid information about the same intended construct for individual students is challenging for both test users and test developers. Furthermore, collecting adequate evidence to document that the same inferences are appropriate under accommodated and nonaccommodated conditions is a difficult task.

> Standard 10.1
>
> In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Typically, accommodations to established conditions are found in three main phases of testing: 1) the administration of tests, 2) how students are allowed to respond to the items, and 3) the presentation of the tests (how the items are presented to the students on the test instrument). Administration accommodations involve setting and timing, and can include extended time to counteract the increased literacy demands for English language learners, or fatigue for a student with sensory disabilities. Response accommodations allow students to demonstrate what they know in different ways, e.g. responding on a computer rather than in a test booklet. Presentation accommodations can include format variations such as fewer items per page, large print, and plain language editing procedures, which use short sentences, common words, and active voice. There is a wide variation in

[122] This is known as construct irrelevance. *See* discussion *infra* Chapter 1 Part (I)(A)(3) (Sources of Validity Error); *Joint Standards*, *supra* note 63, at pp. 173-174.

which accommodations are used across states and school districts. (Appendix C lists many of the accommodations used in large-scale testing for limited English proficient students and students with disabilities. The list is not meant to be exhaustive, and its use in this document should not be seen as an endorsement of any specific accommodations. Rather, the Appendix is meant to provide examples of the types of accommodations that are being used with limited English proficient students and students with disabilities.)

Issues regarding the use of accommodations are complex. When the possible use of an accommodation for a student is being considered, two questions should be examined: 1) What is being measured if conditions are accommodated? 2) What is being measured if the conditions remain the same? The decision to use an accommodation or not should be grounded in the ultimate goal of collecting test information that accurately and fairly represents the knowledge and skills of the individual student on the intended constructs. The overarching concern should be that test score inferences accurately reflect the intended constructs rather than factors extraneous to the intent of the measurement.[123]

B. Testing of Limited English Proficient Students

The *Joint Standards* and several recent measurement publications discuss the population of limited English proficient students and how test publishers and users have handled inclusion in tests to date.[124] This section briefly outlines principles derived from the *Joint Standards* and these publications. It addresses two types of testing situations especially relevant for limited English proficient students: the assessment of English language proficiency and the assessment of academic educational achievement.

---

[123] *See Standard 9.1 and 10.1 in Joint Standards, supra* note 63, at pp. 97, 106; Messick, *Validity, supra* note 78.

Standard 9.1 states, "Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences." *Joint Standards, supra* note 63, at p. 97.

Standard 10.1 states, "In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement." *Joint Standards, supra* note 63, at p. 106.

[124] *See, e.g., Joint Standards, supra* note 63, at pp. 91-97 (Chapter 9); *High Stakes, supra* note 61, at pp. 211-237 (Chapter 9); National Research Council, *Improving America's Schooling for Language Minority Children: A Research Agenda* (Diane August & Kenji Hakuta eds., 1997) (hereinafter *Improving America's Schooling for Language Minority Children*); Rebecca J. Kopriva, Council of Chief State School Officers, *Ensuring Accuracy in Testing for English Language Learners* (2000) (hereinafter Kopriva, *Ensuring Accuracy in Testing*).

1. Assessing English Language Proficiency

Issues of validity, reliability, and fairness apply to tests and other relevant assessments that measure English language proficiency. English language proficiency is typically defined as proficiency in reading, writing, speaking, and understanding English.[125] Assessments that measure English language proficiency are generally used to make decisions about who should receive English language acquisition services, the type of programs in which these students are placed, and the progress of students in the appropriate programs. They are also used to evaluate the English proficiency of students when exiting from services, to ensure that they can successfully participate in the regular school curriculum. In making decisions about which tests are appropriate, it is particularly important to make sure that the tests accurately and completely reflect the intended English language proficiency constructs so that the students are not misclassified. It is generally accepted that an evaluation of a range of communicative abilities will typically need to be assessed when placement decisions are being made.[126]

> Standard 9.10
>
> Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not on a single linguistic skill.

2. Assessing the Academic Educational Achievement of Limited English Proficient Students

Several factors typically affect how well the educational achievement of limited English proficient students is measured on standardized academic achievement tests. For all test takers, any test that employs written or oral skills in English or in another language is, in part, a measure of those skills in the particular language. Test use with individuals who have not sufficiently acquired the literacy or linguistic skills in the language of the test may introduce construct-irrelevant components to the testing process. Further, issues related to differences in the experiences of students may substantially affect test results. In both instances, test scores may not accurately reflect the qualities and competencies that the test intends to measure.[127] While it is very important that the test score inferences are valid, reliable, and fair, the technical issues associated with developing meaningful achievement tests for limited English proficient students are recognizably challenging and difficult to address.

---

[125] *See Improving America's Schooling for Language Minority Children, supra* note 124, at pp. 116-118.

[126] *See Standard 9.10 and comment in Joint Standards, supra* note 63, at pp. 99-100.

[127] *See Joint Standards, supra* note 63, at p. 91.

a. Background Factors for Limited English Proficient Students

The background factors particularly salient in ensuring accuracy in testing for students with limited English proficiency tend to relate to literacy, culture, and schooling.[128]

Limited English proficient students often bring varying levels of English and home-language literacy skills to the testing situation.[129] These students may be adept in conversing orally in their home language, but unless they have had formal schooling in their home language, they may not have a corresponding level of literacy. Also, while students with limited English proficiency may acquire a degree of oral proficiency in English, literacy in English for many students comes later.[130] To add to the complexity, oral and literacy proficiency in either the home language or English involves both social and academic components. Thus, a student may be able to write a well-organized social letter in his or her home language, and may not be able to orally explain adequately in that language how to solve a mathematics problem that includes the knowledge of concepts and words endemic to the field of mathematics. The same

Factors Related to Accurately Testing LEP Students

Literacy Issues
- The student's level of oral and written proficiency in English
- The student's literacy in his or her home language
- The language of instruction

Cultural Issues
- Background experiences
- Perceptions of prior experiences
- Value systems

Schooling Issues
- The amount of formal elementary and secondary schooling in the student's home country and in U.S. schools
- Consistency of schooling
- Instructional practices in the classroom

[128] *See Improving Schooling for Language Minority Children, supra* note 124, at Chapter 5; Kopriva, *Ensuring Accuracy in Testing, supra* note 124, at Chapter 1.

