CLE COMMENTS ON

"THE USE OF TESTS WHEN MAKING HIGH-STAKES DECISIONS FOR STUDENTS"

OFFICE FOR CIVIL RIGHTS (07/6/00 DRAFT)

I. Analysis of Key Issues

It is quite clear that a general appreciation of the basic terms of civil rights analysis -- such as disparity, necessity, and validity -- is not enough. There are, it is true, many places where even this general appreciation is lacking and students suffer from failure of systems to undertake any efforts to examine, justify, or remove the causes of disparate impact on their lives. Nevertheless, providing this general appreciation alone is of little value if, as is often the norm, it is then applied by those institutions and adjudicatory bodies in ways that are superficial and wholly inadequate. In the current wave of school reform, we are already beginning to see this in court decisions that cite the correct legal standards in a very general way but then jump to findings and holdings without spelling out a reasoned application of the general principles to the particular facts, and thus without any way to fairly review the analysis. And, with of course much greater frequency, we are seeing this in educational systems which have only a vague understanding of what it means to justify these decisions and which then move forward with false confidence that they have done so.

Thus, our comments are primarily of two related types. First are proposed changes and additions calling for additional rigorous analysis, so that the logic and implications of both the test measurement principles and the legal principles are more clearly understood -- by education officials, program beneficiaries, and courts and agencies. Second are proposed changes calling for more attention to real world application, and particularly to the ways that high-stakes tests are most frequently being used in the current context of education reform.

The call for a rigorous level of analysis and application of the civil rights principles is frequently confronted by the notion that one shouldn't be *too* rigorous, for fear of stymying much needed education reform. We suggest that this is a wrong-headed notion. It is not merely that civil rights are compatible with good school reform (see pages i and iv of the draft), or that equity is of course a central element of good reform. In a much more specific sense, rather than viewing civil rights as a counterweight that must balance the drive for reform, rigorous application of civil rights principles should be recognized as a great ally of, and powerful engine for, reform because those legal requirements essentially boil down to a demand for two essential elements of standards-based reform -- namely (1) that schools adequately teach students the knowledge and skills we have said all children should learn and (2) that assessment systems truly tell us whether students have learned that knowledge and those skills. In other words, through rigorous application of these legal principles and rights, we can and must ensure that the reform is real.

A. Test Measurement Principles

1. Validity chain of inferences. This is one of the most significant, perhaps most central, advances in the new joint standards. If properly understood and applied by educational and adjudicatory systems, it would advance civil rights and effective school reform several fold. Carefully tease out each of the assumptions underlying the particular way a test is being used, in terms of what must be true in order for the decisions based on the test to be valid; then amass and examine the strength of the evidence for confirming or disconfirming each of those assumptions. This is now the central analytic tool for ensuring that tests are used properly. While some reference is made to this element of the Joint Standards, it is not treated as effectively and systematically as it must be, and, as a result, large numbers of readers will not understand how to use it or the importance of doing so. For example, the draft should be changed to:

a. Provide a clear example of the full chain of inferences that might underlie a particular test use. We have provided one such example delineated by the Joint Standards (in our comment to page 21) -- concerning placement in an advanced class – while emphasizing the importance of highlighting that a part of this inference chain would change when applied to other, currently common uses of high-stakes tests. (Indeed highlighting this change is an important way of helping the reader understand how to work with such validity chains.)

b. Revise the chart designed to address this issue, which leaves out the most critical step altogether (identifying the set of inferences that, if true, would support the use of the test to accomplish the purpose) and uses imprecise, confusing language in many of the subsequent steps (comments to page 33).

c. Use this method of analysis throughout the document to provide more help to readers in looking at each of the major areas of high-stakes use, which will help unify the various pieces of the draft and make it a coherent whole, and, more importantly, will allow readers to better understand and apply the various principles in the draft to the issues which confront them. For example, see our more specific comments [to pages 9-10, and to pages 52-53] concerning application to promotion, graduation, and placement.

2. Examining the consequences of test use. (Pages 24-25 of the draft.) This is another area where the revision of the Joint Standards makes a major advance which is not adequately described and used in the draft. The draft focuses on the "consequence" of disparate impact -- such as students being disproportionately retained in grade or not promoted. But such disparity is, legally, the trigger for requiring the inquiry into *all* aspects of validity of the decision in the first place -- a long-standing basic principle. What the new focus in the Joint Standards on examining consequences does is to highlight one previously underexamined aspect of validity -- namely the need to examine the intended and unintended consequences of the testing initiative itself, and the evidence supporting or disconfirming

each of those intended or unintended consequences. As the Joint Standards make clear, this includes such beliefs as that the high-stakes use of a test will motivate students to learn more, motivate teachers to improve their teaching practice, and be an essential driver for schools to undertake significant reform, or, on the other hand, motivate more students to become discouraged and drop out, motivate teachers to narrowly teach to the test (thereby undermining the validity of the test for representing a broader domain of knowledge and skills), motivate schools to ignore students furthest from the cut score (and/or exclude them from assessments or encourage them to drop out), and become a substitute for, rather than a driver of, fundamental school changes. Much of current policy-making and public-engagement in education now revolves around these questions, and the Joint Standards treatment of testing consequences tell us that they, and the supporting and disconfirming evidence, must be carefully examined. Again, the validity focus in this part of the standards is on these questions that go beyond whether the decisions made on the basis of the test are accurate (e.g., that students denied diplomas do not have the requisite skills) but instead focus on whether the rationale for adopting this use of the test are valid. The draft ignores the heart of the Joint Standards work in this area. Our comments include important language from the Joint Standards designed to remedy this.

3. Connecting the dots . . . on validity, reliability, high-stakes, margin of error, cut scores, test as sole determinant, and multiple measures. These are highly related concerns, and the draft needs a good deal of strengthening in its discussion of each one, in ways that could be addressed by helping the readers making the connections. These points come up in a number of ways:

a. **Margin of error and individual high-stakes use.** The point that the higher the stakes, the greater the demands for validity and reliability of the decision needs to be expanded to emphasize and clarify that (i) the use of a test that may have been validated for program evaluation purposes demands a higher degree of validation, and a *smaller* margin of error, when additional high-stakes consequences for students are attached; but (ii) a test typically has a much *greater* margin of error when disaggregating down to the individual student level than it has when the results were aggregated for a whole grade or whole school. (Indeed at the individual level, the standard margin of error of a test can often be greater than the mean difference between grade levels, raising major caution about the use of such a test in making promotion decisions.) See our comments to page 26 n. 76, and to page 19 at end of paragraph 2 (quoting pages 139-40 of the Joint Standards on higher stakes).

b. **Cut scores.** The discussion of cut points is too cryptic and in major need of clarification, through application of the overall principles of validity and reliability, including the examination of the chain of inferences. Here is where it is important to emphasize that what must be validated is not "the test" per se, or even "the use of the test" in some general way, but the decision that is being made on the basis of the test (and any additional factors) -- e.g., the decision to treat students differently from each other, by promoting some and retaining others. In designing the assessment process, it

is thus critical to identify and examine the evidence for the inferences that support that decision -- including that the test use, including the selection of the cut-off score, is capable of distinguishing with sufficient accuracy (in light of the need discussed in the paragraph above to minimize error) between those students who do and do not have the requisite characteristics. See our comments to page 31 (1st new paragraph) and to pages 31-32. See also our comment (to page 32, fn. 105) that the document should include and highlight the Joint Standards' commentary to Standard 4.19, regarding precision in regions of scores scales where cut points are established.

c. **Sole criterion and multiple measures.** The importance of using other relevant information, including multiple forms of assessment, needs to be put in the context of the overall examination of validity inferences and the discussion noted above about margin of error and cut scores. In particular the discussion of the need to avoid harm to students and thus protect them against those errors which have significant negative consequences, the margin of error at the individual level inherent in tests (particularly in tests originally designed for aggregate conclusions), and the analysis of cut score determinations should all be connected more clearly to explaining the need to avoid reliance on a single test. Comments in the draft about how no test is perfect in this regard, as currently written, can be misconstrued to signal tolerance for a significant degree of error. Instead, by more fully connecting the reliability and cut-score discussion to the discussion of sole criterion and alternative measures, these comments can and should be recrafted to get at their true meaning -- i.e., all the more reason to look at whether available multiple and alternative measures are being used in high-stakes contexts. The tolerable degree of error in high-stakes contexts needs to be understood in light of the availability of such alternatives[1]. This is particularly true for students whose scores fall within the range of potential error. (Also important are (i) properly defining "solely" to include cases where, despite the use of other criteria, test scores alone can nevertheless result in a high-stakes negative consequence, and (ii) speaking to the validity and reliability of the overall decision-making process where multiple measures are used.) See comments to page 31 (last paragraph)-32 (1st paragraph), and page 32 (last paragraph, last sentence). See also our comments to page 3, paragraph 3, and to page 19, paragraph 2.[2]

---

[1]This is thus also an area where the connection between the test measurement principles and legal principles can profitably be made more explicit -- i.e., in regard to the legal principle concerning the availability of alternative approaches for serving the same educational ends with less disparate impact. (See section C. below on linking the two analyses.)

[2]For the reasons expressed here, we also share the central concerns expressed by the Leadership Conference on Civil Rights about the need to avoid reliance on a single test (as discussed above) and the need to recraft language that inadvertently would seem to be endorsing various test

d. See also the comment to page 19 (at end of paragraph 2), proposing that the document include a key paragraph from p. 139-140 of the Joint Standards which helps connect these issues of higher stakes demanding higher test quality, minimizing errors in classifying individuals into categories with consequences, and collecting collateral evidence.

