APPLIED MEASUREMENT IN EDUCATION, *13*(4), 343–385

GI Forum v. Texas Education Agency: Psychometric Evidence[1]

S. E. Phillips
Consultant

The *GI Forum v. Texas Education Agency* court upheld the Texas graduation test against discrimination and due process challenges by Hispanic and African-American students. The court found that the test was valid, reliable, and met applicable professional standards, including notice and opportunity to learn, and that the test was being used to identify and remedy educational inequities. Based on evidence introduced at the trial and the court's decision, this article discusses the plaintiffs' arguments, the state's responses, and the specific findings of the court in the context of the following major themes of the litigation: *Debra P.* requirements, professional standards, historical misuses of tests, validity, reliability, opportunity to learn, setting passing standards, adverse impact, use of a single test score, conjunctive versus compensatory models, differential item performance, item discrimination, factor analyses, and dropout/retention rates. The article concludes with a summary of guidance for graduation testing programs and lessons learned from the case.

BACKGROUND

The *GI Forum* challenge to the Texas high school graduation test[2] alleged racial discrimination in violation of federal Title VI and Equal Education Opportunities Act statutes, Title VI Regulations, the equal protection and due process clauses of the U.S. Constitution under Section 1983 of the Civil Rights Act of 1964, and the federal court order in the *U.S. v. Texas* desegregation case.[3] In response to a motion for summary judgment filed by the defendants, the judge dismissed the claims based on the Title VI statute (failure to show intent to discriminate), Equal Educational Opportunity (lack of a Spanish version of the

Requests for reprints should be sent to S. E. Phillips, Box 384, West Paducah, KY 42086.
Editor's Note. This article does not follow standard journal style. References are given as endnotes.

test is not a barrier to participation in instructional programs), equal protection (failure to show intent to discriminate), and *U.S. v. Texas* (denial of opportunity to graduate not a denial of "educational opportunities"; lack of jurisdiction) and allowed the Title VI Regulations and constitutional due process claims to proceed to trial.[4] In its final decision, the court summarized the case as follows: "The issue before the Court is whether the use of the Texas [graduation test] as a requirement for high school graduation unfairly discriminates against Texas minority students or violates their right to due process."[5]

Due Process

The court reaffirmed the *Debra P.* court's holding that a high school diploma is a property interest "created by the requirement of compulsory education, attendance requirements, and the statute detailing graduation requirements."[6] A *property interest* is a threshold requirement for a due process claim. The factual issue in dispute under the due process claim was the validity of the graduation test; that is, whether the implementation and use of the graduation test was a substantial departure from accepted professional standards, specifically including fairness and opportunity to learn.

Title VI Regulations

Under Title VI Regulations, plaintiffs have an initial burden of demonstrating adverse impact. The burden then shifts to the defendants to show that the test is necessary to achieve an important educational goal. If the defendants satisfy this burden, the burden shifts back to the plaintiffs to show that an equally effective but less discriminatory alternative exists. The court found that there were issues of material fact related to all three areas: adverse impact, educational necessity, and less discriminatory alternatives. Relative to the adverse impact issue, the court stated, "Unfortunately, there is not a clear consensus on what type of statistical analysis is to be used in cases in which racial discrimination is asserted."[7]

The Requested Remedy

The plaintiffs asked the court to issue an injunction prohibiting the state from using the graduation test to award diplomas and requiring the school districts of the named plaintiffs to issue their diplomas. In its final decision, the court denied both requests, stating, "the Court has determined that the use of the [graduation test] does not have an impermissible adverse impact on Texas's minority students and does not violate their right to the due process of law."[8] In reaching this decision, the court recognized that the law requires courts to give deference

to state legislative policy, particularly in an educational context, and that the only controlling and directly on point case was *Debra P.* The plaintiffs elected not to appeal this decision.

Relationship to the *Debra P.* Case

Although the plaintiffs argued for similarity, the situation in Texas when its graduation test was implemented was distinguishable from that of Florida at the time of the *Debra P.* case. Texas had a state-mandated curriculum; Florida did not. Unlike the Florida students in the *Debra P.* case, African-American and Hispanic minority students subject to the graduation test requirement in Texas had not been required by statute to attend segregated schools. Moreover, graduation testing was not a new concept in Texas as it had been in Florida. At the time the *GI Forum* case was filed in 1997, high school graduation tests had been in existence for nearly two decades nationwide and for a decade in Texas, beginning with the challenged test's predecessor implemented in 1985.[9]

Consistent with the ruling in the *Debra P.* case, the *GI Forum* defendants asserted that even if there had been prior discriminatory conduct by some educators in Texas, the graduation test would help to remedy any potential vestiges of past discrimination. In upholding the Florida graduation test once Florida high school students had all been educated in unitary schools and the state had demonstrated the test's curricular validity, the *Debra P.* appeals court held:

> We affirm the district court's findings (1) that students were actually taught test skills, (2) that vestiges of past intentional segregation do not cause the [test's] disproportionate impact on blacks, and (3) that use of the [test] as a diploma sanction will help remedy the vestiges of past segregation. Therefore, the State of Florida may deny diplomas to students.[10]

The Evidence

At trial, the plaintiffs presented a variety of psychometric and statistical arguments related to the quality of the graduation test and its impact on African-American and Hispanic students. These arguments focused on the following areas of contention: historical misuses of tests, validity, reliability, opportunity to learn, setting passing standards, adverse impact, use of a single test score, conjunctive versus compensatory models, differential item performance, item discrimination, factor analyses, and dropout/retention rates. Each major area of contention is discussed in a separate section. The information presented includes a summary of the evidence offered to the court, a description of relevant psychometric standards and procedures, and a critique of the proffered arguments.[11] But first a word about professional standards.

PROFESSIONAL STANDARDS

Professional standards assumed a central role in the trial debate about the psychometric quality of the graduation test. Specific standards from the 1985 American Educational Research Association, American Psychological Association, National Council on Measurement in Education *Test Standards*,[12] applicable at the time the graduation test was developed and implemented, were cited by expert witnesses and their interpretations were debated. Consideration of introductory material appearing with the standards supported the appropriateness of professional judgment in interpreting the standards:

> Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every primary standard in this document, and acceptability cannot be determined by using a checklist. Specific circumstances affect the importance of individual standards. Individual standards should not be considered in isolation. Therefore, evaluating acceptability involves the following: professional judgment that is based on a knowledge of behavioral science, psychometrics, and the professional field to which the tests apply; the degree to which the intent of this document has been satisfied by the test developer and user; the alternatives that are readily available; and research and experiential evidence regarding feasibility. (p. 2)

In the Preface to the *Test Standards*, the Development Committee stated several guidelines that governed the work of the committee:

> The *Standards* should … Be a statement of technical standards for sound professional practice and *not a social action prescription*. … Make it possible to determine the technical adequacy of a test, … and the reasonableness of inferences based on the test results." (p. v, emphasis added)

The *Test Standards* are divided into three categories: primary, secondary, and conditional.

> *Primary standards* are those that should be met by all tests … absent a sound professional reason [to the contrary]. … *Secondary standards* are desirable as goals but are likely to be beyond reasonable expectation in many situations. … Test developers and users are not expected to be able to explain why secondary standards have not been met. (pp. 2–3)

Conditional standards vary in importance depending on the application.

In addition to urging the court to adopt their interpretation of particular standards, the plaintiffs in the *GI Forum* case also urged the court to mandate adherence to secondary and conditional standards. The specifics of these arguments are discussed in relevant sections that follow.

HISTORICAL MISUSES OF TESTS

In its pretrial order, the court eliminated the history of discrimination from consideration in the lawsuit. Nonetheless, the plaintiffs were allowed to present limited evidence alleging discrimination occurring concurrently with the implementation of the graduation test in 1990.

Plaintiffs' Views

The plaintiffs' arguments in this area were primarily ones of guilt by association. They argued that ability tests, such as IQ tests and college entrance examinations, had in the past been inappropriately interpreted as indicating fixed cognitive characteristics for which lower scores by minorities demonstrated inferior capabilities. Similarly, such lowered expectations were postulated for the graduation test and asserted to result in discrimination against minority students.