[129] *See Joint Standards*, Chapter 9, p. 91-100; Kopriva, *Ensuring Accuracy in Testing supra note 124,at* Chapter 1.

[130] *See* National Research Council, *Testing, Teaching and Learning*, p. 61 (Richard F. Elmore & Robert Rothman eds., 1999).

[131] *See Improving America's Schooling for Language Minority Children, supra* note 124, at pp. 113-137.

phenomena may occur in English as well.[131]

Therefore, in determining how to effectively measure the academic knowledge and skills of this population, educators and policymakers should consider how to minimize the influence of literacy issues, except when these constructs are explicitly being measured. Considering the level of linguistic and literacy proficiencies of limited English proficient students in their home language and in English will often affect which achievement tests are appropriate for these students, and which accommodations to standardized testing conditions, if any, might be most useful for which students.[132]

Additionally, diverse cultural and other background experiences, including variations in amount, type and location (home country and United States) of formal elementary and secondary schooling, as well as interrupted and multi-location schooling of students (of the type frequently experienced by children of migrant workers), affect language literacy, the contextual content of items, and the academic foundational knowledge base that can be assumed in appropriately interpreting the results of educational achievement tests. The format and procedures involved in testing can also affect accuracy in test scores, particularly if the test practices differ substantially from ongoing instructional practices in classrooms, including which accommodations are used in the classroom and how they are used.[133]

b. Including Limited English Proficient Students in Large-Scale Standardized Achievement Tests

The *Joint Standards* recognize the complexity of developing educational achievement tests that are appropriate for a range of test takers, including those who are English language learners. Overall, "testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences."[134] When credible research evidence reports that scores may differ in meaning across subgroups of linguistically diverse test takers, then, to the extent feasible, the same form of validity evidence should be collected for each subgroup as for the examinee population as a whole.[135] The *Joint Standards* state, "When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and

[132] *See Improving America's Schooling for Language Minority Children, supra* note 124, at pp. 113-137.

[133] *See* Kopriva, *Ensuring Accuracy in Testing, supra* note 124, at pp. 29-48, 61-70, 95-98.

[134] *See Standard 9.1 in Joint Standards, supra* note 63, at p. 97.

[135] Standard 9.2 states, "When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole." *Joint Standards, supra* note 63, at p. 97.

interpretation."[136] Furthermore, "When testing an examinee proficient in two or more languages for which the test is available, the examinee's relative language proficiencies should be determined. The test generally should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment."[137] Recommended accommodations should be used appropriately and described in detail in the test manual;[138] translation methods and interpreter expertise should be clearly described;[139] evidence of test comparability should be reported when multiple language versions of a test are intended to be comparable;[140] and evidence of the score reliability and the validity of the translated test's score inferences should be provided for the intended uses and linguistic groups.[141]

Providing accommodations to established testing conditions for some students with limited English proficiency may be appropriate when their use would yield the most valid scores on the intended academic achievement constructs. Deciding which accommodations to use for which students usually involves an understanding of which construct irrelevant background factors would substantially influence the measurement of intended knowledge and skills for individual students, and how the accommodations

---

[136] *See Standard 9.6 in Joint Standards, supra* note 63, at p. 99.

[137] *See Standard 9.3 in Joint Standards, supra* note 63, at p. 98.

[138] *See Standard 9.4 and 9.5 in Joint Standards, supra* note 63, at p. 98.

Standard 9.4 states, "Linguistic modifications recommended by test publishers, as well as the rationale for the modifications, should be described in detail in the test manual." *Joint Standards, supra* note 63, at p. 98.

Standard 9.5 states, "When there is credible evidence of score comparability across regular and modified tests or administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores." *Joint Standards, supra* note 63, at p 98.

[139] *See Standard 9.7 and 9.11 in Joint Standards, supra* note 63, at pp. 99-100.

Standard 9.7 states, "When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested." *Joint Standards, supra* note 63, at p. 99.

Standard 9.11 states, "When an interpretation is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process." *Joint Standards, supra* note 63, at p. 100.

[140] Standard 9.9 states "When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability." *Joint Standards, supra* note 63, at p. 99.

[141] *See Standard 9.7 (n.139) and comment in Joint Standards, supra* note 63, at p. 99.

The Comment to Standard 9.7 states "[f]or example, if a test is translated into Spanish for use with Mexican, Puerto Rican, Cuban, Central American, and Spanish populations, score reliability and the validity of the test score inferences should be established with members of each of these groups separately where feasible. In addition, the test translation methods used need to be described in detail." *Joint Standards, supra* note 63, at p. 99.

would impact the validity of the test score interpretations for these students.[142] In collecting evidence to support the technical quality of a test for limited English proficient students, the accumulation of data may need to occur over several test administrations to ensure sufficient sample sizes. Educators and policymakers need to understand that the proper use of accommodations for limited English proficient students and the determination of technical quality are complex and challenging endeavors.

Appendix C lists various test presentation, administration, and response accommodations that states and districts generally employ when testing limited English proficient students. Examples of accommodations in the presentation of the test include editing text so the items are in plain language, or providing page formats which minimize confusion by limiting use of columns and the number of items per page. Presenting the test in the student's native language is an accommodation to a test written in English when the same constructs are being measured on both the English and native-language versions. It is essential that translations accurately convey the meaning of the test items; poor translations will often prove more harmful than helpful.[143] Administration accommodations include extending the length of the testing period, permitting breaks, administering tests in small groups or in separate rooms, and allowing English or native-language glossaries or dictionaries as appropriate. Response accommodations include oral response and permitting students to respond in their native language.