B. Legal Principles

1. Educational necessity

a. Legitimacy of the Goals. While we agree that most of the "action" on educational necessity is around the means, not the goals, the commentary about deference to the institution's statement of its goals is overstated in suggesting that anything goes and leaving readers with no criteria for this part of the inquiry. We suggest that, as a component of "educational necessity," due deference does not eliminate criteria for considering (i) the importance of the goals (in relation to the high-stakes consequences for children), (ii) the need to frame them (particularly in terms of learning outcomes as goals) in non-discriminatory ways, and (iii) the need to consider their consistency with stated public policy. See our comments to page 51 (last paragraph) through page 52 (middle paragraph).

b. Validity of the means. (Page 52 last paragraph and page 53 first paragraph.) Considering that this is the heart of the matter, in terms of civil rights principles applicable to high-stakes testing, the analysis here -- two short paragraphs with footnotes -- is brief in the extreme. Our comments highlight the need for and usefulness of helping readers translate the analysis of specific inferences being made (see comments earlier on the chain inferences) into the context of examining the legal adequacy of particular high-stakes practices.[3] This includes identification in our comments of issues relevant to certain very common phenomena:

(i) **Curriculum placement (including tracking), in the light of new state standards for all students.** In particular, the presence of such standards now changes the analysis, in terms of whether curriculum placements for lower achieving students (including low tracks) provide unequal access to the high-level skills the state has said all students should learn. Earlier, pre-standard analysis, to the extent that it was premised on the acceptability of lower

---

uses. (In this regard, see also our comment below to page 11, (3).)

[3]Alternatively, *these practices, and our comments on them, could be more fully addressed in Chapter 1* (and/or in the portions of the Introduction, pp. 8-12, where many of these practices are discussed). But in that case, there should be clearer statements of the connections to the principles and analysis in that Chapter (including to the commentary summarized here).

expectations for some children, and then examining educational benefit in relation to those lower expectations, is no longer adequate. (Comment to page 52-53 and page 8.)

(ii) **Adequacy of instruction.** (Comments to page 52-53 and page 11(2).) The instructional aspects of validity deserve greater attention in the civil rights, disparate impact analysis of use of test for promotion and graduation.[4] Each of the inferences about the adequacy of instruction need to be articulated and examined -- for example regarding alignment of the curriculum, efficacy of instructional methods for all students, qualifications of teachers to provide this instruction, etc. Two other, related issues concerning adequacy of instruction also deserve attention:

Use of tests simultaneously for program assessment/accountability and for high-stakes student decisions. In particular, looking at the chain of inferences that support the two uses may reveal some fundamental conflicts. The former is premised on the assessment providing useful information about the extent to which the school needs to change in order effectively teach the skills and knowledge in the state's standards.[5] The latter must, as a matter of law, be premised on students' already having been so taught. (And, in terms of the validity obligation to examine the intended and unintended consequences, the fact that the educational institution must justify its high-stakes use by trying to show that students have had an adequate opportunity to learn those skills and knowledge would seem to provide counter-evidence to claims that high-stakes use will advance school reform. That is, the necessary argument that the school already provides instruction adequate to master those standards reduces the urgency of the need for further reform, to say the least.)

Programmatic legal obligations. With the coming of standards for what all students should know and be able to do, school systems have also assumed obligations under various state and federal laws for the provision of educational

---

[4]The statement at the end of the 1st paragraph on page 53 -- that courts "may" consider whether the skills taught have been tested is wholly inadequate in this regard and should be changed or deleted.

[5]A major gap between the skills and knowledge being assessed and what is being taught does not undermine the validity of the assessment for purposes of program evaluation and accountability -- indeed the purpose of the assessment is to detect such gaps. Such a gap does, in contrast, undermine the validity of the assessment for purposes of promotion and graduation, which holds students accountable for what they have purportedly been taught.

programs that effectively teach the content of those standards. For example, Title I (along with some state laws) requires accelerated, enriched curriculum aligned with high standards, effective instructional techniques, highly qualified teachers, and time intervention for individual students having difficulty mastering particular standards -- with the plan for doing so to be jointly developed with the parents of the school. As another example, the Equal Educational Opportunities Act imposes obligations (as delineated in *Castaneda* v. *Pickard*) regarding design, implementation, and review of programs serving students with limited English proficiency, which should now be interpreted in light of what the state has said all students should know. Failure to follow such requirements under Title I, EEOA, or other laws, to the extent that they relate to program quality, should certainly be viewed as relevant evidence in examining the testing premises concerning adequacy of instruction.

Note: **Additional discussion of adequacy of instruction and curricular/instructional match** is found in various other sections throughout our specific comments (below), such as:

Page 9- b. Promotion (including application to students with disabilities
Page 10 - c. Graduation decisions (including overall comment and comment to line 3)
Page 11 - (2) Attribution of cause
Pages 17-18, 63 - Instructional match under Due Process analysis

(iii) **Use of exit exams and exam-based diplomas for subsequent decisions.** In many places, standards-based reform is being accompanied by actions and words to encourage use of exam results and exam-based diplomas for employment and admission to state colleges. The basic requirement of validating the particular use of the exit exams and diplomas (in terms of relatedness to necessary job skills or postsecondary success) should be connected to the decisions that are then based on those exam results (particularly in light of both the facts and the law in the seminal case of test use -- *Griggs* v. *Duke Power Co.*).

(iv) **Differential diplomas.** In some states and districts, failure to pass certain tests results in a different diploma, rather than no diploma. The basic legal and educational analysis must be applied here as well, to the extent that these differences affect important opportunities, such as postsecondary admission or financial aid.

2. Due Process. Among our concerns about the due process section:

(a) There is a key missing question in the legal structure. In moving from "Is the purpose of the testing program legitimate?" to "Have students received adequate notice of the test and its consequences?", the analysis skips over the question of whether the test serves the intended purpose -- that it does the test measure what it purports to measure? See our comments to pages 17-18 and page 62 (after the first paragraph).

(b) The discussion of whether students were actually taught the knowledge and skills measured by the test should be informed by recent developments concerning programmatic obligations, described above, that now accompany standards-based reform and related laws , and the development of a stronger body of research about educational practices that result in high student achievement. See comment to page 63 and pages 17-18.

C. Linking the Analysis of Test Measurement Principles and Legal Principles

We do not question the organization of the guide, dividing the two sets of principles into separate chapters. We understand the benefits of doing so, but want to call your attention to the need to deal with some of the challenges it also creates. Ultimately, readers must be able to integrate the two analyses, and there are numerous important opportunities to help them do so. We have pointed out several of these connections in the analysis above -- showing where legal requirements inform the examination of the validity chain of inferences in Chapter 1 and vice-versa in Chapter 2. Other needs and opportunities arise elsewhere in the document, some of which are identified in our additional comments below.[6]

D. Testing of Students with Disabilities

The core principles discussed above -- aimed at ensuring that high-stakes determinations about students are based on sufficiently valid and reliable information and on students having had adequate opportunity to learn that for which they are being held responsible -- apply fully as the starting point for analysis of testing of students with disabilities. In addition, within this basic framework, there are important principles worthy of additional attention in the guide, including the need to:

---

[6]For example, while the discussion of "flagging" the scores of students with disabilities has been pared back, it remains in a footnote in Chapter 1 citing the Joint Standards provision allowing such flagging "if permitted by law" in those instances where evidence of score comparability is lacking. But the draft contains no legal analysis that would help in understanding what is or is not permitted by law and, in particular, the non-discrimination obligations under Section 504. See our comments to page 42, footnote 133 (highlighting the legal problems when failure to meet the obligation to provide comparable assessments is compounded by then stigmatizing the students who, through no fault of their own, did not have a comparable opportunity to demonstrate their skills).

C Ensure that a particular test score's inferences accurately reflect the intended construct rather than any disability or extraneous disability-related characteristic. (See Joint Standard 10.1.)
C Ensure for any test that the type that items, response formats, scoring procedures, and test administration procedures are selected based on the purposes of the tests, the domain to be measured, and the intended test takers. Based on Standard 3.6, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. Samples of persons with a full range of disabilities must be included in field tests, and the review process must in addition to empirical analyses include expert panels comprised of persons with disabilities and qualified persons with particular knowledge of specific disabilities in relation to test administration and response formats. See also Standard 10.3 requiring, where feasible, tests that have been modified for use with persons with disabilities to be pilot tested on individuals who have similar disabilities so as to investigate the appropriateness and feasibility of the modifications.
C Ensure that persons making decisions about accommodations and modifications of any test being administered to individual students with disabilities are knowledgeable of existing research on the effects of disabilities on test performance. Those who modify tests should have access to psychometric expertise for so doing. See Standard 10.2. The comments point out that in some areas there may be little known about the effects of a particular disability on a particular type of test. (See also Standard 10.8.)
C Ensure for any test that is being used to assign persons to alternative treatment, for example as a basis for deciding that a student with a disability will not participate in the general or regular education curriculum and will be assessed using an alternate assessment, that whenever possible outcomes are compared on a common criterion and supporting evidence of differential outcomes is provided. It is not sufficient to show that the test predicts treatment outcomes, but supporting evidence of differential outcomes must be provided. Support for the validity of the classification or assignment is provided by showing that the test is useful in determining which persons are likely to profit differentially from one educational program over another -- e.g., special education. (Standard 1.19.)
C Ensure that, for any testing application that involves individual interpretation of test scores, a test-taker's score is not accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test-takers performance on the test at that time. While variables such as socioeconomic status, school history, language, ethnicity, culture and gender may be relevant and considered while interpreting test scores for all students, additional disability-related explanations, for example, fatigue, medication, spasticity, or visual impairment, may affect the performance of a test-taker student with a disability on, for example, a paper-and-pencil standardized test. (Standard 7.5.)
C Ensure that in assessing students with disabilities for diagnostic and intervention purposes, no test is used as the sole indicator of student functioning or performance. Standard 10.12 expressly states that "[i]n 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."

Below are our more specific comments tied to the particular pages and paragraphs of the text -- addressing both the issues discussed in this "key issue analysis" and other comments.

II. Specific Comments (by page and paragraph number)

Cover letter

Page iv, last paragraph: "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. . . ."