Defendants' Views

The state defendants countered this line of reasoning by asserting that historical misuses of tests, although unfortunate, lacked relevance to the graduation test. They asserted that the graduation test was an achievement test that measured teachable academic skills, not an intelligence test. They explained that the graduation test identified students who had not learned these skills, and students so identified were provided remedial instruction. Thus, graduation test scores were not fixed but changed over time as students received additional instruction and learned the tested skills. Even though an unsuccessful first attempt to pass the graduation test could be discouraging, defendants' witnesses declared that minority students would be even more disadvantaged if they received high school diplomas without having their skill deficiencies identified and remediated.

The Court's View

> In reviewing the diverse cases that underpin this decision, the Court has had to acknowledge what the Defendants have argued throughout trial—this case is, in some important ways, different from those cases relied upon by the Plaintiffs. … [T]his case asks the Court to consider a standardized test that measures knowledge rather than one that predicts performance. … [T]he TEA's evidence that the implementation of the [graduation test], together with school accountability and mandated remedial follow-up, helps address the effects of any prior discrimination and remaining inequities in the system is both credible and persuasive. (pp. 4, 14)

VALIDITY

Validity refers to the weight of accumulated evidence supporting a particular use of test scores. For the graduation test, scores are used to decide whether students have attained sufficient academic skills in the subject areas of reading, mathematics, and writing for the award of a high school diploma. The most important evidence of validity in this situation is a measure of the degree to which the items on each subject-matter test measure the knowledge and skills prescribed by the state-mandated curriculum. Although this type of validity evidence could be classified under the umbrella term of *construct validity,* in achievement testing applications it is usually referred to as *content validity evidence.* Primary Standards 1.6 and 8.4 described the documentation expected for content validity. The documentation for the graduation test appeared in the *1996–97 Technical Digest* and is summarized in a companion article in this issue.[13]

Content Validity Evidence

Recognizing the importance of the *Test Standards,* the Texas State Board of Education had specified in its 1995–96 Administrative Code: "The commissioner of education shall ensure that each [test developed according to state statute] meets accepted standards for educational testing." (§ 101.1(c)). The Texas Education Code also provided: "The State Board of Education by rule shall establish the essential skills and knowledge that all students should learn" (§ 39.021). Representative committees of Texas educators, business representatives, parents, and the public participated in the establishment of the state Essential Elements tested by the graduation test. By law, all Texas public schools were required to teach this content and to provide remediation to unsuccessful students. The Essential Elements and corresponding state objectives had been widely disseminated to Texas educators, students, parents, and the public.

Determining the content and skills to be mandated for all students in a state is a political decision for which input from a variety of stakeholder groups is desirable. Once the required content objectives and skills have been delineated, psychometric procedures can be applied to evaluate the content validity of the resulting graduation test. Prefatory comments in the *Test Standards* and Conditional Standard 1.7 described relevant documentation for this stage of content validation.

Content validity evidence is typically obtained by professional judgment. Content experts are asked to review each potential test item and classify it according to the objective or skill being measured, check the correctness of the keyed answer, check for ambiguities in wording and other item flaws, evaluate the appropriateness of the content and difficulty for the intended grade level, and identify any in-

appropriate or potentially offensive language or content. Committees of Texas educators performed these functions for all items written for the graduation test.[14] The committees of Texas educators that reviewed the graduation test items were chosen to be representative of the state in terms of geography, size of district, gender, and ethnicity. In addition, each committee member was knowledgeable about the grade level and subject matter being tested. Committee members were trained by the contractor and Texas Education Agency (TEA) staff prior to beginning their reviews. A second review of the items and their statistical data was conducted after field testing.

Criterion and Construct Validity Evidence

Other types of validity evidence have been referred to as criterion and construct validity evidence. Criterion validity evidence, in the form of correlation coefficients, is most appropriate for situations in which test scores are used to predict outcomes such as freshman grade point averages (GPAs). It can also be useful in determining the degree to which two tests measure the same or different skills. Construct validity evidence refers to the sum of research knowledge and experiments designed to define a psychological construct, such as extroversion or locus of control, that an instrument is intended to measure.

Plaintiffs' Views

With respect to the graduation test, the plaintiffs argued that it was not sufficient for the state defendants to demonstrate the content validity of the test. They argued that criterion and construct validity evidence must also be provided. One expert opined,

> Some of the [graduation test] items are irrelevant to any real work or world applications of mathematics, reading and writing, because they have been made "tricky" to lower the proportion (*p*-value) of people who pass these items. Items that are artificially made difficult, for example, by introducing irrelevant information, are items that a disproportionately greater number of disadvantaged and language-minority students are likely to get wrong.[15]

More specifically, in support of a claim of graduation test invalidity, plaintiffs presented the court with evidence of moderate correlations of graduation test scores with sophomore English grades obtained from a volunteer sample from three Texas districts ($N = 3{,}200$). Plaintiffs argued that correlations of .32 to .37 between 10th-grade graduation test subscores and English II grades indicated a lack of criterion-related and curricular validity.

Defendants' Views

Defendants urged that in addition to being a nonrepresentative, small sample of Texas students, the moderate correlations of test subscores with grades could be explained by factors other than lack of test validity. For example, the graduation test was designed to measure eighth- and ninth-grade skill levels. Although sophomore classes might have included some review of prior learning, they probably also included substantial content not tested by the graduation test. Furthermore, grades given by teachers in high school courses may have been based in part on factors besides achievement of skills. These factors, with varying weights assigned to them across teachers and subjects, included attitude, improvement, attendance, and effort. Thus, defendants opined that the graduation subtests and high school grades measured different student characteristics and that grades should not be viewed as substitute measures of tested content.

Subtest Intercorrelations

The plaintiffs also concluded that subtest intercorrelations from this volunteer sample

> cast doubt on the validity and reliability of [graduation test] scores [because the data were] contrary to the expectation that scores of two verbal measures (of reading and writing) should correlate more highly with one another than with a measure of quantitative skills.[16]

When defendants recalculated these subtest intercorrelations (writing/reading = .50; writing/math = .51; reading/math = .69) using the population of all Texas 10th graders tested for the first time that year (N = 210,000), the pattern was more consistent with expectations for achievement tests (writing/reading = .56; writing/math = .48; reading/math = .62). The writing subtest correlated more highly with reading than with mathematics and the mathematics subtest correlated only slightly higher with reading than did the writing subtest. The latter result may have been due to the format similarity of the multiple-choice reading and mathematics subtests as compared to the writing subtest, which was composed of multiple-choice items and a direct writing prompt. The dissimilarity of population and sample results supported defendants' original assertion that the convenience sample relied on by plaintiffs was not representative of the state population.

Predictive Validity

The plaintiffs also analogized the Texas graduation test to ability tests like the college admissions Scholastic Assessment Test (SAT) and argued that standards ap-

propriate for the SAT should also be applied to the graduation test. For example, they discussed predictive validity coefficients provided for the SAT and argued that similar statistics should have been calculated for the graduation test. Defendants responded that such correlations were not required because the graduation test (as an achievement test) was not designed or intended for prediction.

In sum, because the graduation test was intended to measure state-specific content knowledge/skills, and not to predict any other outcome, and because the graduation test was designed to measure specific academic content, not to define more general psychological constructs, defendants argued that criterion and construct validity evidence were not required to demonstrate that the graduation test met professional standards for validity. Defendants' experts testified that the content validity evidence available for the graduation test satisfied the *Test Standards.*

Criterion-Related Validity Data

Even though the defendants believed that criterion-related validity evidence was not required to establish the validity of the graduation test, several types of research data were collected that quantified the relationship between test scores and other variables such as courses taken, grades, and eighth-grade test scores.

Common sense suggested that students who began high school with adequate prerequisite skills and took more academic courses were more likely to pass the graduation test. Data collected by the state supported these relationships. For the 1995 10th-grade cohort, 80% of African-American, 83% of Hispanic, and 90% of White students who had passed all Grade 8 tests in 1993 passed the graduation test. This relationship was not perfect, the defendants maintained, because attainment of prerequisite skills did not guarantee that students would learn new material at a satisfactory level, nor did lack of attainment preclude remediation and future success.