C. Testing of Students with Disabilities

The *Joint Standards* and several recent measurement publications discuss the population of students with disabilities and how test publishers and users have handled inclusion in tests to date.[144] This section briefly outlines principles derived from the *Joint Standards* and these publications. It addresses three types of testing situations especially relevant for students with disabilities: tests used for diagnostic and intervention purposes, the assessment of academic educational achievement, and alternate assessments for elementary and secondary school students with disabilities who cannot participate in districtwide academic achievement tests.

---

[142] *See* Kopriva, *Ensuring Accuracy in Testing*, *supra* note 124, at pp. 49-66, 71-76 (discussing which accommodations might be most beneficial for students with various background factors).

[143] *See* President's Advisory Commission on Educational Excellence for Hispanic Americans, *Testing Hispanic Students in the United States: Technical and Policy Issues, Executive Summary*, p. 8 (2000).

[144] *See, e.g.*, *Joint Standards*, *supra* note 63, at pp. 101-106 (Chapter 10); *High Stakes*, *supra* note 61, at pp. 188-210 (Chapter 8); National Research Council, *Educating One and All: Students with Disabilities and Standards-Based Reform* (Lorraine M. McDonnell, Margaret J. McLaughlin & Patricia Morison eds., 1997) (hereinafter *Educating One and All*); Martha Thurlow, Judy Elliott & Jim Ysseldyke, *Testing Students with Disabilities* (1998) (hereinafter *Testing Students with Disabilities*).

1. Tests Used for Diagnostic and Intervention Purposes

All issues of validity, reliability, and fairness apply to tests and other assessments used to make diagnostic and intervention decisions for students with disabilities. Tests that yield diagnostic information typically focus in great detail on identifying the specific challenges and strengths of a student.[145] These diagnostic tests are often administered in one-to-one situations (test taker and examiner) rather than in a group situation. In many cases, they have been designed with standardized adaptations to fit the needs of individual examinees. In making decisions about which tests are appropriate to use, it is important to make sure that the tests accurately and completely reflect the intended constructs, so that the interventions are appropriate and beneficial for the individual students. Proper analyses should be conducted to yield correct interpretations of results when differential prediction for different groups is likely.[146]

> Standard 10.12
>
> In testing individuals with disabilities for diagnostic and intervention purposes, the test should not be used as the sole indicator of the test taker's functioning. Instead, multiple sources of information should be used.

2. Assessing the Academic Educational Achievement of Students with Disabilities

Several factors affect how well the educational achievement of students with disabilities is measured on standardized academic achievement tests. Test scores should accurately measure the students' knowledge and skills in academic achievement rather than factors irrelevant to the intended constructs of the test.[147] While it is very important that the test score inferences be valid, reliable, and fair, the technical issues associated with developing meaningful achievement tests for students with disabilities are recognizably complex and difficult to accomplish. Under federal law, students with disabilities must be included in statewide or districtwide assessment programs and provided with appropriate accommodations if necessary. Guidance about testing elementary and secondary school students with disabilities is addressed by the individualized education program (IEP) process or other applicable evaluation procedures. The IEP or Section 504 plan addresses how students should be tested, and identifies testing

[145] *See Joint Standards, supra* note 63, at pp. 101-106, 119-145 (Chapters 10, 12, and 13); *High Stakes, supra* note 61, at pp. 13-28 (Chapter 1).

[146] *See Standard 7.6 (n.82) in Joint Standards, supra* note 63, at p. 82.

[147] *See Standards 10.1 (n.123) in Joint Standards, supra* note 63, at p. 106.

accommodations that would be appropriate for individual students. The IDEA also requires state or local education agencies to develop guidelines for the relatively small number of students with disabilities who cannot take part in statewide or districtwide tests to participate in alternate assessments.

a. Background Factors for Students with Disabilities

The background factors particularly important to students with disabilities are generally related to the nature of the disabilities or to the schooling experiences of these students.[148]
Within any disability category, the type, number, and severity of impairments vary

Factors Related to Accurately Testing Students with Disabilities

Disability Issues
- Types of impairments
- Severity of impairments

Schooling Experiences
- Overlap of individualized educational goals and general education curricula in elementary and secondary schooling
- Pace of schooling
- Instructional practices in the classroom

greatly.[149] For instance, some students with learning disabilities have a processing disability in only one subject, such as mathematics, while others experience accessing, retrieval, and processing impairments that affect a broad number of school subjects and contexts. For many of these students, one or more of the impairments may be relatively mild, while for others one or more can be significant. Further, different types of disabilities yield significantly different constellations of issues. For instance, the considerations surrounding students with hearing impairment or deafness may overlap significantly with limited English proficient students in some ways and with other students with disabilities in other respects. For example, the *Joint Standards* discuss provisions regarding the testing and validation of tests for English language learners that apply to students who have hearing impairments or deafness, as well.[150] This complexity poses a challenge not only to educators, but also to test administrators and developers. In general, in determining how to use academic tests appropriately for

[148] *See Educating One and All, supra* note 144, at Chapter 3; *Testing Students with Disabilities, supra* note 144.

[149] *See Joint Standards, supra* note 63, at pp. 101-105 (Chapter 10); *Testing Students with Disabilities, supra* note 144.

[150] *See Standard 9.2 (n.135) and 9.10 (n.126) in Joint Standards, supra* note 63, at pp. 97, 99-100.

students with disabilities, educators and policymakers should consider how to minimize the influence of the impairments in measuring the intended constructs.