Suggested Clarifications:

1. This paragraph summary of the law is likely to be misread by many readers, for several related reasons.

a. The first sentence statement that test score disparity does not alone constitute discrimination is still overbroad -- it does constitute a prima facie case, which may by itself result in a finding of discrimination if sufficient evidence relating to educational necessity is not provided. At a minimum, change "does not alone constitute discrimination" to "does not necessarily constitute discrimination."

b. In a related vein, the second sentence, stating that the "guarantee under federal law is for equal opportunity, not equal results" is a very frequently stated but misleading sentiment that is often taken to mean that considerably less is required by the law than is actually the case. In any event, as indicated in the first point, results do matter under the civil rights laws. We suggest dropping this sentence altogether as highly subject to misinterpretation, and letting the rest of the paragraph speak for itself.

c. The statement that "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..." can be viewed as implying, incorrectly, that school systems are free to make all sorts of educational decisions with disparate impact so long as the factors producing the disparities are not in their control (even though the educational decisions are). This is simply not true (and indeed race, gender, and disability are themselves beyond school systems' control, which of course does not free them from legal obligations in this regard). This sentence (and the short one immediately thereafter) should be deleted.

d. Given the space devoted to this paragraph, the legal standard that is stated -- that disparities (including on the basis of disability) should result in thoroughly examining the educational practices at issue "to ensure that they are in fact non-discriminatory and educationally sound" -- is not as informative as it should be. A short version of the educationally necessity and no feasible, less disparate alternative standard would be more so.

2. At the end of the paragraph -- after the statement that the law is not designed to water down or frustrate the establishment and application of educational standards -- add:
"In fact, properly understood, the legal standards are an aide to meaningful educational reform -- by insisting that assessments are structured fully to measure, and instructional programs are structured fully to teach, the high-level skills and knowledge that sensible and rigorous standards seek for all children."

Introduction

Page 3. paragraph 2: The following statement is incomplete, thus, misleading, and raises a significant issue: "When educators and policy makers consider 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."

The fact that the same test has been independently validated for each particular purpose for which it is being used doesn't address, let alone, resolve the problem. Rather, it is the inappropriate simultaneous use of the same test for making high stakes decisions that creates a conflict. Results from the same tests indicating inadequate student performance are, on one hand, being used to deny students promotion or graduation while simultaneously identifying as underperforming the schools that have failed to prepare adequately the students. These simultaneous uses are typically based on mutually conflicting premises or inferences. (This issue is discussed more fully in part (2)(a)(ii) of our comments to pages 51-53, and summarized in I.B.1.b.ii above.)

Page 3, paragraph 3: The following statement is confusing: "When high stakes decisions are made, *test scores* are often used in conjunction with other criteria, such as grades and teacher recommendations. A test should not be used as the sole criterion for making a high stakes decision unless it is validated for this use. The Joint Standards state that '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.'"[footnote omitted]

What this paragraph ought to convey is that no standardized test -- whether administered once or multiple times - ought to be used as the sole criterion for making a high stakes decision. Rather, other relevant information must be considered whenever it will enhance the overall validity of the decision. (See I.A.3.c above)

Test Use Principles

Page 8 - a. Placement decisions

1. Application to placement decisions needs a great deal more clarification. We have addressed this through comment to page 52-53 below, but some or all of those comments could go here.

2. After the reference to the Joint Standard that there should be adequate evidence documenting the relationship among test scores, appropriate instructional programs and beneficial outcomes [citing Standard 13.9], clarify by using examples. For example, when a student is determined, in part, based on test scores, to have a particular type of disability and educational needs requiring specialized instruction and related services, provision of such of instruction must be monitored to ensure that it is beneficial and effective. Also, unlike the child taking a test of mathematical skills to determine whether s/he should participate in a math program for gifted students, where the test should adequately cover content and thought processes that are essential to the instruction students will actually receive in that program, no single test may be used to determine a child's disability or the educational program tailored to meet his/her needs and the child must be placed to the maximum extent possible in the regular education classroom with his/her non-disabled peers.

Page ~~9~~ - b. Promotion decisions

1. Application to promotion decisions needs a great deal more clarification. We have addressed this through comment to page 52-53 below, but some or all of those comments could go here.

2. Other Suggested changes: Given the extent that high stakes decisions are being implemented concerning promotion or retention of students based on test outcomes, further explanation is needed to ensure that the issues are being properly understood and steps taken consistent with law and professional practice. As written the text provides: "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 date and a determination regarding whether the student will be able to master the content at the next grade level (a placement decision). At present, the focus of most school districts and states with promotion policies has been primarily on assessing mastery of curriculum taught at a given grade level."

Instead of just citing to the *Joint Standards* at fn. 19, additional explanation would be helpful. We suggest the following language: "This means that any test used for promotion purposes must have been validated to demonstrate that the body of knowledge and particular skills are necessary to participate in the next grade successfully; that the test actually assesses the knowledge and skills determined to be prerequisites to promotion with a high degree of reliability across race, gender, ethnicity, limited English proficiency and disability; that the cut off score(s) accurately identify with a small standard error of measurement persons who are

successful. To ensure that the test is fair assessment of the knowledge and skills of all students, the test must have been validated." As noted, when a test is used for such purpose, its use must adhere to professional standards for certifying knowledge and skills for all students [*Joint Standards* 1999, Standards 13.5, 13.6 and 13.9, High Stakes, 1999:287].

Also, state explicitly that: "This means that all students must be provided with multiple measures for demonstrating mastery; that the items on the test are generally representative of the content and skills that students ***have actually been taught*** [*not adequate to say "covered"*] at their current grade level [*Joint Standards*, Standard 13.5, High Stakes, 124-25] .

Also, state : "For students with disabilities, and other students based on race, ethnicity, or gender who may be disparately affected by any mastery requirement, in part, because they may have been denied multiple years of adequate and appropriate education designed to teach them the cumulative educational skills expected to be learned by all other students, that a compensatory education program be designed and implemented prior to using a test given for promotion purposes."

Page 10 -- c. Graduation Decisions

1. Application to graduation decisions needs a great deal more clarification. We have addressed this through comment to page 52-53 below, but some or all of those comments could go here.

2. Suggested change: **Add** here and presumably in introduction as follows:

"Although this document is intended to address only student accountability, not teacher or school/school district accountability, before students can be fairly denied a diploma because they failed to meet the expected standard of proficiency, the school system must be able to demonstrate its proficiency in meeting the standards of teaching and learning that enable all students without regard to race, gender, language, ethnicity or disability to receive and participate in a sufficient quality program or curriculum that prepare them to learn the knowledge, skills required as a condition of graduation and as measured by the test."

Page 10, Section c., line 3:

With respect to students with disabilities, and other students, *it is not adequate to state*: "When large scale standardized tests are used in making graduation decisions, there should be evidence that the test adequately covers the content and skills that students have had an opportunity to learn. 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 curriculum, instruction, and the material covered on the test used as a

condition for graduation."

Rather, further information is needed to explain the variety of ways that a school could demonstrate that it has provided a curriculum that provided all students the opportunity to learn to the standards required for mastery. Without more detail and examples, this document runs the risk of proselytizing without giving the kind of direction and identifying the kinds of interventions and strategies needed to promote change.

For example, it would be helpful to state the following: "Schools must be able to demonstrate that students, including students with disabilities, have been provided a meaningful opportunity to acquire through effective teaching and learning the content and skills that all students are expected to know and be able to do. Evidence of meaningful opportunities include students receiving for such periods of time as necessary and consistent with IEPs or other individualized plans, a program of teaching and learning that modifies curriculum and instruction, e.g., as by breaking down standards into components of learning, identifying short-term objectives or benchmarks aligned with the curriculum and established standards, tailoring instructional strategies and methods of learning to each student's instructional needs, providing effective one-to-one intervention, compensatory programming, specialized instruction and supportive services through extended school day and school year program or other manner."

Overarching Principles

Pages 10-11 -- (1) Measurement Validity

What is here is good, but we believe that further expansion on validating each of a change of inferences, and providing examples, would be important for helping readers understand the principles and how to apply them.

Page 11 -- (2) Attribution of cause.

[Regarding appropriate instruction, see I.B.1.b(ii) above and our comments to page 52-53 and fn. 175.]

1. In the second sentence, before "equal opportunity to acquire" insert "appropriate instruction and". This tracks the language and the heading and helps supplement "equal opportunity" which otherwise could be read to countenance equally poor instruction.

2. At lines 5-6, ". . . all students have an equal opportunity to acquire the knowledge and skills that are being tested."

Comment: While this is a useful statement here, this point should be reiterated and

explained in the body of Chapter 1.

3. Additionally, please clarify here that this sentence is about equal opportunity to *acquire* the knowledge and skills being tested. The next sentence, which seems intended to illustrate the point, however, is about accommodations in order to provide equal opportunity to *demonstrate* those skills and knowledge on the test. Both concepts are important and should be addressed in both the first sentence (equal opportunity both to acquire and to demonstrate . . .) and in the second (modifications, services, accommodations, etc. both in instruction and in testing in order to provide that equal opportunity).

4. It should also be noted that "equal" does not always mean "the same." For example, children with disabilities or who have limited English proficiency must be given supplementary supports and services, tutoring, extended school services, or even compensatory education to make up for the deficiencies in the cumulative education program.

.

Page 11 --(3) Effectiveness of treatment - Do test scores lead to placements [instructional programs and services]

Suggested change: In this pitch for tests being perceived as beneficial instruments to help children learn effectively, use this as an opportunity to introduce the use of alternative performance measures, other forms of authentic assessments, portfolios, and projects, as multiple measures for assessment.

Page 12, line 3. Footnote 27 citing the importance of education from Brown v. Board of Education is a good footnote which deserves text status, particularly in light of the other examples provided regarding using tests for advanced courses.

Suggested change: Move fn 27 into text.

Page 12, end of section (3), fn. 29 -- re research on low-track classes.

Suggested change: move fn 29 re research concerning low track classes into text. Add to footnote accompanying citation a reference to the Congressional findings regarding 20 years' low expectations held for students with disabilities that were recently codified in the IDEA Amendments of 1997 at 20 U.S.C 1400(c)(4)-(5).

Legal Principles

Page 13 -- a. Different Treatment

Suggested change: Provide specific examples re/LEP, gender, and disability where students were not given a meaningful opportunity to participate in the curriculum (e.g., Lau and Keyes, Carter). Using actual examples from case law will help insure that those reading this document for guidance will actually understand this issue and get it -- that students, given an opportunity to learn and necessary support services, can and do succeed.