For a subset of the Spring 1995 Grade 10 students, the passing rates increased for each higher level math course taken. Students receiving credit for Algebra II had the highest mathematics passing rates, whereas those receiving credit for Pre-algebra had the lowest passing rates. Minority students who received credit for Algebra II passed the mathematics subtest at a rate four to five times higher than those only receiving credit for Pre-algebra.

However, minority students took significantly fewer advanced math courses than majority students. Although the percentage of African-American and Hispanic students receiving credit for Algebra I was similar to that for White students, the percentage of White students receiving credit for Geometry and Algebra II was, respectively, 1.5 and 2.0 times that for African-American and Hispanic students.

The defendants emphasized two facts about the data depicting a relationship between mathematics courses taken and mathematics graduation subtest performance. First, advanced math courses were not required to pass the mathematics subtest, which tested content through eighth-grade math. The higher passing rates

for students receiving credit for more advanced mathematics courses may have been due to instructional reinforcement of prerequisite lower level content in the higher level courses or completion of these courses by students with higher levels of achievement in the lower level courses. Second, Algebra I is now a requirement for high school graduation for all students in Texas.

Testing Linguistic Minorities

Primary Standard 13.1 stated that tests administered to nonnative English speakers should be designed to minimize threats to reliability and validity related to English language proficiency. The comment to this standard stated: "Careful professional judgment is required to determine when language differences are relevant" (p. 74).

Plaintiffs argued that requiring Hispanic students with limited English proficiency (LEP) to pass the graduation test was discriminatory because the test measured both content knowledge/skills and English language proficiency. Defendants argued that the purpose of the graduation test was to measure achievement of reading, mathematics, and writing skills in English. They further argued that there was no discrimination because native English speakers with poor language skills were also required to demonstrate skill mastery in English. The appropriateness of holding LEP students to the same graduation standards as non-LEP students is discussed in a companion article in this issue.[17]

The Court's View

The court made extensive findings of fact about graduation test construction consistent with the defendants' trial testimony and evidence presented in the *Technical Digest.* Finding the defendants' claims of test invalidity based on technical considerations and broader educational factors unpersuasive, the Court stated:

> [T]he [graduation test] meets currently accepted standards for curricular (sic) validity. In other words, the test measures what it purports to measure, and it does so with a sufficient degree of reliability.
>
> …
>
> The Court also finds that the Plaintiffs have not demonstrated that the [graduation test] is a substantial departure from accepted academic norms or is based on a failure to exercise professional judgment. … Educators and test-designers testified that the design and the use of the test was within accepted norms. … In addition, the State need not equate its test on the basis of standards it rejects, such as subjective teacher evaluations.
>
> In short, the Court finds, on the basis of the evidence presented at trial, that the disparities in test scores do not result from flaws in the test or in the way it is adminis-

> tered. Instead, as the Plaintiffs themselves have argued, some minority students have, for a myriad of reasons, failed to keep up (or catch up) with their majority counterparts. It may be, as the [state] argues, that the [graduation test] is one weapon in the fight to remedy this problem. At any rate, the State is within its power to choose this remedy. (pp. 29–30, citations omitted)

RELIABILITY

Reliability is an indicator of consistency of measurement. Errors of measurement are minimized and decision consistency is maximized by a reliable test. Reliability is a necessary but not sufficient condition for validity. There are two major procedures for calculating test reliability: repeat testing and measures based on a single test administration. The *Test Standards* did not specify any particular type of reliability estimate as mandated or preferred. Rather, Primary Standard 2.1 stated: "[E]stimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test" (p. 20). Primary Standards 2.2 and 2.3 specified documentation of sample characteristics and methods used to calculate reported reliability estimates.

Conditional Standard 2.9 recommended separate reliability estimates for subpopulations and Secondary Standard 2.10 recommended reporting standard errors at the cut score. For the graduation test, KR_{20} reliabilities and total test standard errors of measurement were reported by subgroup for each subtest. Standard errors at the cut scores were available from the contractor but were not published in the *Technical Digest.* The standard errors at the cut scores were approximately equal to the overall standard errors for their respective subtests. For the 1997 spring administration of the graduation test to first-time test takers (200,000 total 10th graders, including 27,000 African-Americans and 68,000 Hispanics), the KR_{20} reliabilities were slightly higher for minority subgroups than for the majority subgroup.[18]

Plaintiffs' Views

The plaintiffs argued that the KR_{20} estimates were insufficient and that test–retest and alternate forms reliability estimates should also have been provided: "[W]hile the technical report bases the calculation of standard error of measurement on internal consistency reliability estimates, they should have been based on test–retest or alternate forms reliability estimates."[19] This opinion was based in part on a quote from a measurement textbook that stated: "[E]vidence based on equivalent test forms should usually be given the most weight in evaluating the reliability of a

test."[20] The plaintiffs pointed out that if test reliability were reduced from .90 to .85 by using alternate forms rather than KR_{20} estimates, the standard error for a test with a standard deviation of 6.3 would increase from 2.0 to 2.4. The plaintiffs also presented the court with multiple measures of reliability for the Grade 5 Iowa Tests of Basic Skills from a table in a measurement textbook as evidence that the KR_{20} reliability coefficient had been chosen because it would produce the largest reliability estimates.[21]

Defendants' Views

The defendants noted that in a subsequent paragraph of the measurement text cited by the plaintiffs, the author stated: "Practical considerations often dictate that a reliability estimate be obtained from a single administration of the test." Regarding test–retest reliabilities for a reading comprehension test, this text also stated that "[students'] answers the second time will involve merely remembering what answers [the students] had chosen the first time and marking them again."[22] Furthermore, the defendants argued that alternate forms testing was impractical for the graduation test for several reasons: (a) Decreased student motivation on a second testing that does not count alters performance; (b) schools are unwilling to devote additional instructional time to unnecessary double testing of students; (c) naturally occurring repeat testing takes place after remediation, creating an expectation of changed scores; and (d) unlike achievement test batteries where students may take different forms at a single administration, all students take the same form of the graduation test at a single administration. Parallel forms for subsequent administrations are equated to a common scale to maintain a consistent passing standard.

If the state could have convinced several districts to do so, alternate forms under unmotivated conditions could have been administered to the same groups of students. However, the defendants indicated that in their view the limited utility of this data for generalizing to test administrations that "count" would be insufficient to justify the substantial costs of collecting such data.

The Preface to the *Test Standards* stated, "The *Standards* is not meant to prescribe the use of specific statistical methods. Where specific statistical reporting requirements are mentioned, the phrase 'or equivalent' should always be understood" (p. 2). One way to compute reliability for alternate forms of a single-administration test is to split the test into two parallel halves. The KR_{20} reliability estimate is an average of all such possible splits so it includes errors related to item sampling. Thus, KR_{20}s provide an estimate of alternate forms administered concurrently. Because students in Texas are expected to continue receiving instruction between test administrations, it would be inaccurate to obtain alternate forms reliability estimates based on the administration of test forms separated by a time period. Indeed, the prefatory material in the chapter on reliability in the *Test Standards* stated: "Differences between scores from … one occasion to another …

are not attributable to errors of measurement if maturation, intervention, or some other event has made these differences meaningful. …" (p. 19).

The Court's View

The court held that "the [graduation test] measures what it purports to measure, and it does so with a sufficient degree of reliability" (p. 29, citations omitted).

OPPORTUNITY TO LEARN

In the *Debra P.* case, the court instituted two additional requirements for graduation tests: notice and curricular validity. The curricular validity requirement, also referred to as opportunity to learn, was included as Standard 8.7 in the 1985 revision of the *Test Standards.*

Notice

Notice requires the state to disseminate information about graduation test requirements to all affected students well in advance of implementation. This responsibility is codified in the *Texas Administrative Code* as follows:

> The superintendent of each school district shall be responsible for the following: (1) notifying each student and his or her parent or guardian in writing no later than the beginning of the student's seventh grade year of the essential skills and knowledge to be measured on the … [graduation] tests administered under the [Texas Education Code]; (2) notifying each 7th–12th grade student new to the district of the testing requirements for graduation, including the essential skills and knowledge to be measured; and (3) notifying each student required to take the … [graduation] tests and out-of-school individuals of the dates, times, and locations of testing. (§ 101.2(a))

The notification provided to students and their parents occurs more than 3 years before the first graduation tests are administered in the spring of the 10th grade and more than 5 years prior to the expected graduation of these students in the spring of the 12th grade.