Educating One and All explains that the schooling experiences of students with disabilities vary greatly as a function of their disability, the severity of impairments, and expectations of their capabilities.[151] Two sets of educational experiences, in particular, affect how educators and policymakers accommodate tests and use them appropriately for this population. First, the individualized education program (IEP) teams identify individual educational plans for students with disabilities that have different degrees of overlap with the general education curricula. This alignment will affect what opportunities students with disabilities will have to master the material being tested on the schoolwide academic achievement tests. Second, the IEP team also recommends appropriate accommodations for students, and these accommodations are usually consistent with classroom accommodation techniques. However, while special educators have a long history of accommodating instruction and evaluation to fit student strengths, not all the instructional or testing practices in the classroom are appropriate in large-scale testing. Additionally, some students may not have been exposed routinely to the types of accommodations that would be possible in large-scale testing.[152]

b. Including Students with Disabilities in Large-Scale Standardized Achievement Tests

The *Joint Standards* recognize the complexity of developing educational achievement tests that are appropriate for a range of test takers, including students with disabilities. The interpretation of the scores of students with disabilities should accurately and fairly reflect the academic knowledge, skills, or abilities that the test intends to measure. The interpretation should not be confounded by those challenges students face that are extraneous to the intent of the measurement.[153] Rather, validity evidence should document that the inferences of the scores of students with disabilities are accurate. Pilot testing and other technical investigations should be conducted where feasible to

[151] *See Educating One and All, supra* note 144, at Chapter 3.

[152] *See Educating One and All, supra* note 144, at Chapter 5.

[153] *See Standard 10.1 (n.123) and 10.10 in Joint Standards, supra* note 63, at pp. 106, 107-108.

Standard 10.10 states, "Any test modifications adopted should be appropriate for the individual test taker, while maintaining all feasible standardized features. A test professional needs to consider reasonably available information about each test taker's experiences, characteristics, and capabilities that might impact test performance, and document the grounds for the modification." *Joint Standards, supra* note 63, at pp. 107-108.

ensure the validity of the test inferences when accommodations have been allowed.[154] Feasibility is always a consideration, although the *Joint Standards* comment, "the costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities."[155]

Providing accommodations to established testing conditions for some students with disabilities may be appropriate when their use would yield the most valid scores on the intended academic achievement constructs. Deciding which accommodations to use for which students usually involves an understanding of which construct irrelevant background factors would substantially influence the measurement of intended knowledge and skills for individual students, and how the accommodations would impact the validity of the test score interpretations for these students.[156] In collecting

[154] Several standards discuss the appropriate types of validity evidence, including Standards 10.3, 10.5, 10.6, 10.7, 10.8, and 10.11. Because of the low-incidence nature of several of the disability groups, such as hearing loss, vision loss, or concomitant hearing and vision loss, especially when different severity levels and combinations of impairments are considered, this type of evidence will probably need to be accumulated over time in order to have a large enough sample size.

Standard 10.3 states, "Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications." *Joint Standards*, *supra* note 63, at p. 106.

Standard 10.5 states, "Technical material and manuals that accompany modified tests should include a careful statement of the steps taken to modify the test to alert users to changes that are likely to alter the validity of inferences drawn from the test scores." *Joint Standards*, *supra* note 63, at p. 106.

Standard 10.6 states, "If a test developer recommends specific time limits for people with disabilities, empirical procedures should be used, whenever possible, to establish time limits for modified forms of timed tests rather than simply allowing test takers with disabilities a multiple of the standard time. When possible, fatigue should be investigated as a potentially important factor when time limits are extended." *Joint Standards*, *supra* note 63, at p. 107.

Standard 10.7 states, "When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them." *Joint Standards*, *supra* note 63, at p. 107.

Standard 10.8 states, "Those responsible for decisions about test use with potential test takers who may need or may request specific accommodations should (a) possess the information necessary to make an appropriate selection of measures, (b) have current information regarding the availability of modified forms of the test in question, (c) inform individuals, when appropriate, about the existence of modified forms, and (d) make these forms available to test takers when appropriate and feasible." *Joint Standards*, *supra* note 63, at p. 107.

Standard 10.11 states, "When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores." *Joint Standards*, *supra* note 63, at p. 108.

[155] *See comment to Standard 10.7 in Joint Standards*, *supra* note 63, at p. 106.

[156] *See Testing Students with Disabilities*, *supra* note 144, for a discussion of which accommodations might be most beneficial for students with various impairments and other background factors.

evidence to support the technical quality of the test results for students with disabilities, the accumulation of data may need to occur over several administrations to ensure sufficient sample sizes. Educators and policymakers need to understand that the proper use of accommodations for students with disabilities and the determination of technical quality are complex and challenging endeavors.

Appendix C lists various presentation, administration, and response accommodations that states and districts generally employ when testing students with disabilities. Examples of presentation accommodations are the use of Braille, large print, oral reading, or providing page formats that minimize confusion by limiting use of columns and the number of items per page. Administration accommodations in setting include allowing students to take the test at home or in a small group, and accommodations in timing include extended time and frequent breaks. Variations in response formats include allowing students to respond orally, point or use a computer.

3. Alternate Assessments

Alternate assessments are assessments for those elementary and secondary school students with disabilities who cannot participate in state or districtwide standardized assessments, even with the use of appropriate accommodations and modifications.[157] For the constructs being measured, the considerations with respect to validity, reliability, and fairness apply to alternate assessments, as well. Appropriate content needs to be identified, and procedures designed to ensure technical rigor need to be followed.[158] In addition, strong evidence should show that the test measures the knowledge and skills it intends to measure, and that the measurement is a valid reflection of mastery in a range of contextual situations.

---

[157] The IDEA requires use of alternate assessments in certain areas. *See* 34 C.F.R. § 300.138. These assessments may or may not be used in decisions that have high-stakes consequences for students.

[158] *See Educating One and All*, *supra* note 144, at Chapter 5, and *Testing Students with Disabilities*, *supra* note 144, for a discussion of the issues and processes involved in developing and implementing alternate assessments.

CHAPTER 2. Legal Principles

It is important for educators and policymakers to understand the test measurement principles and the legal principles that will enable them to ask informed questions and make sound decisions regarding the use of tests for high-stakes purposes. The goal of this chapter is to explain the legal principles that apply to educational testing.