Pages 14-15 -- b. Disparate Impact

Suggested change: **insert** disability and parallel references to implementing regulations under Section 504, 34 C.F.R. 104.4 after questions (1), (2) and (3).

Comment: Paragraph one refers to a further discussion of issues related to students with disabilities being covered elsewhere in the document....presumably recognizing some special issues and concerns relating to disability. Paragraph 3 discusses disparate impact with respect to race, ethnicity, and sex, but is silent with respect to disability. Though left unresolved by the Supreme Court, legal analysis for disparate impact remains no different regarding the paragraph 3 questions to be raised and examined for race, sex, ethnicity. See Alexander v. Choate, 469 U.S. 287, 301, 302-09 (1985); Oberti v. Board of Education, 801 F. Supp. 1392, 1405 (D.N.J. 1992).

Principles Relating to Inclusion and Accommodations

Page 15, fn. 40.

Suggested change: Move footnote into the text.
Comment: Although the Title I requirement is not *per se* about high-stakes testing, for purposes of this guide, the continued extent of confusion about Title I affirmative obligations with respect to LEP students warrants highlighting the provisions, particularly with respect to assessment.

Page 16, first paragraph, line 6 & page 59, first full paragraph

Comment: This section needs to be clear and to send the right message concerning the inclusion of students with disabilities, with accommodations, if appropriate; to more fully address the issues the decision-making process and the analysis as it relates to use of state or district wide assessments, including alternate assessments that ,ensure different content and alternative performance assessments (designed to measure same content/constructs), the independent and individual decision whether or not the student ought to be participating in a "high stakes" testing based, in large part, on validity issues as well as opportunity to learn; and

consequential provision of instruction and FACE.

Note that Congress recognized in the legislative findings of the IDEA Amendments of 1997, that implementation of students' rights to participate in the general education curriculum and to meet the standards established for all other students has been impeded for more than 20 years by low expectations of students with disabilities and the failure to apply replicable research on proven methods of teaching and learning for children with disabilities. 20 U.S.C. 1400(c)(4)-(5). Furthermore, history has shown that IEP teams, which are dominated by school personnel, cannot be given unfettered authority to decide when students with disabilities shall be included in state or district-wide assessments. In no case shall any student with a disability be exempt or otherwise excluded from participation in large scale state or district-wide assessments on the basis of limited expectations based on disability status, type of disability, time spent in the general curriculum, prior inadequate access to learning or at the unfettered discretion of the IEP team. Significantly, the decision to participate is independent from how the assessment information is used, particularly with respect to whether it has or should have high stakes consequences for the student.

Moreover, students with disabilities are entitled under the civil rights statutes, specifically Section 504 and the ADA, to challenge any decision to exclude them from the State accountability system, and to participate equitably and effectively in any State or district-wide assessment, with accommodations or other support services, if necessary, as they are entitled to comparable aids, benefits, and services provided their non-disabled peers. 34 C.F.R. 100.4(b). For the small number of children who cannot participate in the State or district-wide assessments, even with accommodations, alternate assessments must be developed and conducted by July 1, 2000. 20 U.S.C. 1412(a)(17); 34 C.F.R. 300.138; 34 C.F.R. 100.4(b).

It is noteworthy that most disabled students will be able to participate in a State or district wide assessment with no accommodations. Other students with disabilities are also capable of learning what the large scale assessment tests, but need a different way of demonstrating their mastery of, e.g., reading or mathematics skills. They too are entitled to accommodation, as through an alternative assessment - a portfolio or other performance assessment -that permits them to demonstrate their level of mastery -that they are capable of learning what the State or district-wide assessments test, but need a different way of showing it. These students do not require an alternate assessment (defined here as an assessment that measures different content), but merely a different way of demonstrating what they know and can do - the same content being assessed through paper and pencil on the standard administration of the assessment being used. For example, a student who has mastered the basic concepts of geometry but, who because of This particular neurological impairment, cannot demonstrate what he knows and can do on a standard administration of the test using paper and pencil, must be provided the opportunity to demonstrate his knowledge, skills and proficiencies using manipulatives. If the content being assessed is the same, then demonstrating

learning by doing it through an alternative performance assessment should be conceived of as an accommodation that allows this particular student to equitably and effectively participate in the large scale assessment. Presumably this type of accommodation would, as all others, need to be validated. OCR needs to offer guidance to test users when the content being assessed is virtually identical.

If the definition of accommodation is not construed to encompass the needs of students with disabilities who because of the nature of their disabilities need a different way to demonstrate their full degree of mastery of the knowledge, skills or understanding being assessed by the standard administration of the test, then these students must be provided such an opportunity to participate in a different type of assessment that enables them to demonstrate that they have learned the same skills, knowledge and understanding expected to be learned by all other students participating in the standard administration of the test or the particular test with such accommodations, as necessary.

Suggested clarification: Given that the focus of this guide is on the use of assessments to make high stakes determinations *about students*, it is important to bear in mind that the consequence of the system includes what is typically perceived of as negative actions for the student (potential retention, failure to graduate, etc.), in addition to potential benefits of the accountability system (ensuring that schools provide the program that teaches students to high standards). Just as it would with any student, the margin of error tolerated would have to be much lower than when used, with many other student scores, to evaluate a school or program. The test must be valid and reliable for students with disabilities, and particularly for students with this particular disability. Appropriate accommodations, as described above, must be considered in assessing the student and provided, as needed, including in the form of an alternative performance assessment.

Further, the student must have been given a full opportunity to learn what is being assessed, as ensured by the IEP Team‘s providing the student with such specialized instruction, related services, and supplementary aids and services as may be needed to enable the student to overcome any barrier as may exist as a result of his/her disability. Significantly, the question of whether a student participates in a particular assessment is distinguished from how the assessment information is used, **particularly with respect to whether it has or should have high stakes consequences for the student**.

Thus, while the IEP team will in most instances determine that the student ought to participate in the assessment and be included in the accountability system, the IEP team must ensure that the test is valid for purposes of determining whether the individual student should be retained or promoted, if it is to be so applied. And to the degree that the test is not validated, the IEP team, other qualified persons, including the parents and the child, as appropriate, must as part of its annual review and revision of the student’s IEP, determine whether or not based

on a variety of information and data, including the assessment, the degree to which the student has reached his/her performance goals and indicators, and other benchmarks of learning, whether the student should be retained, promoted and with what kind of interventions and strategies for improving teaching and learning. Just as in the case of any other student, the school or district would be responsible for ensuring that the student has had the opportunity to learn and that the assessments used are valid and reliable.

Federal Constitutional Questions

Page 17, **bottom** (the three questions for examining due process -- legitimate and reasonable purpose, adequate notice, and students actually taught the tested knowledge and skills), and
p. 18, **first paragraph** ("Federal courts have typically deferred to educators' judgments about the beneficial purposes of a testing program. . . .").

Suggested changes:

(1) On page 17, add a new bullet after the first one: "Is the test use a reasonable way of achieving that purpose?"

(2) At the top of page 18, begin the first sentence with: "For due process questions,".

(3) On page 18, after the first paragraph, add a new paragraph explaining the suggested new bullet, in terms of accurate and appropriate (i.e., valid) means of achieving the purpose (comparable to the other paragraphs on the page explaining the last two bullets).

Reason: There is a missing question and level of analysis. It is true that, for due process analysis, federal courts have typically deferred to educators' judgments *about "the beneficial educational purposes" or "reasonable goals,"* such as "improving the quality of education, ensuring that students can compete, . . . " etc. But the question of whether the test is reasonably related to that purpose is a separate issue subject to scrutiny.[7] (*Debra P.*, et al.) Indeed, it is more typically the means, not the ends, that is the focus of due process inquiry, both substantively and procedurally. Just as instructional or curricular aspects of validity are implicated in the question about whether students were actually taught the knowledge and skills being measured, so too under the case law other aspects of validity are implicated in the question about whether the tests as used are measuring those things related to the educational

[7]To cite an extreme or absurd example, a graduation test instituted for these laudable purposes that consisted of a single short-answer question to a question from Trivial Pursuits (or ten such questions) would not seem to meet this standard, regardless of whether the separate standard of having been taught the matter were met. (Or suppose the test is one that produces demonstrably unreliable results from one student to the next. That has all the hallmarks of lack of due process analogous to a fact-finding process for deprivations of liberty or property that regularly fail to find the facts accurately.) More useful examples for the reader would deal with less obviously absurd measures.

purpose.

See also comment to analogous due process analysis on page 62 (after the first paragraph).

The other suggested change -- adding "In due process questions," -- is proposed in order to help readers stay oriented, given the necessarily rapid pace at which shifts are made in the text through the various legal doctrines, to keep in mind that this is due process analysis, distinct, for example, from the earlier Title VI analysis.

Page 18, last paragraph -- on reasonable opportunity to learn the material

See comment to page 63 question (3) for additional issues regarding instructional match. Those comments should be incorporated here as well.

dd

Chapter 1. Test Measurement Principles

Page 19, para 2. At line 5, the text reads as follows: "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."

Suggested clarification. Before the words "Information may include", insert: *"When used to make high stakes decisions,"*
Before "test results", **insert**: *"valid and reliable"*,
Delete "as well as other"and **insert**: *"but must include"*

Reason. The permissive language here could mislead jurisdictions using assessments for high stakes. It would help to clarify that the principle is actually that when using an assessment with high stakes for students, decisions "should not be made on the basis of a single test score. Other relevant information should be taken into account. . . ." In fact, the use of the word *must* would be appropriate here as well, given the legal principles involved.