Opportunity to Learn

Opportunity to learn (OTL) means that students must be taught the skills tested on a graduation test. In practice, evidence of OTL is often gathered by examining the official curricular materials used in instruction and by surveying teachers to determine whether they are teaching the tested content.

Plaintiffs' Views

The plaintiffs argued that Texas should have collected exactly the same data as Florida presented to the court in the *Debra P.* case, which included formal surveys of districts' curricular materials, teachers, and students. They also argued that differential passing rates, moderate subtest/grade correlations, and few predominantly minority schools with exemplary and recognized accountability ratings demonstrated lack of OTL for minority students.

The evidence presented by the plaintiffs included a longitudinal study of the readability of passages included in the graduation reading tests. The passage readability estimates presented in the study indicated that the passages included in recent tests were easier than those included in earlier tests.

The plaintiffs also offered the results of a "Testing and Teaching survey" designed "to obtain the opinions of a representative sample of [Texas secondary teachers] about the relationships between mandated testing and teaching and the effects of mandated testing."[23] These results were based on a 15% response rate from 1,000 Texas secondary teachers randomly sampled from a sample of 4,000 such teachers obtained by randomly sampling from a list compiled by a market data firm in Connecticut. The survey questions asked respondents about mandated testing in general. In addition to responses demonstrating a perceived ineffectiveness of mandated testing, the plaintiffs also presented the court with examples of specific negative comments made by individual teachers selected from a compilation of positive and negative comments from survey respondents. No data on follow-up of nonrespondents' opinions were reported.

Defendants' Views

Defendants responded that Texas had implemented an equivalent OTL procedure as allowed by the *Test Standards*. For the graduation test, OTL was established through the state-mandated curriculum, surveys of teachers and curricular materials for the prior graduation test based on the same curriculum, and adequacy of preparation reviews by Texas educator committees and separate bias review panels. Furthermore, district surveys had been conducted for the prior graduation test based on the same mandated state curriculum.

The testimony by named minority plaintiffs in the case was also cited. Most had received average or below-average grades in academic subjects, had failed one or more academic courses, were interested in nonacademic pursuits, and chose not to participate in remediation options offered by their schools. They also acknowledged that they had sat next to majority and minority students in their classes who had passed the graduation test. Some had made up incomplete course work or completed remediation and passed the test after their scheduled graduation. Those

who had not said they were too busy with work or family obligations. When asked under oath if they felt the graduation test was fair, several said "Yes."[24]

Defendants questioned how OTL could have been adequate for some minority students but not for others in the same classes and schools. In the *Debra P.* case, the court held that the appropriate standard for curricular validity was that "the [tested] skills be included in the official curriculum and that the majority of the teachers recognize them as being something they should teach."[25] The *Debra P.* court also found that (a) it was not constitutionally unfair that some students had mediocre teachers and (b) proving instructional validity for each individual student was an impossible burden.

Defendants asserted that collectively, the following provided sufficient evidence of OTL for the graduation test:

1. Well-publicized, state-mandated graduation test objectives that all schools were required to teach.
2. Wide dissemination to students, parents, and educators.
3. Positive adequacy of preparation reviews by educator committees and bias review panels asked to respond "yes or no" to the following question for each test item: "Would you expect students in your class to have received sufficient instruction by the time of the test administration to enable them to answer this item correctly?"
4. Mandated remediation.
5. Distribution of study guides.
6. Availability of released tests.
7. Teacher survey data from the preceding graduation test based on the same state-mandated curriculum.

In response to the readability study data introduced by the plaintiffs, the defendants responded that the analysis did not consider reading test difficulty as a function of the interaction between passages and items. Therefore, readability analyses alone were not sufficient for judging the relative difficulty of different graduation test forms. Small year-to-year equating constants for the graduation test indicated that the overall difficulty of the reading subtest had changed little over time.

In response to the survey data presented by the plaintiffs, the defendants cited a research methods text consistent with prevailing professional views that stated:

> As a result of low returns in mail questionnaires, valid generalizations cannot be made. … If they are used, every effort should be made to obtain returns of at least 80 to 90 percent or more, and lacking such returns, to learn something of the characteristics of the nonrespondents.[26]

Defendants argued that the survey data were nonrepresentative and irrelevant to the specific issue of OTL for the graduation test.

Remediation

Remediation efforts were persuasive in the *Debra P.* case where the appeals court stated:

> [The state's] remedial efforts are extensive. … Students have five chances to pass the [graduation test] between 10th and 12th grades, and if they fail, they are offered remedial help. … All [of the state's experts] agreed that the [state's remediation] efforts were substantial and bolstered a finding of [adequate OTL].[27]

The *Texas Education Code* provided: "Each school district shall offer an intensive program of instruction for students who did not [pass the graduation test]" (§ 39.024(b)). According to estimates based on TEA data, more than 13,000 African-Americans and 31,000 Hispanics who were unsuccessful on their initial attempt to pass the graduation test in 1995 were successfully remediated by their scheduled graduation in 1997.

Defendants questioned whether it would be fair to the substantial majority of African-American and Hispanic students, who worked hard to attain the skills needed to pass the graduation test, to allow minority students who were not successful on multiple retests to also receive a high school diploma. A judge in the *Debra P.* case put it this way:

> It is undoubtedly true that the appearance of having been educated may be accomplished by the conferring of a diploma. Nevertheless, if [the student has not learned the tested skills], even the most emphatic judgment and order of the most diligent court cannot supply [the missing achievement].[28]

Furthermore, it would be unfair for those minority students with high school diplomas who were qualified because employers might use their experiences with unskilled diploma holders to discount the credentials of all minority applicants in a return to the stigmatizing assumption that minority students are incapable of achieving at the same level as White students.

The Court's View

> [A]ll students in Texas have had a reasonable opportunity to learn the subject matters covered by the exam. The State's efforts at remediation and the fact that students are given eight opportunities to pass the [graduation test] before leaving school support this conclusion.
>
> …
>
> The Court has determined that the use and implementation of the [graduation test] does identify educational inequalities and attempts to address them. While lack of ef-

> fort and creativity at the local level sometimes frustrate those attempts, local policy is not an issue before the Court. The results of the [graduation test] are used, in many cases quite effectively, to motivate not only students but schools and teachers to raise and meet educational standards. (pp. 29, 31, citations omitted)

SETTING PASSING STANDARDS

The responsibility for setting passing standards on the graduation test resided with the State Board of Education. The *Texas Education Code* stated: "The State Board of Education shall determine the level of performance considered to be satisfactory on the assessment instruments" (§ 39.024(a)). Nothing in the law, administrative code, or *Test Standards* prescribed what information the State Board should consider or how it should weight the information in arriving at a passing standard. The State Board acted within its authority granted by statute when it established the 70% passing standard for the graduation test.

Primary Standard 6.9 required that the procedures used to establish the passing standard on a graduation test be documented and explained but did not require any specific method to be used. Documentation provided by the contractor and contained in the *Technical Digest* indicated that educator committees provided recommendations to the state agency and the commissioner. The commissioner in turn provided a recommendation to the State Board that included field test estimates of passing rates by subgroup at passing standards of 60% and 70% correct. After a lively and extended debate, the State Board made the final decision to set the passing standard at 60% for the first year and 70% thereafter.

With minor modifications, the 1990 graduation test was constructed to measure the same state curriculum as the 1985 graduation test that preceded it. The major difference between the 1985 and 1990 graduation tests was the level and complexity of the skills assessed. The earlier graduation test focused on basic skills; the newer graduation test covered the same curricular areas but placed more emphasis on higher order thinking and problem-solving skills. Thus, by design, the 1990 graduation test was more difficult than its predecessor.

For their discussions with the commissioner regarding passing standards for the new graduation test, the state agency received input from the educator committees that reviewed the specifications and items for the more difficult 1990 graduation test. The state agency also had results from an equating study that related 1990 scores to their equivalents on the 1985 scale. This information, together with student performance data from the field test, provided the basis for the commissioner's recommendation to the State Board.