The primary focus of this chapter is four federal nondiscrimination laws, enacted by Congress, and their implementing regulations: Title VI of the Civil Rights Act of 1964 (Title VI), Title IX of the Education Amendments of 1972 (Title IX), Section 504 of the Rehabilitation Act of 1973 (Section 504), and Title II of the Americans with Disabilities Act of 1990 (Title II).[159] Within the U.S. Department of Education, the Office for Civil Rights has responsibility for enforcing the requirements of these four statutes and their implementing regulations. Although the Office for Civil Rights does not enforce federal constitutional provisions, an overview of these constitutional principles, including under the Fifth and Fourteenth Amendments of the U.S. Constitution, has also been included for informational purposes because of their importance to sound test use. The discussion of legal principles in this chapter is intended to reflect existing legal principles and does not establish new requirements.[160]

[159] Title VI prohibits discrimination on the basis of race, color and national origin in the programs and activities of recipients that receive federal financial assistance. The U.S. Department of Education's regulation implementing Title VI is found at 34 C.F.R. Part 100. Title IX prohibits discrimination on the basis of sex in educational programs and activities of recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title IX is found at 34 C.F.R. Part 106. Section 504 prohibits discrimination on the basis of disability in the programs and activities of recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Section 504 is found at 34 C.F.R. Part 104. Title II prohibits discrimination on the basis of disability by public entities, regardless of whether they receive federal funding. The U.S. Department of Justice's regulation implementing Title II is found at 28 C.F.R. Part 35.

[160] Consistent with this approach, court decisions are not cited if the case is still on appeal or the time to request an appeal has not ended.

Some of the issues that have been considered by federal courts in assessing the legality of specific testing practices for making high-stakes decisions include:

- The use of an educational test for a purpose for which the test was not designed or validated[1]
- The use of a test score as the sole criterion for the educational decision[2]
- The nature and quality of the opportunity provided to students to master required content, including whether classroom instruction includes the material covered by a test administered to determine student achievement[3]
- The significance of any fairness problems identified, including evidence of differential prediction of a criterion and possible cultural biases in the test or in test items[4]

The educational basis for establishing passing or cut-off scores[5]

---

[161] *See Sharif v. New York State Educ. Dep't.*, 709 F. Supp. 345, 354-55, 364 (S.D.N.Y. 1989) (in granting a motion for preliminary injunction, where girls received comparatively lower scores than boys, court found that the state's use of SAT scores as the sole basis for decisions awarding college scholarships intended to reward high school achievement was not educationally justified for this purpose in that the SAT had been designed as an aptitude test to predict college success and was not designed or validated to measure past high school achievement).

[162] *See United States v. Fordice*, 505 U.S. 717, 735-39 (1992) (Recognizing that "[a]nother constitutionally problematic aspect of the state's use of the ACT test scores is its policy of denying automatic admission if an applicant fails to earn the minimum ACT score specified for the particular institution, without also resorting to the appicant's high school grades as an additional factor in predicting college performance," the United States Supreme Court, taking into account evidence that the American College Testing Program discourages reliance on ACT scores as a sole admissions criterion, rejected the state's reliance on ACT scores as the sole criterion for denying applicants admission to historically segregated colleges, where the ACT requirement was originally adopted for discriminatory purposes, the current requirement is traceable to that decision and continues to have segregative effects, and the state failed to show that the "ACT-only" admissions standard was not susceptible to elimination without eroding sound educational policy); *see also Sharif*, 709 F. Supp. at 364.

[163] *See Lau v. Nichols*, 414 U.S. 563, 566-69 (1974) (finding a violation of the Title VI regulations where limited English proficient students were taught only in English and not provided any special assistance needed to meet English language proficiency standards required by the state for a high school diploma); *see also Debra P. v. Turlington*, 644 F.2d 397, 406-08 (5th Cir. 1981) (holding that use of a graduation test that covered material that had not been taught in class would violate the due process and equal protection clauses and that, under the circumstances of the case, immediate use of the diploma sanction for test failure would punish black students for deficiencies created by an illegally segregated school system which had provided them with inferior physical structures, course offerings, instructional materials, and equipment).

[164] *See Larry P. v. Riles*, 793 F.2d 969, 980-81, 983 (9th Cir. 1984) (finding that IQ tests the state used had not been validated for use as the sole means for determining that black children should be placed in classes for educable mentally retarded students); *Sharif*, 709 F. Supp. at 354 (observing that the SAT under-predicts success for female college freshmen as compared with males); *see also Parents in Action on Special Educ. v. Hannon*, 506 F. Supp. 831, 836-37 (N.D. Ill. 1980) (court's analysis of items on I.Q. test found only minimal amount of cultural bias not resulting in erroneous mental retardation diagnoses given other information considered in process).

[165] *See Groves v. Alabama State Bd. of Educ.*, 776 F. Supp. 1518, 1530-31 (M.D. Ala. 1991) (finding test required for admission to undergraduate teacher training program would not be educationally justified if the passing score is not itself a valid measure of the minimal ability necessary to become a teacher); *Richardson v. Lamar County Bd. of Educ.*, 729 F. Supp. 806, 823-25 (M.D. Ala. 1989) (evidence revealed that cut off scores had not been set through a well-conceived, systematic process nor could the scores be characterized as reflecting the good faith

I. Discrimination Under Federal Statutes and Regulations

Congress has enacted four statutes prohibiting discrimination based on race, color, national origin, sex, and disability in elementary and secondary schools, colleges, and universities. Title VI prohibits discrimination based on race, color, or national origin; Title IX prohibits discrimination based on sex; and Section 504 and Title II of the Americans with Disabilities Act (ADA) prohibit discrimination based on disability. Title VI, Title IX, and Section 504 apply to all educational institutions that receive federal funds. Title II of the ADA applies to public entities, including public school districts and state colleges and universities.[166] The Title VI, Title IX, Section 504, and Title II statutes and their implementing regulations as well as the equal protection clause of the Fourteenth Amendment to the United States Constitution, prohibit intentional discrimination, based on race, national origin, sex, or disability.[167] In addition, the regulations that implement Title VI, Title IX, Section 504 and Title II prohibit policies or practices that have a discriminatory disparate impact on students based on their race, national origin, sex, or disability.[168]

---

exercise of professional judgment), *aff'd sub nom., Richardson v. Alabama State Bd. of Educ.*, 935 F.2d 1240 (11th Cir. 1991).