Pages 19, after paragraph 2 -- As the stakes for individual students increase

Suggested clarification and addition: Add (either here *or elsewhere* in the section):
Joint Standards Commentary on "Stakes of Testing":

"The higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality. In particular, when the stakes for an individual are high, and important decisions depend substantially on test performance, test needs to exhibit higher standards of technical quality of its avowed purposes than might be expected of tests for lower-stakes purposes. . . .Although it is never possible to

achieve perfect accuracy in describing an individual's performance, efforts need to be made to minimize errors in estimated individual scores in classifying individuals in pass/fail or admit/reject categories. Further, enhancing validity for high stakes purposes, whether individual or institutional, typically entails collecting sound collateral information both to assist in understanding the factors that contribute d to test results and to provide corroborating evidence that supports; inferences based on test results" (*Joint Standards*, pages 139-140)

Reason: This is much more helpful to readers in understanding the issues than only quoting the portion currently in paragraph 2. To ensure that the reader more fully appreciates how these issues are connected and how higher stakes demand higher standards of quality and assurances, including collecting collateral evidence, so as to minimize errors and mitigate harmful consequences -intended or unintended, this commentary is especially helpful.

Page 20, continuing paragraph from page 19:

Suggested clarification: After "When test results are used to make high-stakes decisions" insert ", for example."

Reason: For example is needed, as this point is not limited to high stakes decisions about promotion or graduation exclusively, but presumably pertains to such decisions as referral for evaluation for eligibility for special education, eligibility for gifted program, placement in exam schools or other programs having criteria for inclusion.

Page 21, 1. Validity of the Inferences of the Scores

Suggested clarification: Expand the comments here, which while quite useful, ought to be discussed and explained more explicitly in the context of and consistent with the *Joint Standards* -- for example, to help the reader understand how to identify the full set of inferences about the test that are at play in the way the test results are actually used. See for example the *Joint Standards*, page 9 that read as follows:

> "Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. (page 9)
>
> " The decision about what types of evidence are important for validation in each instance can be clarified by developing a set of propositions that support the proposed interpretation for the particular purpose of testing. For instance, when a mathematics achievement test is used to assess readiness for an advanced course, evidence for the following propositions might be deemed necessary: (a) that certain skills are prerequisite for the advanced course; (b) that the content domain of the test is consistent with these prerequisite skills; (c) that test scores can be generalized across

relevant sets of items; (d) that test scores are not unduly influenced by ancillary variables, such as writing ability; (e) that success in the advanced course can be validly assessed; and (f) that examinees with high scores on the test will be more successful in the advanced course than examinees with low scores on the test [but see comment below] (pages 9-10)

"Identifying the propositions implied by a proposed test interpretation can be facilitated by considering rival hypotheses that may challenge the proposed interpretation. (page 10)

"Because a validity argument typically depends on more than one proposition, strong evidence in support of one in no way diminishes the need for evidence to support the others." (page 11)

In using the above, useful example, it is ***ESSENTIAL*** for the OCR Guidance to underscore to the reader that high-stakes testing as actually practiced in schools is often for a somewhat different use that will change the specifics of the inferences or propositions, *particularly in regard to proposition (f) above*. The Guidance should, therefore, emphasize this point, including with language such as the following **Suggested addition**:

"In the context of a set of rigorous standards for what all children show know and be able to do, a test that limits access to courses which provide more or higher quality access to instruction related to those rigorous standards -- either in the context of ability grouping or in the context of promotion or graduation tied to test results -- cannot be justified by the proposition that students with high scores will be more successful in those courses. Once it has been determined that these are the standards that all students should meet, it would be inconsistent to then deny access to higher quality instruction for meeting those standards to the very students most in need of assistance."

The focus on how proposition (f) would necessarily change would not only help in avoiding a very harmful misapplication of particular propositions. It would also help the readers more generally improve their overall ability to make such applications in a variety of contexts.

Pages 20-24, Validity

Suggested addition: **Add** to the text the following relevant standards as they help the reader more fully understand the issues being addressed.

"Standard 1.4
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."
"Standard 11.2
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." **For clarification regarding these points, add:** "For example, if a test has been validated in terms of its ability to reflect certain types of reading or math skills, but has not been validated for use in determining promotion, an additional set of inferences about the test results would need to be identified and validated -- such as, for example, inferences about the ability of the test use to distinguish who will or will not be able to function in the next grade level, the impact of this test use on instruction in the desired skills, etc."

"Standard 11.20
In educational, clinical, and counseling settings, a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered."

Pages 24-25 -- 3. Considering consequences of test use

Para. 2. Much of the focus of the new Joint Standards' provision on consequences of test use has been overlooked in the draft. Before the example given here -- identifying differences in placement test scores based on race, gender or national origin as a "consequence" -- add other examples that explain and offer clarification about both intended and unintended consequences of test uses. [Also at line 1, after "gender" insert "disability."]

Suggested addition: "Some intended consequences of high stakes test use might be that teachers would better align instruction with the underlying skills and knowledge in the standards if students are subject to high-stakes consequences, that higher-quality attention will be focused on students experiencing most difficulty, or that students will be more diligent in learning those underlying skills and knowledge. Unintended consequences might be that teachers narrow instruction to something close to the test items (thereby actually undermining the validity that a test might otherwise have had, absent high stakes, in terms of the items' ability to capture the underlying skills and knowledge domains of which they are but a sample), that students having the most difficulty are disproportionately not included in school accountability measures, or that students get discouraged and drop out of school (either, for example, after being retained or after failing the test even though they may have other opportunities to take the test and avoid retention)."

"It is especially important to examine the consequences of the high-stakes aspect of the test use, separate and distinct from consequences that may flow from the administration of the test without attaching high stakes (e.g., for diagnostic purposes)."

Add to text Joint Testing Standard 1.22 and particularly 1.23 and the accompanying comment for the latter:

"Standard 1.22

"When it is clearly stated or implied that recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence."

"Standard 1.23

"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.

"*Comment:* For example, certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices or would clarify students' understanding of the kind or level of achievement that they were expected to attain. To the extent that such claims enter into the justification for a testing program, they become part of the validity argument for test use and so should be examined as part of the validation effort. Due weight should be given to evidence against such predictions, for example, evidence that under some conditions educational testing may have a negative effect on classroom instruction." *[Joint Standards, page 23. See also page 142]*

Reason: While racial disparities in program placement, if they occur, are a consequence, using this example alone is likely to leave many readers confused about consequential validity and its relationship to the overall in the guide, all of which is premised on disparities by race, national origin, disability, or gender. The examples above add clarity about the nature of inquiry into intended or unintended consequences and, in particular, the focus here on consequences that are over and above the interpretation or use of the test results. (The example on the top of page 21, in contrast, focuses directly on the validity of the test interpretation and use, which is of course indistinguishable from the main line of validity analysis but fails to capture the additional inquiry generally associated with the testing principles' inquiry into consequences.) As the comment indicates, articulating, and then examining the evidence for, the intended consequences is an important aspect of assessing the validity of the inferences that underlie the test use; and examining the unintended consequences is part of examining contrary evidence.

Reliability

Page 26, fn. 76

Suggested clarification:

1. Move fn. 76 to text

2. Elaborate with language to the following effect:

"A test which may be fairly reliable for aggregate score purposes will typically be much less reliable for individual purposes, with a much greater standard error or variation in test scores at the individual level. Because the standard error at the individual level can approach or exceed the average difference in test scores between grade levels, the use of a particular test for promotion purposes may be called into question. At the same time, the standard for scrutinizing a test's adequacy is higher when the test is used for high-stakes individual purposes. This means that when a *higher* degree of validity and reliability is needed -- as when making high-stakes individual decisions -- a single test score is typically much *less* reliable. *The need to proceed cautiously in use of such tests for individual high-stakes purposes is particularly strong."*

Reason: While one margin of error may be acceptable for the use of a test for school accountability, that same margin of error may be wholly unacceptable when the test is used for high stakes decisions for individual students. There may be a relatively small margin of error when scores are considered in the aggregate – as in a whole grade, or a whole school – which become quite large when disaggregated to the individual student basis. And, at the same time, the level of error that is acceptable becomes smaller as the negative consequences for individuals stemming from erroneous decisions becomes larger -- i.e., for high-stakes testing. Thus, the size of an acceptable margin of error must be considered in context. (See the *Joint Standards* language on high stakes quoted in our comment to Page 19, after paragraph 2, above.)

<u>Fairness</u>

Page 30, first new paragraph -- "feasibility"

As the draft notes here, the Joint Standards make considerable use of the term "to the extent feasible" and the like. (See, for example, page 11 n. 25, page 28, and 29.) While we believe that this term and others like it are generally intended to create only a narrow exception to the general standard at issue (e.g., where there simply is no way to develop an adequate sample size), we strongly believe that the language will get misconstrued and applied overly broadly. In the educational settings with which we are most familiar, such terms are often perceived to be an indication that the standard which they accompany can largely be safely ignored. Therefore,

we suggest that, in the various places where such a term is referenced, the reference be either deleted or given a more precise and narrow construction, as appropriate. On page 30, there is some effort to address this question, but the language needs to be stronger and clearer if readers are to come away with a narrower interpretation.

Cut Scores

Pages 31, first new paragraph and fn. 98

Move the footnote into the text, and expand on it to the following effect:
"It is necessary to examine the validity of the inferences that underlie the specific decisions being made on the basis of the cut scores. For example, is there an inference that a certain test is capable of distinguishing who is proficient in certain areas and who is not, and that students who score above a certain score on the test are proficient in the standards for that grade and are ready for the next while students who score below it are not? If so, what is the evidence for that inference? This is the actual use that must be validated -- treating students differently on the basis of this score, based on the proposition that the cut-off score adequately distinguishes between the two groups. This is particularly significant where large numbers of scores tend to cluster around the cut-off point, and where a shift in answers on just one or two questions is sufficient within that range to have a big impact on scores.

"Thus, the issues of selection and use of a cut-off score need to be addressed as part of the overall construction of the test and the designing of its use. In particular, where a bright line decision is going to be made (e.g., between proficiency and non-proficiency for purposes of promotion or graduation), the test user needs to develop an approach to assessment capable of making that distinction with sufficient accuracy, rather than to first decide to adopt the test and then be bound what kind of evidence concerning the cutoff is "feasible" despite its inadequacy."

Reasons: The language in the footnote is key to understanding cut scores. The additional suggested text helps explain the connection between the cut scores and the method for examining the inferences behind the decisions being made, and grounds it in practical usage.