As was pointed out to the State Board members, field test estimated passing rates must be viewed cautiously because they represent student performance under conditions of low motivation. In its pretrial order, the court observed that initially

> Defendants were also willing to tolerate up to 50 percent failures among white students. While this projected failure rate is surely not as large as that projected for minorities, it is significant and, in the Court's view, supports the Defendants' argument that its goal was remedial, not discriminatory.[29]

Furthermore, graduation test data presented by the defendants at trial indicated that the majority of students in all ethnic groups were currently meeting this standard on their first attempt and that remediation for nonpassing students had been successful.

Plaintiffs' Views

Plaintiffs argued that the passing standard for the graduation test was invalid because

> (1) The process was not based on any of the professionally recognized methods for setting passing standards on tests; (2) It appears to have failed completely to take the standard error of measurement into account; and, (3) [the standard setting process] yielded a passing score that effectively maximized the adverse impact of the [test] on Black and Hispanic students.[30]

Defendants' Views

The second point is discussed in the next section. The defendants countered the first point as follows. The *Test Standards* did not require any particular methodology. The most commonly used methods applied to the same data yield different passing standards. The purpose of such methods is to assist educator groups unfamiliar with the test to consider individual test items carefully when providing recommendations to decision-makers. The educators who provided recommendations to the commissioner had previously reviewed and accepted test items with the knowledge that the passing standard was likely to be set at 70%. Thus, a separate procedure for considering test items and their statistics was unnecessary for the educators' deliberations regarding an appropriate passing standard.

If the educators had recommended a standard other than 70% (with or without the aid of a formal methodology), the State Board would have been free to disregard the recommendation and set the passing standard at a different value. State Board members had an opportunity to review the content of test items. Both formal standard-setting methods and final judgments by designated decision- makers rely on human judgment. Based on experience from other states, reliance on a formal standard-setting procedure would probably have produced a higher passing standard.

Regarding the final point, plaintiffs offered a "cut score study." Nine Texans had been given graphs of field test distributions by ethnicity for mathematics and

reading and were told to choose the raw score that most clearly differentiated White students from African-American and Hispanic students. The responses ranged from 33 to 44 (69%–92% correct) on the 48-item reading subtest and from 34 to 44 (57%–73% correct) on the 60-item mathematics subtest. Median responses of 35 for reading and 38 for mathematics were near the corresponding 70% passing standards of 34 and 42, respectively. From these data, plaintiffs inferred that if the State Board had intended to maximize the difference in passing rates between majority and minority students, it would have chosen the 70% standard. Choosing a passing standard that did not differentiate between majority and minority students would have required passing standards substantially below 50% or above 90%.

Relating the Passing Score to the Standard Error

Some professionals have advocated an alternative passing standard that is a number of standard errors below the passing score set by a policy-making board. The rationale for this recommendation is to minimize false negatives.

Plaintiffs justified such a recommendation for the graduation test using a "risk analysis study."[31] The purpose of the study was to demonstrate that false negatives were more serious than false positives. A survey form sent to a random sample of 500 secondary teachers in Texas asked them to rate, on a scale of 1 (*minimum harm*) to 10 (*maximum harm*), the relative harm to the individual of denying a qualified high school student a diploma versus granting a diploma to an unqualified high school student.[32] Based on a 13.2% response rate, plaintiffs concluded that the possibility of a false negative posed a greater risk than a false positive and that the cut score should be adjusted downward accordingly, although other data from the survey suggested the opposite effect when participants were asked to evaluate the harm to society.

The risk of false negatives would be a concern if passing decisions had been made based on a single attempt because negative errors of measurement could cause a student with true achievement at or slightly above the passing score to fail a single administration of the graduation test. However, consistent with Primary Standard 8.8, which mandated that students be afforded multiple opportunities to pass a graduation test, students in Texas had eight attempts to pass the graduation test prior to their scheduled graduation and could continue retaking the test at subsequent administrations (*Texas Education Code* at § 39.025(b), (c)). These multiple attempts made a false negative (denying a diploma to a student whose true achievement met the passing standard) an exceedingly rare event. Conversely, multiple retakes significantly increased the probability that a student with true achievement below the cut score would have passed on one attempt due to random positive errors of measurement. Plaintiffs disputed the latter conclusion on the grounds that

> Students who fail the [graduation test] more than once or twice are likely to be held back in grade and to drop out of school long before they reach grade 12 by which time they would have had a chance to take the [test] eight times.[33]

However, at the high school level, being "held back" is typically a function of insufficient academic credits, not poor test performance.

Texas students were expected to learn all the material covered on the graduation test. When considering where to set the passing standard, board members likely considered the possibility that students might misunderstand, make careless errors, forget something, or be momentarily distracted for some items. The Board expressly lowered the passing standard by 10% for the first year to allow extra time for all tested objectives to be fully incorporated into instruction. Board meeting minutes demonstrated that members explicitly considered alternative adjustments to the recommended passing standard and were aware that students would be given multiple opportunities to retake the test. Given the historical record of a vigorous debate of these factors, lowering the passing standard another 1 or 2 standard errors below the Board's judgment of 70%, as advocated by plaintiffs, would have adjusted the passing standard twice for the same potential errors.

The Court's View

> Whether the use of a given cut score, or any cut score, is proper depends on whether the use of the score is justified. In *Cureton,* a case relied on heavily by the plaintiffs in this case, the court found that the use of an SAT cut score *as a selection practice for* the NCAA must be justified by some independent basis for choosing the cut score. …
>
> Here, the test use being challenged is the assessment of legislatively established minimum skills as a requisite for graduation. This is a conceptually different exercise from that of predicting graduation rates or success in employment or college. In addition, the Court finds that it is an exercise well within the State's power and authority. The State of Texas has determined that, to graduate, a senior must have mastered 70 percent of the tested minimal essentials.
>
> … The Court does not mean to suggest that a state could arrive at *any* cut score without running afoul of the law. However, Texas relied on field test data and input from educators to determine where to set its cut score. It set initial cut scores 10 percentage points lower, and phased in the 70-percent score. While field test results suggested that a large number of students would not pass at the 70-percent cut score, officials had reason to believe that those numbers were inflated. Officials contemplated the possible consequences and determined that the risk should be taken. The Court cannot say, based on the record, that the State's chosen cut score was arbitrary or unjustified. Moreover, the Court finds that the score bears a manifest relationship to the State's legitimate goals. (pp. 24–26, citations omitted, emphasis in original)

ADVERSE IMPACT

Differential performance occurs when passing rates for African-American and Hispanic students (minority groups) are lower than the passing rates for White students (majority group). When the differential performance between minority and majority groups becomes too great, it is labeled *adverse impact.* An important issue in this context is determining when differential performance becomes large enough to qualify as adverse impact.

In employment testing, two types of significant differences are commonly used to assess adverse impact: practical significance and statistical significance. Statistical significance is important when the group differences being used to evaluate potential adverse impact represent samples from their respective populations. In such cases, the relevant question is whether the sample differences are the result of random error or true population differences. Statistical tests can be used to evaluate whether the differential performance among the samples is large enough to justify the conclusion that there is differential performance among the respective minority and majority populations.

Once differential performance has been established for a minority population, one must decide if it is large enough to justify the label of adverse impact. This requires a judgmental evaluation of the practical significance of the population differences. The *Uniform Guidelines* for employment testing label differential performance as adverse impact when the passing rate for the minority group is less than 80% of the passing rate for the majority group.[34] One of the issues in the *GI Forum* case was whether this employment standard should be applied to state graduation tests.

Cumulative Versus Initial Passing Rates

Defendants' views. Employment cases typically involve hiring or promotion decisions based on a single administration of a test instrument. Applicants typically are not given additional opportunities to retake the test. For the graduation test, however, students had repeated opportunities to pass with targeted remediation in between. Thus, the defendants argued that when an adverse impact standard is applied to the graduation test, comparisons should be based on cumulative passing rates rather than on the initial passing rates urged by the plaintiffs.