[166] OCR enforces five nondiscrimination statutes. Title VI of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000d, *et seq.* (2000); Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq.* (1999); Section 504 of the Rehabilitation Act of 1973, as amended, 29 U.S.C. § 794 (1999); Title II of the Americans with Disabilities Act of 1990, 42 U.S.C. §§, 12131, *et seq.* (1995 & Supp. 1999); and the Age Discrimination Act of 1975, as amended, 42 U.S.C. §§ 6101, *et. seq.* (1995 & Supp. 1999). Regulations issued by the United States Department of Education implementing Title VI, Title IX, and Section 504, respectively, can be found at 34 C.F.R. Part 100, 34 C.F.R. Part 106, and 34 C.F.R. Part 104. These regulations can be found on OCR's web site at *www.ed.gov/offices/OCR*. Regulations implementing Title II of the ADA can be found at 28 C.F.R. Part 35. Title III of the ADA, which is enforced by the U.S. Department of Justice, prohibits discrimination in public accommodations by private entities, including schools. Religious entities operated by religious organizations are exempt from Title III.

[167] The United States Supreme Court has held that "Title VI itself directly reached only instances of intentional discrimination . . . [but that] actions having an unjustifiable disparate impact on minorities could be redressed through agency regulations designed to implement the purposes of Title VI." *Alexander v. Choate*, 469 U.S. 287, 293 (1985) (discussing *Guardians Ass'n v. City Service Comm'n of N.Y.*, 463 U.S. 582 (1983)). The United States Supreme Court has never expressly ruled on whether Section 504, Title II and/or Title IX directly reach disparate impact discrimination. *See, e.g., Alexander*, 469 U.S. at 294 & n. 11 (noting the possibly different congressional purpose with respect to Section 504 as compared with Title VI). Section 504 and Title II require reasonable modifications where necessary to enable persons with disabilities to participate in or enjoy the benefits of public services. Nonetheless, the regulations implementing Section 504, Title II, and Title IX, like the Title VI regulations, explicitly prohibit actions having discriminatory effects as well as actions that are intentionally discriminatory.

[168] *See* 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. §§ 106.21(b)(2), 106.36(b), 106.52 (Title IX); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); and 28 C.F.R. § 35.130(b)(3) (Title II).

The authority of federal agencies to issue regulations with an "effects" standard has been consistently acknowledged by United States Supreme Court decisions and applied by lower federal courts addressing claims of discrimination in education. *See, e.g., Alexander*, 469 U.S. at 289-300 (1985); *Guardians Ass'n*, 463 U.S. at 584-93; *Lau*, 414 U.S. at 568; *see also* 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 of 1964* (July 14, 1994).

This section describes two central analytical frameworks for examining allegations of discrimination as set forth in federal nondiscrimination regulations: different treatment and disparate impact.[169] It also includes a further discussion of legal principles that apply specifically to students with limited English proficiency and to students with disabilities.

A. Different Treatment

Under federal law, policies and practices generally must be applied consistently to similarly situated individuals or groups, regardless of their race, national origin, sex, or disability.[170] For example, a federal court concluded that a school district had intentionally treated students differently on the basis of race where minority students whose test scores qualified them for two or more ability levels were more likely to be assigned to the lower-level class than similarly situated white students, and no explanatory reason was evident.[171]

In addition, educational systems that were previously segregated by race in violation of the Fourteenth Amendment and have not achieved unitary status have an obligation to dismantle their prior *de jure* segregation. In such instances, school districts are under "a 'heavy burden' of showing that actions that [have] increased or continued the effects of the dual system serve important and legitimate ends."[172] When a school district or other educational system uses a test or assessment procedure for a high-stakes purpose that has racially disparate effects, the school district can justify the test use only by showing that the test results are not due to the present effects of prior segregation or that the practice or procedure remedies the adverse effects of such segregation by offering better educational opportunities.[173]

---

[169] Intentional racial discrimination is a violation of both the Fourteenth Amendment to the United States Constitution and federal civil rights statutes in cases where evidence demonstrates that an action such as the use of a test for high-stakes purposes is motivated by an intent to discriminate. *See Elston v. Talladega County Bd. of Educ.*, 997 F.2d 1394, 1406 (11th Cir. 1993). As explained further in this section, the regulations promulgated under the federal civil rights statutes prohibit the use of neutral criteria having disparate effects unless the criteria are educationally justified. *See Guardians Ass'n*, 463 U.S. at 598.

[170] For example, under the Fourteenth Amendment and Title VI, different treatment based on race is permitted only when such action is narrowly tailored to further a compelling state interest. *See Adarand Constructors, Inc., v. Pena*, 515 U.S. 200 (1995); *Regents of the Univ. of Cal. v. Bakke*, 438 U.S. 265 (1978).

[171] *See People Who Care v. Rockford Bd. of Educ.*, 851 F. Supp. 905, 958-1001 (N.D. Ill. 1994), *remedial order rev'd, in part*, 111 F.3d 528 (7th Cir. 1997). On appeal, the Seventh Circuit Court of Appeals stated that the appropriate remedy based on the facts in the case was to require the district to use objective, non-racial criteria to assign students to classes, rather than abolishing the district's tracking system. 111 F.3d at 536.

[172] *Dayton Bd. of Educ. v. Brinkman*, 443 U.S. 526, 538 (1979) (quoting *Green v. County School Bd.*, 391 U.S. 430, 439 (1968)).

[173] *See Debra P.*, 644 F.2d at 397 ("[Defendants] failed to demonstrate either that the disproportionate failure [rate] of blacks was not due to the present effects of past intentional segregation or, that as presently used, the

B. Disparate Impact

Discrimination under federal law may also occur where the application of neutral criteria has discriminatory effects and those criteria are not educationally justified. Even if the criteria are educationally justified, discrimination may be found if there are alternative practices available that are equally effective in serving the educational institution's goals and that have less disparate impact. The federal nondiscrimination regulations provide that a recipient of federal funds may not "utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination."[174] It is important to understand that disparities in student performance based on race, national origin, sex, or disability, alone, do not constitute disparate impact discrimination under federal law. Furthermore, nothing in federal law guarantees equal results.