Page 31 (last paragraph) - page 32 (first paragraph)

Suggested Clarification: Add language, regarding the two paragraphs generally, and the question raised at the end of the second paragraph specifically, to the following effect:

"In answering such questions and addressing the question of test error, it is also important to look at further steps taken, in light of available alternatives, to determine the actual knowledge and skills of students with failing scores, particularly those within the range of potential error -- such as, for example, the offering of more extended, direct measures of performance for that subgroup of students who have failed the test. In this context, it must be

remembered that any rights to be assessed in a manner that is fair, valid, and reliable are *individual* rights that cannot be denied on the basis of having been provided to other students.

Page 31-32, footnote 100 -- some measurement errors more serious.

The statement that, "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," is not always true in many cases. While special concern for the validity of scores close to the cut point is very important, the obverse is not necessarily true. A student with the requisite skills could nevertheless do extremely poorly on a test and fail by a huge margin because of an endemic problem (for that student) in the format of the test, its administration, or similar factors. This is particularly a concern for students with limited English proficiency or students with disabilities, but is by no means limited to them. Delete the sentence.

Page 32, last paragraph, line 4

Suggested Clarification: Add examples after footnote 104.

Reason: It's not clear what this means. Perhaps some examples of methods to be used to validate cut scores would help. Are there some methods of validation that are preferable to others? What are they? Without knowing what they are, it is hard for us to comment on the substance, though we do think further elaboration would be useful.

Page 32, fn. 105 and accompanying text

Suggested change: Add to text of the final paragraph on page 32 the following quote from the commentary to Section 4.19 of the *Joint Standards.*:

"Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories." [*Joint Standards*, p. 59.]

Reason: This is a particular clear statement in support of the overall point the draft is seeking to make here.

Page 32, last paragraph, last sentence -- on using other relevant information, rather than making decisions "*solely* or automatically on the basis of a single test score."

Suggested Clarification: Define: "solely" as used above. **First, distinguish** between systems in which failure to meet the test criteria can be overridden by meeting other criteria *and* systems in which the additional criteria act as additional requirements that students must meet (e.g.,

where graduation depends upon both a certain test score and course completions). In the latter case, the test does function independently as a sole criterion. **Second, clarify** that the system of criteria for the decision as a whole must be valid and reliable -- a test that itself is not valid and reliable for that purpose is not necessarily saved by the addition of other measures that are also not valid and reliable. What evidence has been amassed to show that the particular way that the factors are combined results in a valid and reliable method for distinguishing between who does and does not meet the underlying important criteria?

Test Measurement Principles: Questions About Appropriate Test Use

Page 33, chart

a. **Missing Step:** An important step -- namely, articulating the inferences -- is necessary in between question 2 and question 3. **Ask:**

"3. What are the particular inferences, or set of propositions, that if true, would support the use of the test to accomplish this purpose? What inferences are being made in using the test and the test results, including inferences about students who score at particular levels?" (and renumber subsequent paragraphs)

Reason: The questions jump from what is the purpose (#1) to is there adequate evidence to support the inferences(#3-6). That jump cannot be made without first carefully articulating what those inferences are in relation to the intended use of the test. (For further discussion see comments to page 21, concerning the chain of inferences whose validity needs to be examined.)

b. Clarifying language for questions 3 and 5: The sentence structure in regard to the use of the term "inferences" in these two paragraphs is somewhat confusing: "that the test score differences are accurate and meaningful for the students," "that the inferences accurately reflect the specific knowledge and skills," "that the inferences are measuring the same constructs," etc. "Inferences" do not, for example, "measure constructs" or perform the other tasks described. The concept of inferences -- and the need to examine the evidence supporting those inferences -- is very important here, as are the related points being made in these paragraphs; the language just needs to be reworked to accurately reflect them. For example, in #3 "is there adequate evidence of validity to document that these score inferences are accurate and meaningful for the students taking the test?" could read "is there adequate evidence to document the validity of the inferences that the test scores are accurate and meaningful for the students taking the test?" The statement "Does the evidence support that the inferences accurately reflect the specific knowledge and skills the test says it measures?" could read instead "Does the evidence support the inferences that the test accurately reflects the specific

knowledge and skills the test says it measures?" And so forth. (For further discussion see comments to page 21.)

c. **Clarifying language for questions 4 and 6:** It would probably useful to provide some additional text here (on reliability and cut scores) analogous to the additional text in paragraphs 3 and 5 (on validity and fairness), drawing from the prior pages to spell out a parallel "That is . . . ."

d. **General Point:** The purpose of this list of questions is unclear. State up front that the entity administering the test should be able to answer these questions adequately. This will clarify who has the obligation to do what.

Accuracy in Testing LEP Students and Students with Disabilities

Page 35, first full paragraph, reference to Appendix C

Suggested clarification: 1. Determine whether Appendix C is really valuable to folks.
2. If so, make some determination about whether the various accommodations were actually implemented in ways that were legal.
3. Indicate that the list of accommodations is not intended to be exhaustive, and there may be other accommodations not on the list that are, nevertheless, required by IDEA for given students.

Reason: Appendix C is referred to here and elsewhere in the document in a way that implies that this is a good list of what states do to accommodate both LEP and disabled students. Yet, the list may not be particularly useful because it neither ensures that such accommodations are actually provided in ways that are legal, nor is it a complete list of possibilities. On the other hand, Appendix C does provide a non-exhaustive list that identifies what some states are doing, and thus, encourages exchange of ideas and perhaps, effective strategies and interventions. In some ways, a shorter list of some examples might serve the purposes here better, without risking ED's use of examples from states that are not complying with the law.

Page 36-40, LEP students and accuracy

Suggested clarification: In this section, **add** language to the effect that: " It is necessary to acknowledge the interaction between limited English proficiency and other needs that affect the fairness, validity, and reliability of assessment. Significantly, students with limited English proficiency can face two risks that must be minimized in order to ensure that assessments are fair, valid, and reliable for them. One is the risk that limited English proficiency will be confused

with low academic skills or ability. The other is the risk that inadequacies in assessment will result in failing to recognize any special needs that students with limited English proficiency may have, e.g., in relation to a learning or other disability that may independently pose a barrier to their opportunities for either learning or demonstrating their skills. Students with limited English proficiency must be afforded the same fair, valid, and reliable assessments of such needs, and accommodations and modifications to meet them, to which all other students are entitled."

Page 40-45 -- Disability section --

Page 42, footnote 133 (quoting Joint Standard 10.11) under Students with Disabilities -- flagging

Suggested addition: Further explanation of Standard 10.11 is essential, as by itself, it is inconsistent with current discrimination law. Flagging is a serious issue that warrants careful attention. Further clarification and guidance consistent with principles of fairness and non-discrimination are needed as reliance on Joint Standard 10.11 has the effect of singling out those students most likely with low incidence disabilities for the system's failure to provide them with comparable opportunities to participate in assessments. Because a comparable assessment does not in all instances exist that provides necessary modifications and credible evidence of score comparability across regular and modified administrations, the score is "flagged" to provide information about the nature of the modification. Because there is an affirmative obligation to ensure that students with disabilities have access to equally effective means of demonstrating their proficiencies on the same underlying skills, and thus to develop such accommodations and modifications, it is a violation of these students' civil rights under section 504 when the SEA or school district fails to provide them such comparable opportunity to demonstrate mastery of the same knowledge and skills as they are entitled to, or fails to determine whether the accommodations or alternative assessments are comparable, then further disadvantages them by attaching a stigmatizing note on their records because of the system's failure to provide the comparable opportunity. Flagging or singling out these students is not consistent with equal opportunity, and also does not comport with the non-discrimination requirements in Section 504 of the Rehabilitation Act against coerced disclosure of disabilities.

Page 42-- 1. Tests Used for Diagnostic and Intervention Purposes

Suggested clarification: This section needs to be supplemented to address issues concerning the purpose and use of tests, including, e.g., tests used for purposes of classifying students and identifying their eligibility and need for special education - but not placement -since special education is neither a place nor a curriculum. For example, when a student is determined to have a particular type of disability that results in educational needs that purportedly can be effectively addressed by a type of specialized instruction, provision of that form of instruction is warranted if it has been proven effective in addressing the educational needs related to the

particular disability at issue. The members of the child's IEP team shall draw on the specialized knowledge of those individuals who participated in the multidisciplinary evaluation of the student and other qualified individuals to assess the appropriate selection of assessment instruments, educational planning and implementation of the child's IEP so as to ensure that the child is provided a full educational opportunity with such support services as needed to learn what other non-disabled students are expected to know and be able to do.

Page 44, first paragraph

Need for clarification: The following two sentences need clarification: "*Second, classroom instructional techniques affect large scale testing. While special educators have a long history of accommodating instruction to fit student strengths, not all the instructional practices are appropriate to large scale testing.*" It is not clear what the text here is getting at. If the point is that accommodations used in the classroom for purposes of instruction are not necessarily transferable or appropriate for students taking tests of otherwise participating in large scale assessments, the issues and concerns need to be identified with specificity and more thoroughly explored.

Page 44, at end of first paragraph

Suggested clarification: Finally, a factor that is frequently overlooked when educators and policy makers are involved in how to accommodate tests and how to use them appropriately for students with disabilities, is that the students have been denied the opportunity to learn the material for so long that this cumulative deficit poses a major barrier.

Page 44 - b. Accommodations for Students with Disabilities, reference to Appendix C

Suggested Clarification – Same point regarding Appendix C as comment to page 35**.**

Page 44, Alternate assessments.

Suggested clarification. This section is inadequate; it fails to recognize those students who cannot for a variety of reasons, including inaccessible assessment, participate in large scale academic achievement tests with accommodations, but could, if given an alternative means of demonstrating mastery, demonstrate proficiency through a different means of performance assessment. These students must be given an opportunity to demonstrate mastery over same content standards as assessed on large scale assessment but not alternate assessment - if latter is defined as measuring different content standards,

While in most instances States are recognizing accommodations that a child uses on a daily basis in the classroom, in some instances these accommodations, if provided, would have

the effect of invalidating the test or making it unreliable [inconsistent] or even an invalid measurement. If the requested accommodation/modification is rejected because it will invalidate the test as a measure of what it purports to be or on reliability grounds, the child may need to be provided an "alternative assessment" - i.e., a performance assessment that allows the child to demonstrate what he or she has mastered but using a different performance measurement. Such a performance assessment is likely to be different from the alternate assessment if, for example, the alternate assessment is intended to measure different, frequently functional skills of children who are otherwise unable to demonstrate any progress toward meeting even the most basic levels of proficiency being assessed by the large scale state or district-wide assessment.