State accountability data provided estimates of cumulative passing rates for seniors who had taken the graduation test for the first time 2 years earlier. However, these values were conservative estimates because they did not account for students who would not graduate because they had failed to complete all required courses; students whose special education status had exempted them from passing the test af-

TABLE 1
Texas Graduation Test Cumulative Passing Rates, All Tests

Class	*White*	*African-American*	*Hispanic*	*80% White*
1998	94	82	83	**75**
1997	93	79	79	**74**
1996	92	76	76	**74**

TABLE 2
Texas Graduation Test Initial Passing Rates, All Tests

Year	*White*	*African-American*	*Hispanic*	*80% White*
1994	67	35	29	**54**
1998	85	59	55	**68**

ter the initial administration; or other factors, such as changing school districts, that caused otherwise successful students not to be counted. Despite these limitations, as indicated in Table 1, the state cumulative passing rates data indicated that the 80% rule was satisfied for the 1996, 1997, and 1998 graduating classes.

Plaintiffs' views. Disagreeing, the plaintiffs argued that the 80% rule from the employment context should be applied to initial passing rates on the graduation test. This analysis treated scores from initial test administrations as if they determined whether a student would receive a high school diploma.

Trends in initial passing rates. Longitudinal data for 1994 through 1998 indicated that the initial passing rates for Whites, African-Americans, and Hispanics increased over time and that the largest gains were made by African-American and Hispanic students. The percentage increase in initial passing rates was greatest in mathematics, where African-American and Hispanic passing rates increased 85% and 63%, respectively, compared to a 26% increase for Whites over the 5-year period.[35] From 1994 to 1998, both minority groups also closed the gap between their overall initial passing rates and the 80% standard. As demonstrated in Table 2, African-Americans moved from 25 points below the 80% standard in 1994 to 13 points below in 1998. The Hispanic group closed the gap from 19 points below the 80% standard in 1994 to 9 points below the standard in 1998.

Moreover, data for 1998 indicated that approximately 10,400 seniors statewide were still trying to pass the graduation test at the last administration prior to their scheduled graduation and about one third were successful. Those seniors who did not pass the test at this administration may also have not completed all the course work required for graduation. Assuming about half the nonpassing students had

completed all the other requirements for graduation, approximately 1.7% may have been prevented from graduating due to failure to pass the graduation test.

Calculation Methodology

Plaintiffs' views. As an alternative to initial passing rates, the plaintiffs argued that if cumulative passing rates were considered, they should be based on percentages of students from the class of students initially administered the test in 10th grade who received high school diplomas 2 years later. In their calculations, students who dropped out, were repeat test takers, moved out of state, received diplomas by satisfying a special education Individualized Educational Program (IEP), or chose to spread their high school courses over more than the traditional 4 years were counted as failures.

The plaintiffs also urged the court to consider statistical significance tests for evaluating adverse impact. They used a statistical test for the difference in two proportions, appropriately applied in employment cases when only a sample of the potential applicant pool had been tested, to calculate the probability of population differences based on the full population. For the 1998 graduation test, using this statistical procedure and state total enrollments by ethnicity "as estimates of sample sizes for calculating tests of statistical significance," the plaintiffs reported *z* values across subtests and for all tests taken from 55.6 to 105.1 for African-American/White differences and from 98.4 to 133.4 for Hispanic/White differences, respectively. Another of the plaintiffs' experts calculated similar group difference statistics across subtests for 1993 through 1997 and reported *z* values from 8,502 to 12,504. Plaintiffs' experts testified that these large *z* values indicated that adverse impact was very highly significant for these group comparisons.

Defendants' views. The state data on which the plaintiffs' calculations were based consisted of subgroup passing rates for all Texas students. For example, for 1998 first-time test takers in 10th grade, initial passing rates for all tests taken of 85%, 55%, and 59% for White, African-American, and Hispanic students, respectively, were based on statewide populations of 110,893, 27,921, and 73,044, respectively. By using sample inferential statistical tests on population data involving large numbers of students, the magnitude of the plaintiffs' obtained *z* values was predictable but not appropriate for judging whether subgroup differences were large enough to create a presumption of adverse impact.

Arguing that inferential statistical tests such as the difference between two proportions should not be applied to population data, the defendants calculated that the difference in cumulative passing rates between Whites and Hispanics would have to decrease to 0.36 percentage points (93.41% White to 93.05% Hispanic) for

the number of standard deviations of the difference (z) to be less than 3 using the plaintiffs' methodology and criteria. The plaintiffs countered this view by asserting that the subgroups were samples from the total population and the items administered to students were samples from the domain of all possible items. The defendants rejected these assertions because all subgroups took the same test (items were not sampled and the plaintiffs' standard error calculations used the number of students) and because all members of each subgroup were included in the calculation of the passing rate for that subgroup (the total population was partitioned into subpopulations).

Articulating a legal standard for evaluating adverse impact. In the defense view, no statistical tests were required to determine that the Texas data reflected actual differences among subpopulations. The relevant questions were whether the observed differences were large enough to meet a legal standard of adverse impact (i.e., were practically significant), and if so, whether the observed adverse impact was caused by the graduation test or other factors. Application of the 80% rule was one possible way of judging practical significance. The plaintiffs argued that another indicator of practical significance was the number of minority students who would have passed the test if their passing rates had been equal to the passing rate for White students. No evidence was offered to support the reasonableness of assuming that passing rates should be identical in all subgroups. One of the plaintiffs' experts addressed the causation question, stating, "[M]y own view … is that social, economic, and educational factors are the main determinants of the relative standing of ethnic groups on test results."[36]

The Court's View

> The Court finds that, on balance, remedial efforts are largely successful. TEA's expert … estimates that 44,515 minority students in 1997 were successfully remediated after having failed their first attempt at the [graduation test] in 1995. The Court finds this evidence credible. … [T]he Court also finds that it is highly significant that minority students have continued to narrow the passing rate gap at a rapid rate. In addition, minority students have made gains on other measures of academic progress. …
> …
> In determining whether an adverse impact exists in this case, the Court has considered and applied the [EEOC's 80% Rule]. … Plaintiff's statistical analysis, while somewhat flawed, demonstrates a significant impact on first-time administration of the [graduation test]. However, cumulative pass rates do not demonstrate so severe an impact and, at least for the classes of 1996, 1997, and 1998, are not statistically significant under the [EEOC's 80% Rule].
> In considering how to handle the dilemma of choosing between cumulative and single-test administration, the Court has taken into account the immediate impact of

initial and subsequent in-school failure of the exam—largely successful educational remediation. In addition, the Court has considered the evidence that minority scores have shown dramatic improvement. These facts would seem to support the TEA's position that cumulative pass rates are the relevant consideration here.

The Plaintiffs argue that successful remediation and pass-rate improvement should not be considered in determining whether an adverse impact exists. To support their argument, the Plaintiffs point to case law holding that a "bottom line" defense is insufficient to combat a showing of adverse impact. The Court is not convinced that this argument is applicable to the case before it.

…

Having said all that, however, the Court finds that, whether one looks at cumulative or single-administration results, the disparity between minority and majority pass rates on the [graduation test] must give pause to anyone looking at the numbers. … Disparate impact is suspected if the statistical significance test yields a result, or z-score, of more than two or three standard deviations. In all cases here, on single and cumulative administrations, there are significant statistical differences under this standard. Given the sobering differences in pass rates and their demonstrated statistical significance, the Court finds that the Plaintiffs have [met their burden of showing] significant adverse impact.

[The Court then considered whether TEA had met its burden of producing evidence of a manifest relation between the graduation test and a legitimate educational goal and whether the plaintiffs had demonstrated equally effective alternatives to the current use of the graduation test. The Court held that TEA had met its burden but that the plaintiffs had not.]