Courts applying the disparate impact test have examined three questions to determine if the practices at issue are discriminatory: (1) Does the practice or procedure in question result in substantial differences in the award of benefits or services based on race, national origin, or sex? (2) Is the practice or procedure educationally justified? and (3) Is there an equally effective alternative that can accomplish the institution's educational goal with less disparity?[175] (For a discussion of disability discrimination,

---

diploma sanction was necessary [in order] to remedy those effects."); *McNeal v. Tate County Sch. Dist.*, 508 F.2d 1017, 1020 (5th Cir. 1975) (ability grouping method that causes segregation may nonetheless be used "if the school district can demonstrate that its assignment method is not based on the present results of past segregation or that the method of assignment will remedy such effects through better educational opportunities"); *see also GI Forum v. Texas Educ. Agency*, 87 F. Supp. 2d 667, 674 (W.D. Tex. 2000) (inequalities in education did not cause disproportionate failure rate since all students have an equal opportunity to learn the items on the test, and the testing program, along with school accountability and remedial follow up, helps to address the effects of any prior discrimination and remaining disparities); *Fordice*, 505 U.S. at 731 ("If the State [university system] perpetuates policies and practices traceable to its prior system that continue to have segregative effects . . . and such policies are without sound educational justification and can be practically eliminated, the State has not satisfied its burden of proving that it has dismantled its prior system.").

[174] *See* 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); and 28 C.F.R. § 35.130(b)(3)(i) (Title II); *see also* 34 C.F.R. § 106.31 (Title IX). In *Guardians Association*, the U.S. Supreme Court upheld the use of the effects test, stating that the Title VI regulation forbids the use of federal funds, "not only in programs that intentionally discriminate on racial grounds but also in those endeavors that have a[n] [unjustified racially disproportionate] impact on racial minorities." *Guardians Ass'n*, 463 U.S. at 589-90.

[175] *See Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1417 (11th Cir. 1985); *see also Elston*, 997 F.2d at 1407 & n.14; *Larry P.*, 793 F.2d at 982 & n. 9; *Groves*, 776 F. Supp. at 1523-24, 1529-32; *Sharif*, 709 F. Supp. at 361. Many courts use the term "equally effective" when discussing whether the alternative offered by the party challenging the test is feasible and would effectively meet the institution's goals. *See, e.g., Georgia State Conf.*, 775 F.2d at 1417; *Sharif*, 709 F. Supp. at 361. Other courts use the term "comparably effective" in evaluating proposed alternatives. *See, e.g., Elston*, 997 F.2d at 1407; *Fitzpatrick v. City of Atlanta*, 2 F.3d 1112, 1118 (11th Cir. 1993). Review of the decisions in these cases indicates that the courts appear to be using the terms synonymously.

including disparate impact discrimination, *see* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).[176])

The party challenging the test has the burden of establishing disparate impact. If disparate impact is established, the educational institution must demonstrate the "educational necessity" of the practice in question.[177] If a sufficient educational justification is established, then the party challenging the test must demonstrate that an alternative with less disparate impact is equally effective in meeting the institution's educational goals or needs in order to prevail.[178]

1. Determining Disproportionate Impact

The first question in the disparate impact analysis is whether there is information indicating a significant disparity in the award of benefits or services to students based on race, national origin, or sex. A variety of methods is commonly used by courts to distinguish differences between outcomes that are statistically and practically significant from those that are random.[179] To determine if a sufficient disparate impact exists, courts have focused on evidence of statistical disparities.[180] Generally, a test has a disproportionate adverse impact if a statistical analysis shows a significant difference from the expected random distribution.[181] There is no rigid mathematical threshold regarding the degree of disproportionality required; however, the statistical evidence must identify disparities

> Generally, if a statistical analysis shows that the success rate for a particular group of students is significantly lower (or the failure rate is significantly higher) than what would be expected from a random distribution, then the test has disproportionate adverse impact.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 59 (Jay P. Heubert & Robert M. Hauser 1999).

[176] Disparate impact disability discrimination may take forms that are not always amenable to analysis through the three-part approach usually applied in race or sex discrimination cases. For example, statistical evidence showing the effect of architectural barriers on persons of various types of disabilities may not be necessary. *See Alexander*, 469 U.S. at 297-300. For this reason, disability discrimination is discussed separately. *See* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).

[177] *See Elston*, 997 F.2d at 1412.

[178] *See Georgia State Conf.*, 775 F.2d at 1417; *see also* Department of Justice, *Title VI Legal Manual*, p. 2.

[179] Some courts have used an 80 percent rule whereby disparate impact is shown when the rate of selection for the less successful group is less than 80percent of the rate of selection for the most successful group. Another type of statistical analysis considers the difference between the expected and observed rates in terms of standard deviations, with the difference generally expected to be more than two or three standard deviations. Another test is known as the "Shoben formula" in which the difference or Z-value in the groups' success rates must be statistically significant. *See Groves*, 776 F. Supp. at 1526-28 (discussing these methods and the cases in which they were used).

[180] *See Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994-97 (1988) (O'Connor, J., plurality opinion).

[181] *See Watson*, 487 U.S. at 995; *Groves*, 776 F. Supp. at 1526-28.

that are sufficiently substantial to raise an inference that the challenged practice caused the disparate results.[182] To establish disparate impact in the context of a selection system, the comparison must be made between those selected for the educational benefit or service and a relevant pool of applicants or test takers.[183]

In general, a specific policy, practice or procedure must be identified as causing the disproportionate adverse effect on the basis of race, national origin, or sex.[184] For example, when a particular use of a test is being challenged, the evidence should show that the test use, rather than other selection factors, accounts for the disparity.[185]

2. Determining Educational Necessity

Where the use of a test results in decisions that have a disparate impact on the basis of race, national origin, or sex, the test use causing the disparity must significantly serve the legitimate educational goals of the institution.[186] This inquiry is usually referred to as determining the "educational necessity" of the test use or determining whether the test is "educationally justified."[187]

---

[182] *See Watson*, 487 U.S. at 994-95; *Groves*, 776 F. Supp. at 1526-27.