Particularly given the Appendix's list of accommodations, the **Guidance should expressly state that whether or not an accommodation is on a State's list is not conclusive**. Rather, if the accommodation does not violate the technical properties of the assessment, is reasonable and necessary for the student to have a fair opportunity to demonstrate what she has learned, the student is entitled to the accommodation or modification under IDEA, section 504 and the ADA. The IEP team, including the parent and other qualified persons with knowledge and understanding of the student's disability (e.g., professional with expertise in an area of specific disability, teacher who works with and observes the student on a daily basis), must make any decisions about the accommodations needed by the student and about participation in the assessment, and must document their decision.

Specifically, each child's IEP [section 504 plan] must include a statement of any individual modifications (a term that remains undefined and which, in some states, means accommodations) in the administration of State or district-wide assessments of student achievement that are needed for the student to participate. In addition, the IEP must include a statement of why the assessment is not appropriate for the student if the IEP team and/or other qualified persons determines it is not, and *how* the child will otherwise be assessed.

Chapter 2. Legal Principles

Page 48 -- fn. 155, last sentence

The following sentence, while much improved, still needs revision. "On appeal the Seventh Circuit Court of Appeals stated that the appropriate remedy in this case was to require the district to use objective, non-racial criteria to assign students to classes, rather than abolishing the districts tracking system." While the sentence has been made clearer since the last draft, it will still lead some non-lawyer readers to think that the court said it would inappropriate for the school system to abolish tracking -- as distinct from saying the lower court could not require (based on the facts in that case).

Page 49 -- B. Disparate Impact -- first paragraph -- disparities not enough / equal results not required

> **Suggested Clarification:** The two sentences stating that "disparities in student performance . . . alone, do not constitute disparate impact discrimination under federal law," and that "nothing in federal law guarantees equal results" should be changed or deleted. For explanation, see our comments to similar language in the cover letter, page iv.

Page 49, paragraph 2, last line: Paragraph in setting out the test for determining if testing practices are discriminatory based on the disparate impact standard fails to reference disability. Add reference to "disability" to question raised about discrimination on basis of race, ethnicity, and sex.

Page 51, footnote 166 -- disparate impact: inseparable elements of a decision-making process

> **Suggested clarification:** Move footnote 166 into text. The footnote allows elements of a decision-making process that interact and thus cannot be separately analyzed to be analyzed together. It deserves as much attention as the point in the text that it qualifies -- the general need to isolate the policy, practice, or procedure causing the disparate impact.

Page 51-53, Determining educational necessity

> **Suggested clarification**: Separately address the two components of this analysis -- the goals or purpose and the validity of the use of the test for serving that purpose. The text and footnotes tend to merge the two in ways that are confusing. On the one hand, by not distinguishing them clearly enough, the statement that courts rarely question the educational goals (see especially text and footnotes 171-176) can be misread to apply to examination of the validity of the means for achieving those goals. On the other hand, the references to test validity, reliability and fairness could be construed to mean that these concepts are all that educational necessity boils down to, and that no examination of the adequacy of the purpose is ever appropriate.
>
> **Clarify in text as follows**: "A testing scheme can, in fact, fail either (i) because, although its purpose or goal is entirely proper, the test is not a valid means of achieving that goal; or (ii) because even though the test is a perfectly valid way of measuring the characteristics necessary to achieve the purpose, the purpose (despite the deference normally given to educators' articulation of goals) is not adequate to justify the burden disparately imposed on the basis of race, gender, or disability."

Additional text should be drafted to include:

Page 51-52, Legitimacy of purposes:

1. Importance. To meet the educational necessity standard, the educational institution must identify the importance of the purpose being served. One could have a valid and reliable measure of something that was trivial, or indeed not relevant to the function of a school, but that would not meet the educational necessity test. This goes to the underlying function of the disparate impact standard -- the obligation of government entities not to do things that increase inequality without a very important reason. The necessity standard thus needs to be talked about in those terms, before getting to the question of whether the test is a valid measure. Schools should think through the question, how essential to the mission of the school or school system is the underlying purpose for which the test is going to be used?

To give another example, in light of the state's adoption of standards for what all children are expected to know and do, there can be no sufficient educational necessity for a decision to assign many students to a low-track program that fails to teach adequately and fully those very skills and knowledge.

2. Non-discrimination. Not only must the underlying purpose must be identified and shown to be educationally necessary (giving due deference to educational judgment), it must be non-discriminatory. Just as a high-school graduation requirement to be able to lift a heavy weight would be subject to scrutiny for why it was educationally necessary, in light of its disparate impact by gender, regardless of whether the test of weight-lifting ability were valid and reliable -- so too would other standards or requirements that have disparate impact because, perhaps inadvertently, discrimination was built into the goals or standards themselves. Test users should be thinking through the question of the underlying skills, etc. that they really want to capture, and for what purpose.[8]

3. In light of (1) and (2),the statement [p. 52] needs to be examined that "[i]n evaluating educational necessity, [although] 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", "[c]ourts generally allow educational institutions to define their own educational goals and focus

---

[8]Take, for example, standards that either seek to measure, or assume, visual spatial ability. A relevant question should be whether the construct should properly be defined in terms of "visual spatial ability" or "spatial ability." (If it's really the former, then there really is no need to accommodate the sight-impaired.) While the correct answer may not be predetermined for all settings, we should recognize that the latter version has less disparate impact, and that we must be very careful not to define the constructs in a way which inherently discriminates against a group if the element that does so is in fact not essential to the underlying purpose -- particularly in terms of its exclusion of visually impaired individuals who can show that they have developed strategies that allow them to display the critical spatial abilities despite their lack of sight.

on whether the challenged test serves the institution's articulated objectives.".

4. **Relationship to publicly declared purposes.** As public institutions with statements of educational purposes and goals, dedicated to the education of the students in their charge (including goals of equal educational opportunity), issued by governing bodies (legislatures, school boards, etc.), public school systems' testing programs may be examined in relationship to those various publicly declared purposes [see comment below regarding that relationship]. This may include claims that various other public purposes of the institution, in addition to those claimed for the test, are actually being advanced or thwarted by the test. This is consistent with educational testing principles discussed earlier (comments to pp. 19-25); see *Joint Standards*, Standards 1.23 and 1.24 and page 142, concerning the need to examine unintended consequences. Similarly, in comparing other feasible alternatives with less disparate impact, the effect on achieving these other declared purposes may be considered.

Pages 52-53, Validity of the means to achieve that purpose, namely of the test use:

Insert text along the following lines:

1. "The inquiry must focus on what are the inferences being made -- in actual practice, not just as stated -- and what is the evidence that those inferences are valid? If a student is not promoted based on a test score, what is the inference being made? Is it that this student will not be able to master the material in the next grade, even with extra help? If so, what is the evidence that this score translates into that inability? And if another student is promoted based on a higher test score, what is the evidence that this higher score translates into the ability to master that next grade's material?"

2. The last paragraph in **footnote 175** -- on testing for placement and the relationship between the test and the services provided -- is very important. It should be moved up into the text and expanded. This should include a look at this and the other most common forms of high-stakes testing currently in, or coming into, use across the country, particularly in relationship to school reform initiatives:

[N.B. EACH OF THE OVERLAPPING AREAS DISCUSSED BELOW CAN AND SHOULD ALSO INFORM CHAPTER 1, OR THEY COULD BE ADDRESSED PRIMARILY IN <u>CHAPTER 1</u>, WITH SUPPLEMENTARY LANGUAGE HERE[9].]

---

[9]*Currently, the draft addresses many of them primarily in the <u>Introduction</u>, instead. For reasons discussed both in our comments to that portion and in our key issue analysis above, the treatment there of these areas and practices is not adequate. If, however, that remains the main place where they are treated, the comments here (and in the key issue analysis) should apply.*

a. Curriculum Placement (including tracking) In particular, as elsewhere in the document, it should be recognized that the existence of new state or district-wide standards for what all students should learn changes the context for examining use of assessments for *curriculum placement or tracking*. Assignment of students to low-track curricula which fails to teach adequately and fully those skills and knowledge or provides less intensive opportunity to master them cannot generally meet a standard of *educational necessity* when the state or district has said all students should master those skills -- even if the assessment used to make the decision has some correlation with either past or future achievement.