…

Ultimately, resolution of this case turns not on the relative validity of the parties' views on education but on the State's right to pursue educational policies that it legitimately believes are in the best interests of Texas students. The Plaintiffs were able to show that the policies are debated and debatable among learned people. The Plaintiffs demonstrated that the policies have had an initial and substantial adverse impact on minority students. The Plaintiffs demonstrated that the policies are not perfect. However, the Plaintiffs failed to prove that the policies are unconstitutional, that the adverse impact is avoidable or more significant than the concomitant *positive* impact, or that other approaches would meet the State's articulated legitimate goals. In the absence of such proof, the State must be allowed to design an educational system that it believes best meets the need of its citizens. (pp. 12, 15, 20–21, 22–23, 27–28, 7, citations omitted, emphasis in original)

USE OF A SINGLE TEST SCORE

Primary Standard 8.12 stated:

In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score. Other relevant information for the decision should also be taken into ac-

> count by the professionals making the decision.
> Comment:
> A student should not be placed in special classes or schools, for example, solely on the basis of an ability test score. Other information about the student's ability to learn, such as observations by teachers or parents, should also play a part in such decisions. (p. 54)

Primary Standard 8.8 stated:

> Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have multiple opportunities to demonstrate the skills. (p. 53)

Plaintiffs' Views

Plaintiffs argued that the graduation test violated Standard 8.12 because passing decisions were based on a single test score. They argued that other data such as high school grades should be used in conjunction with graduation test scores to award high school diplomas in Texas. In particular, plaintiffs cited low correlations between high school grades and graduation test scores as evidence that the graduation test lacked validity and should not be used in isolation to determine which students qualified for a high school diploma.

Defendants' Views

Contrary to plaintiffs' assertions, the defendants argued that the graduation test was not used in isolation to make graduation decisions. In addition to passing the graduation test, students were expected to successfully complete all required course work and other graduation obligations imposed by their districts. Students were required to meet both testing and course requirements because each represented a different kind of accomplishment that was valued in a high school graduate.

The defendants further argued that the plaintiffs' interpretation of Standard 8.12 as applying to graduation tests was inconsistent with the intent of the drafters as indicated by the comment to Standard 8.12 and the inclusion of Standard 8.8. The comment to Standard 8.12 indicated that the drafters were concerned about the use of a single test score, with no opportunity for retesting, to place students in an instructional program. The inclusion of Standard 8.8 suggested that the drafters considered graduation tests separately from educational placement tests and found such tests acceptable as long as students had multiple opportunities to pass. It appeared that the drafters envisioned different methods of compensating for potential measurement error in the two situations and did not intend that both multiple retakes and the inclusion of nontest data be required for graduation tests.

Moreover, students who failed a single course may have been unable to graduate on time just as those who did not pass the graduation test may have had to delay graduation. In both cases, students had multiple opportunities to complete the failed course or retake the failed graduation subtest. Furthermore, students whose graduations were delayed due to not having passed the graduation test had not had their graduation delayed based on a single piece of data. Rather, the denial was based on the opportunity to obtain at least eight scores from eight forms of the test administered on eight different occasions. In addition, it was virtually impossible for the true achievement of such students to be above the graduation test passing standard.[37] Thus, these students were not false negatives and the decision to delay award of their high school diplomas until they had attained the required skills and passed the graduation test was justified.

The defendants also disagreed with the plaintiffs' suggestion that grades be used to compensate for low test scores when making passing decisions. As indicated in the discussion on validity, grades can reasonably be assumed to include factors other than achievement of the tested skills. Furthermore, grades were already included because students had to pass all required course work to receive a high school diploma. Using high school grades again to compensate for low test scores would further exacerbate the pressures for grade inflation and alter the interpretation of the test result as a measure of essential knowledge and skills. Thus, defendants argued that it would be inappropriate to allow high grades to compensate for low scores on the graduation test.

The Court's View

> [T]he failure to pass the [graduation test] does serve as a bar to graduation, and the exam is properly called a "high-stakes" test. … [But] a single [graduation test] score does *not* serve as the sole criterion for graduation. The TEA presented persuasive evidence that the number of testing opportunities severely limits the possibility of "false negative" results and actually increases the possibility of "false positives," a fact that arguably advantages all students whose scores hover near the borderline between passing and failing.
>
> …
>
> In addition, the State need not equate its test on the basis of standards it rejects, such as subjective teacher evaluations. (pp. 15, 30, emphasis in original)

CONJUNCTIVE VERSUS COMPENSATORY MODELS

Under a Title VI challenge, the burden of proof alternates between plaintiffs and defendants. Plaintiffs establish a presumptive violation with adverse impact data. Defendants can counter this presumption with evidence of educational necessity.

Plaintiffs may still prevail if they can demonstrate that an equally valid, less discriminatory alternative is available.

Plaintiffs' Views

In the *GI Forum* case, the plaintiffs offered two compensatory models as possible replacements for the conjunctive model used by Texas to award diplomas based on passing the graduation test and completing required course work. One compensatory model offered by the plaintiffs was the use of high school grades to offset poor graduation test performance. This option was critiqued earlier.

Justifying their proposed use of high school grades along with graduation test scores, the plaintiffs cited a case involving the use of SAT test scores to award college scholarships.[38] Data indicated that female students were receiving such scholarships significantly less often than male students and these data formed the basis of an argument of gender discrimination. Because female students tended to have higher grades in high school than male students, the court was persuaded that a fairer method for awarding scholarships was to use a combination of SAT scores and GPAs similar to the weighted decisions made by college admissions officers. The court ruled that predictions of college success based on GPAs and SAT scores were less discriminatory than predictions based on test scores alone.

Defendants' Views

The defendants argued that the SAT college scholarship case was not applicable to the graduation test situation because the types and purposes of the two tests were different. The SAT is an ability test designed to predict college success; the graduation test is an achievement test designed to measure attainment of teachable skills. As a proxy measure for motivation and effort, grades may contribute to more accurate predictions of college success. However, because grades may measure attributes other than achievement of skills, they are not a good proxy measure of the skills the graduation test is intended to measure.

Sliding Scale

The other compensatory model offered by the plaintiffs was the creation of a sliding scale based on composite performance across subtests. For example, if there were two subscores, reading and mathematics, with individual passing standards of 70 on a 100-point scale, a student who achieved a total score of 140 would be judged to have met the standard. This model provided that poor performance in one subject (e.g., mathematics) could be offset by a good performance in another subject (e.g., reading). Under this compensatory model, a student with a score of 55 in mathematics and a score of 85 in reading would pass the gradua-

tion test. Defendants asserted that this procedure would defeat the purpose of the graduation test, which was to ensure a minimal level of competence in both mathematics and reading. In other words, greater reading skill could not compensate for lack of knowledge of mathematics.

Furthermore, in the defendants' view, both of the "alternatives" proposed by the plaintiffs had another deficiency. The legal standard for equally valid alternatives contemplates a less discriminatory but equally valid test.[39] Rather than proposing an alternative testing instrument, the plaintiffs were proposing a dilution of the graduation test scores with less valid and less reliable nontest information.

The Court's View

> In spite of projected disparities in passing rates, the TEA determined that objective measures of mastery should be imposed in order to eliminate what it perceived to be inconsistent and possibly subjective teacher evaluations of students. The TEA offered evidence at trial that such inconsistency exists. The TEA also presented testimony that subjectivity can work to disadvantage minority students by allowing inflated grades to mask gaps in learning.
>
> …
>
> In considering whether the Plaintiffs have shown that there are equally effective alternatives to the current use of the [graduation test], the Court must begin with the State's articulated, legitimate goals in instituting the examination. Those goals are to hold students, teachers, and schools accountable for learning and for teaching, to ensure that all students have the opportunity to learn minimal skills and knowledge, and to make the Texas high school diploma uniformly meaningful.
>
> …
>
> Plaintiffs did offer evidence that different approaches would aid the State in measuring the acquisition of essential skills. Among these approaches was a sliding-scale system that would allow educators to compensate a student's low test performance with high academic grades or to compensate lower grades with outstanding test scores. However, Plaintiffs failed to present evidence that this, or other, alternatives could sufficiently motivate students to perform to their highest ability. In addition, and perhaps more importantly, the present use of the [graduation test] motivates schools and teachers to provide an adequate and fair education, at least of the minimum skills required by the State, to all students. The Plaintiffs produced no alternative that adequately addressed the goal of systemic accountability. (pp. 12, 27–28, citations omitted)

DIFFERENTIAL ITEM PERFORMANCE

When minority and majority students exhibit differential levels of performance on an achievement test, some observers believe that the test items are "biased" against members of the lower scoring minority group. However, an equally plausible ex-

planation for the differential performance is a true difference in average achievement levels for the two groups. To address the issue of differential performance for the graduation test, item response theory (IRT) and Mantel–Haenszel statistics were calculated for each item and presented to the educator review panels.