[183] When determining disparate impact in the context of a selection system, the comparison pool generally consists of all minimally qualified test takers or applicants. When tests are used to determine placement or some other type of educational treatment, the comparison is between those identified by the test for the placement or educational treatment and the relevant pool of test takers. The precise composition of the comparison pool is determined on a case-by-case basis. *See Wards Cove Packing Co. v. Antonio*, 490 U.S. 642, 650-51 (1989); *Watson*, 487 U.S. at 995-97; *Groves*, 776 F. Supp. at 1525-26.

[184] As noted by Justice O'Connor in *Watson*, courts have found it "relatively easy," when appropriate statistical proof is presented, to identify a standardized test as causing the racial, national origin, or sex related disparity at issue. *See Watson*, 487 U.S. at 994; *see also GI Forum*, 87 F. Supp. 2d at 677-79 (given legally meaningful differences in the pass rates of minority and majority students, plaintiffs made a prima facie showing of disparate impact resulting from a minimum graduation test).

[185] Elements of a decision-making process that cannot be separated for purposes of analysis may be analyzed as one selection practice. *See* Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(k)(1)(B)(i). This is necessary because limiting the disparate impact analysis to a discrete component of a selection process would not allow for situations "where the adverse impact is caused by the interaction of two or more components of the process." *Graffam v. Scott Paper Co.*, 870 F. Supp. 389, 395 (D. Me. 1994), *aff'd*, 60 F.3d 809 (1995).

[186] *See Wards Cove*, 490 U.S. at 659.

[187] *See Board of Educ. v. Harris*, 444 U.S. 130, 151 (1979); *Elston*, 997 F.2d at 1412.

In evaluating educational necessity, both the legitimacy of the educational goal asserted by the institution and the use of the test as a valid means to advance this goal may be at issue. Courts generally give deference to educational institutions to define their own legitimate educational goals[188] and focus more directly on whether the challenged test supports those goals.[189] While the test need not be "essential" or "indispensable" to achieving the institution's educational goal,[190] the educational institution must show a manifest relationship between use of the test and an important educational purpose.[191]

In conducting this analysis, courts have generally considered relevant evidence of validity, reliability, and fairness[192] provided by the test developer and test user to determine the acceptability of the test for the purpose used, giving deference, as appropriate, to the educational institution's testing practices that are within professionally accepted standards.[193] The educational justification inquiry thus generally looks at technical questions regarding the test's accuracy in relation to the

[188] *See Groves*, 776 F. Supp. at 1529 (citing *Wards Cove*, 490 U.S. at 659).

[189] *See, e.g., Debra P.*, 644 F.2d at 402 (indicating that the court is not in a position to determine education policy, and the state's efforts to establish minimum standards and improve educational quality are praiseworthy).

[190] *See Wards Cove*, 490 U.S. at 659; *Elston*, 997 F.2d at 1412 (citing *Georgia State Conf.*, 775 F.2d at 1417-18).

[191] *See Georgia State Conf.*, 775 F.2d at 1418 (showing required that "achievement grouping practices bear a manifest demonstrable relationship to classroom education"); *Sharif*, 709 F. Supp. at 362 (defendants must show a manifest relationship between use of the SAT and recognition of academic achievement in high school). As explained in *Elston*, "from consulting the way in which . . . [courts] analyze the 'educational necessity' issue, it becomes clear that . . . [they] are essentially requiring . . . [the educational institution to] show that the challenged course of action is demonstrably necessary to meeting an important educational goal." *Elston*, 997 F.2d at 1412. In other words, the institution can defend the challenged practice on the grounds that it is "supported by a 'substantial legitimate justification.'" *Elston*, 997 F.2d at 1412 (quoting *Georgia State Conf.*, 775 F.2d at 1417); *see also Georgia State Conf.*, 775 F.2d at 1417-18; *Groves*, 776 F. Supp. at 1529-32.

[192] In general, courts have said that validity refers to the accuracy of conclusions drawn from test results. *See Allen v. Alabama State Bd. of Educ.*, 976 F. Supp. 1410, 1420-21 (M.D. Ala. 1997) ("Generally, validity is defined as the degree to which a certain inference from a test is appropriate and meaningful", quoting *Richardson v. Lamar County Bd. of Educ.*, 729 F. Supp. 809, 820 (M.D. Ala. 1989), *aff'd*, 164 F.3d 1347 (11th Cir. 1999), *injunction granted*, 2000 U.S. Dist. LEXIS 123 (M.D. Ala.)); *see also Richardson*, 729 F. Supp. at 820-21 ("[A] test will be valid so long as it is built to yield its intended inference and the design and execution of the test are within the bounds of professional standards accepted by the testing industry."); *Anderson*, 520 F. Supp. at 489 ("Validity in the testing field indicates whether a test measures what it is supposed to measure.").

[193] *See, e.g., United States v. LULAC*, 793 F.2d 636, 640, 649 (5th Cir. 1986) (pointing to substantial expert evidence in the record, including validity studies, indicating that the tests involved were valid measures of the basic skills that teachers should have). The sponsors of the newly revised *Joint Standards* advise that the *Joint Standards* are intended to provide guidance to testing professionals in making such judgments. *See* American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, p. 4 (1999) (hereinafter *Joint Standards*). The *Joint Standards* are discussed more fully in Chapter One of this guide.

Where the evidence indicates that the educational institution is using a test in a manner that does not lead to valid inferences, educational justification may be found lacking. *See Fordice*, 505 U.S. at 736-37 (ruling that Mississippi's exclusive use of ACT scores in making college admissions decisions was not educationally justified, since, among other factors, the ACT's administering organization discouraged this practice); *Groves*, 776 F. Supp.