Even were disproportionate assignment of some groups of students to a curriculum that is less intensive, less connected to the higher level skills and knowledge reflected in standards or otherwise inferior not a problem, other issues would still require attention -- including the accuracy of the assessment outcome in predicting the likely education effects of different treatments, the validity and reliability problems in exclusive reliance on test scores, and in some cases the relying on a test in one subject (typically reading) as the basis for assignment in other subjects.

b. Adequacy of instruction. Simultaneous use for school improvement/accountability and high-stakes student decisions. One of the biggest issues now, in the context of school reform, is the *relationship between* use of individual assessment scores for *high-stakes student promotion and graduation decisions and* use of aggregate scores for *school improvement and accountability*. In particular, in some systems, the same test is being used at the same time in the same schools for both purposes in ways that bring into question the validity of the inferences being made in using the tests. On the one hand, the use of the assessment to trigger improvements and interventions in the school is premised on the notion that a low aggregate score indicates that the school is failing to teach adequately the skills and knowledge reflected in the state and district standards. On the other hand, the use of the assessment to impose high-stakes consequences on students must be premised on the notion that the students *have* been adequately taught those skills and knowledge (as indicated in the last paragraph). Thus, not only is it the case that an assessment that is valid for purposes of providing information about a school's need for improvement may not be valid for high-stakes student consequences, in fact to the degree that the assessment results indicate inadequacies in curriculum and instruction they provide evidence of the impropriety of using the test for high-stakes student purposes until those inadequacies are corrected.[10]

c. The issue of whether students have adequately been taught the skills and knowledge which

[10]In light of the discussion here and elsewhere in the document and our comments, the statement at the end of the top paragraph on page 53, that "courts *may* also consider whether the skills testing have been taught in the program" should be **deleted or changed** substantially -- falsely conveying the notion that this is an optional inquiry when students are denied promotion or graduation.

the test is designed to measure -- instructional aspects of validity -- deserve attention in this civil rights, disparate impact analysis. (Although the focus on adequacy of instruction is more visible in the due process section, it needs highlighting in this section as well.) Each of the inferences about the adequacy of instruction for all students, in relationship to the identified skills and knowledge -- for example, regarding alignment of the curriculum, qualifications of teachers to teach them effectively to all students, efficacy of instructional methods used, efficacy of systems of assistance for identifying and assisting students having difficulties in particular skill areas, etc. -- need to be articulated and then the evidence for that inference needs to be examined.

d. Under civil rights law analysis, failure to follow relevant requirements of other laws in the programs in which the groups disproportionately subject to the high-stakes sanctions are enrolled, should be viewed as *relevant evidence of lack of sufficient justification for the disparate treatment*. For example, Title I (along with some state mandates) requires enriched, accelerated curriculum aligned with high standards, effective instructional techniques, highly qualified teaching staff, and timely effective intervention for individual students having difficulty mastering particular standards -- with the plan for doing so jointly developed with parents.

e. Promotion. In addition to the discussion above on the relationship to school improvement, in examining the validity arguments underlying use of tests for promotion purposes, attention should be paid to the distinction between promotion based on mastery of the grade below versus readiness for the grade that follows -- the identity of the two cannot be assumed. Additional issues here include the need to validate the actual decision being made (1), the reliability of individual scores, (see our comment to page 26, above), and the discussion of cut-off scores (see our comment to page 31, first new paragraph, above).

f. In examining the validity of the *use of a test* in terms of the actual decisions being made and the evidence for the inferences that underlie them, it is important to look at the significance and use of *exit exams and diplomas*. Employers are increasingly being encouraged to look at students' school credentials, including diplomas, scores and grades, and state postsecondary institutions may use them for admission, etc. There is, however, an obligation *first to validate the information for that purpose*. In fact, the seminal case in this area - *Griggs* v. *Duke Power Co.* -- was specifically on that topic: the improper use of high-school diplomas as an employment criteria without showing that the presence or absence of a diploma had been validated as capturing the presence or absence skills or knowledge needed in the particular jobs.

g. Differential diplomas. In some cases, test performance is not being used to deny diplomas altogether but to determine the kind of diploma a student receives. To the extent that these differences affect important opportunities such as admission to postsecondary institutions or eligibility for financial aid, the same criteria for validity and reliability apply.

Page 56, second new paragraph

Suggested Clarification: State that under Title I, schools must include LEP students in the assessment. (20 U.S.C. § (b)(3)(F)).

Reason: Whenever possible, the Department should underscore the basic obligation of education agencies to assess all students. The specifics regarding accommodation can follow, but the first point should be "you have to include all kids."

Page 57-58 Performance assessment re: children who need alternative way of demonstrating what they know and are able to do - see discussion concerning accommodations and alternative assessments at our comments to page 16 and to page 44.

Page 58, first paragraph, first sentence -- on prerequisite skills

Clarification: Need to clarify and circumscribe the meaning of "functions . ~. . necessary for participation in the program." ["When tests are used for other purposes, such as making decisions about placement in gifted and talented programs, it is important that tests measure the skills and abilities needed in the program, rather than the disability, unless the test purports to measure skills or functions which are impaired by the disability and such functions are necessary for participation in the program."]

Reason: Even if such *functions* are impaired by the disability and are necessary for participation in the program, this cannot be the basis for exclusion if the student could function in the program with appropriate accommodations, supplementary aids and services, special education and related services, assistive technology services or devices, etc. so that with such assistance or accommodation s/he could either acquire those skills simultaneously or function in the program without them.

Page 58 (second paragraph) - page 59 (first paragraph) -- participation in state- or district-wide assessment programs

Suggested clarification: {See comments above to page 16; criteria for the individual determination} Inclusion in the general accountability system is important as a matter of civil rights. Once a system of accountability is established for all children, as a matter of law, children with disabilities must be included. This is required by Goals 2000 and Title I as well as under section 504 of the Rehabilitation Act and the Americans with Disabilities Act that prohibit discrimination on the basis of disability. In addition, section 1412(a)(17) of IDEA, as amended, requires children with disabilities to be included in any general State and district-wide

assessment, with appropriate accommodations and modifications, if necessary. Assuming the state or district has adopted a high stakes system, students with disabilities are expected to participate. However, such a determination shall be on an individual basis by each youth's IEP team, which in the case of high stakes testing, must consider whether or not the student has been taught the material consistent with his/her individual due process rights.

States and local educational agencies (as appropriate) must develop guidelines for participation of those children with disabilities in alternate assessments who cannot participate in State and district-wide assessment programs. Alternate assessments need to be conducted by July 1, 2000. For a child who, for example, because of severe cognitive disability, is unable to participate in the regular assessment, the child's progress, however significantly modified, must be assessed -- even if the so-called alternate assessment has not yet been developed. Assessments serve many purposes -- whether we are describing a regular assessment or an assessment designed to assess a child who requires an alternate assessment. SEAs/LEAs are now obligated to conduct alternate assessment for those children unable to participate in the regular assessment; they are required to find ways to provide them benefits comparable to their non-disabled peers, including, e.g., by designing a means to monitor the child's progress, hold schools and teachers accountable, certify student skills and capabilities, possibly achieve improved alignment of the modified curriculum that the child is receiving, instruction, and assessment; and informing curriculum decisions and instructional practice.

Furthermore, if a child with a disability is unable to participate in the general assessment, e.g., because the child is medically fragile or has such limited cognitive ability that the child cannot demonstrate any progress toward even minimal levels of proficiency on the general assessment with accommodations and modifications, the SEA/LEA has an obligation to ensure that the child participates in an alternate assessment. In this scenario, we are usually talking about an assessment that measures different content than the general large-scale assessment. On the other hand, if a child with a disability is unable to demonstrate his level of mastery or proficiency using the general assessment even with modifications and accommodations, but is able to demonstrate some progress toward proficiencies using a different performance measurement, the child must be provided with a different type of assessment - in all likelihood a performance assessment. SEA/LEAs are required to have alternate assessments established by July 2000, all students' IEPs must include a provision as to how these students will be assessed in the interim.

Page 62, after first paragraph -- due process analysis

Clarification - Missing Step: After identifying the purpose of the testing program and determining that it is legitimate, there is a critical next question [either as a new paragraph (2) or as a subpart of a redefined paragraph (1)]: Is the testing program reasonably related to that

purpose? That is, does the test measure what it purports to measure? Just as instructional aspects of validity are posed in relationship to the third question ("Are students actually taught the knowledge and skills measured by the test?"), issues of construct validity (see pages 20-24) as well as reliability are implicated in answering this question. See also our comments to the due process discussion on pages 17-18 of the draft, above. (In a sense, the need to address more distinctly the legitimacy of the means for achieving the purpose from the legitimacy of the purpose is analogous to the need to address those questions more distinctly in the Title VI educational necessity analysis. See comments to pages 51-53, above.)

Page 63, question (3) (and pages 17-18) -- Are students actually taught the knowledge and skills measured by the test?"

Suggested clarification: Regarding alignment of the skills and knowledge being tested with actual instruction, it is worth noting developments since *Debra P.* that further shape the basic analysis. First, the existence in each state of standards for what all students should know and be able to do, as a central feature of the system, and (as with Title I noted above) certain obligations concerning the alignment of curriculum, instruction, and assistance with those standards helps to sharpen the analysis of what is adequacy (and also to rule out certain kinds of arguments that are premised on assumptions that contrary to the notion that all students can achieve at these levels). Second, there is a much stronger body of research now available concerning the kinds of instructional practices that result in high student achievement and the institutional practices that in turn support those instructional practices.

Page 63, question (3), last sentence

Suggested change: In the last sentence, delete the phrase ", but may not expect proof that every student has received has received the relevant instruction."

Reason. In the sentence, "In cases examining system-wide administration of a test, courts require evidence that the content covered by the test is actually taught, but may not expect proof that every student has received the relevant instruction," the last phrase still remains confusing in that it seems to imply that these are not individual rights -- e.g. that a student who was not given the relevant instruction may not have a claim so long as most of the students in her class DID receive the relevant instruction. Due process is indeed a right belonging to the individual and calling for fair treatment of that individual.

Appendix A: Glossary of Legal Terms

Page 65, "Educational necessity"

Clarification: The last sentence in the definition needs to be modified, for reasons discussed previously (see comments to page 51-53). After properly saying that "educational necessity generally refers to a showing that practices or procedures are necessary to meet an important educational goal," the next and final sentence says, "In the context of testing this means the test or assessment procedure must be valid and reliable for [the] purpose for which [it] is being used." As discussed earlier, the former sentence cannot be reduced to the latter. The focus on validity and reliability is an essential element, but not sufficient -- there is the independent question of whether that purpose is an important educational goal, and even with due deference to educators' judgments in setting goals, they must, particularly in light of public policy on equal educational opportunity and civil rights, be framed non-discriminatorily and be of sufficient importance in light of any disparate impact.

Cross-cutting point. Finally, in a variety of places in the document, terms such as "where feasible" are used. While we suspect that this term and others like it are generally intended to create only a narrow exception to the general standard at issue (e.g., where there simply is no way to develop an adequate sample size), we strongly believe that the language will get misconstrued and applied overly broadly. In the educational settings with which we are most familiar, such terms are often (and correctly) perceived to be an indication that the standard which they accompany can largely be safely ignored. Therefore, we suggest that, in the various places where such a term appears, you either delete it or give it a more precise and narrow construction. (Some, but not all, such instances are cited in our specific comments.)