Plaintiffs' Views

The plaintiffs argued that the graduation test was "biased" because *p*-value differences between majority and minority groups correlated highly with total group point biserials (a difference correlated with another correlation). This led the plaintiffs to argue that test development procedures for the graduation test were flawed because the "methods employed by defendants are not designed to reduce racial/ethnic differences in either item performance or passing rates."[40] The plaintiffs argued that the point-biserial statistic should not have been used in the selection of items for the graduation test because items with larger point biserials tended to be items with greater White–minority *p*-value differences. This argument implied that the defendants had a duty to minimize majority/minority *p*-value differences. The plaintiffs argued that this could be done by using minority group point biserials rather than majority group point biserials (see discussion in next section).

Defendants' Views

Defendants responded that only items approved by ethnically representative educator committees after review of all statistical data were eligible for selection. For the reviews, items with point biserials less than .30 were flagged to alert committee members that the item might be miskeyed or ambiguous. The committee members made a final decision to accept or reject each item based on a review of its content and all the accompanying statistical data, including two measures of differential item performance for African-Americans and Hispanics that quantified racial differences with ability held constant. An item that exhibited statistically significant differential performance between minority and majority students could be retained for use on a graduation test form only if, in the professional judgment of the item review committee, the item was a fair measure of its corresponding state objective for all students and was free of offensive language or content that might differentially disadvantage minority students.

When items were selected for a test form from the pool of eligible items, the primary criteria were following the content representation specified by the test blueprint, choosing a set of items for each objective with an average difficulty approximately equal to a target value from prior test forms, and ensuring that the

answer to an item being considered for selection would not be cued by material contained in any other selected item. All else being equal, after the aforementioned criteria were satisfied, the item with the higher point-biserial statistic would be preferred. The reason for the preference was that items with higher point biserials were more likely to be answered correctly by students who had achieved the skills measured by the test and incorrectly by those who had not, thus increasing the reliability of the test.

The defendants also characterized the plaintiffs' use of evidence tied to *p*-value differences as an attempt to institute the *Golden Rule* procedure previously renounced by the Educational Testing Service and discredited by measurement professionals.[41] Furthermore, defendants criticized the correlation of *p*-value differences with point biserials because: (a) "bias" measures require groups of equal achievement, and to the extent that *p*-value differences are based on groups of unequal ability, the purported measure of "bias" is confounded by achievement differences in the two groups; (b) when the effects of unequal achievement were removed by using IRT differences, the correlations and differential effects across groups decreased substantially; (c) item *p*-value differences and point biserials for groups of unequal achievement are positively correlated because both measure the ability of the item to distinguish students who have learned the tested content from those who have not; (d) total and minority group point-biserial distributions and rank orderings of items were similar; and (e) differences between highly correlated measures, such as *p*-values by subgroup, are unreliable, and the correlation between an unreliable measure and another measure has little interpretive validity.

Summary of Plaintiffs' Criticisms and Proposed Alternatives

Plaintiffs' experts urged the court to find defendants' test construction procedures in violation of professional standards because item point-biserial statistics were positively correlated with majority/minority *p*-value differences, and total group point-biserial statistics were utilized rather than minority subgroup point-biserial statistics. An expert for the plaintiffs stated, "The effects of the specific item pre-testing and selection procedures employed by defendants are to retard the reduction of racial differences in item percent-correct values and test passing-rates."[42] From the defendants' perspective, the court was being asked to invalidate use of the total group point-biserial statistic for test development because across items, total group point biserials were positively correlated with *p*-value differences between subgroups (the discredited *Golden Rule* measure of item bias), or in the alternative, to require test developers to use subgroup rather than total group point-biserial statistics. The former alternative was discussed earlier; the latter alternative proposed by the plaintiffs is discussed in the next section.

The Court's View

> The Court also finds that the Plaintiffs have not demonstrated that the [graduation test] is a substantial departure from accepted academic norms or is based on a failure to exercise professional judgment. …
> The Court, in reaching this conclusion, has considered carefully the testimony of Plaintiffs' expert … demonstrating that the item-selection system chosen by TEA often results in the favoring of items on which minorities will perform poorly, while disfavoring items where discrepancies are less wide. The Court cannot quarrel with this evidence. However, the Court finds that the Plaintiffs have not been able to demonstrate that the test, as validated and equated, does not best serve the State's goals of identifying and remediating educational problems. Because one of the goals of the [graduation test] is to identify and remedy problems in the State's educational system, no matter their source, then it would be reasonable for the State to validate and equate test items on some basis other than their disparate impact on certain groups. (pp. 29–30)

SUBGROUP ITEM POINT-BISERIAL STATISTICS

Plaintiffs' Views

Plaintiffs presented data from the 1997 mathematics graduation subtest to bolster their argument that subgroup point-biserial statistics should have been used for test development. Plaintiffs argued that different items would have been selected for the graduation test had African-American or Hispanic point biserials been used rather than the total group point-biserial statistic.

Defendants' Views

Analyzing plaintiffs' data, defendants' experts noted that: (a) There must have been a calculation error because the plaintiffs included point-biserial values smaller than the lowest value for any item on that particular test; (b) the items analyzed represented an intact test form for which all items were used, not a pool of items from which a subset might be selected; (c) the total group point biserial did not allow data for White students (52%) to swamp that from minority students (48%) as claimed by the plaintiffs; and (d) the reported correlations between African-American–total group (.91) and Hispanic–total group (.93) point biserials indicated that the minority point biserials were rank ordering the items similarly to the total group point biserials.

Based on these considerations, the defendants believed that item selection decisions made using minority point biserials would be similar to those obtained based on total group point biserials. To illustrate this point using the plaintiffs' data, defendants considered the argument that the 60 items on the 1997 mathematics grad-

uation subtest represented a pool of items from which 50% were to be selected and that item selections were to be made based solely on the point-biserial correlation (these assumptions did not represent actual practice but were adopted to simulate plaintiffs' proposed procedure). The defendants then rank ordered the items in turn based on the total group, African-American, and Hispanic point biserials and selected the 31 items (there were ties) with the highest point biserials in each group.

The results of this analysis are shown in Table 3. All but three items (10%) selected using the minority point biserials were identical to the selections made using the total point biserial. Where there were differences, the minority point biserials favored less complex computation items over more complex problem-solving items. Therefore, these changes illustrate the potential for decreased content validity, a major problem previously identified for *Golden Rule* type procedures. Choosing items with smaller total point biserials also decreases test reliability. As a result, defendants urged the court not to invalidate use of the point-biserial statistic as one of many factors considered in the selection of test items, and not to adopt the use of minority point biserials as a remedy for differential graduation test performance by minority groups.

The Court's View

The court did not specifically address the issue of which item statistics should be preferred. However, as indicated in the previous section, the court found that TEA and its contractors adhered to accepted professional practices in constructing the graduation test and were not required to choose methods that would minimize selected items' disparate impact on certain groups.

FACTOR ANALYSES

Plaintiffs' Views

Plaintiffs' experts argued that test development procedures for the graduation test were flawed because factor analyses were not used to select the items. To illustrate their criticism of the test, they ran a principal components analysis with varimax rotation separately for African-Americans, Hispanics, and Whites on the 1997 mathematics graduation subtest identifying all factors with eigenvalues greater than 1. The same number of factors (60) as items on the test were identified in each analysis. Using the associated factor loadings and a criterion of 0.3, items were grouped into four major and four minor factors for each ethnic subgroup.

The plaintiffs' experts then used this information to argue that the test was flawed because each item did not load cleanly on a single factor. They argued that factor analyses should have been used to eliminate items loading on more than one factor or failing to show a significant factor loading on any factor. In addition, they

TABLE 3
Selection of 50% of Items Based on Point-Biserial Statistic[a]

Item	*Total Point Biserial*	*Rank*	*African-American Point Biserial*	*Rank*	*Hispanic Point Biserial*	*Rank*	*Item Content*
48	.58	1	.53	8	.55	3	
33	.57	2	.55	4	.55	3	
54	.57	2	.57	1	.58	1	
15	.55	4	.55	4	.54	7	
42	.54	5	.52	12	.54	7	
37	.54	5	.54	7	.54	7	
12	.54	5	.53	8	.56	2	
18	.54	5	.53	8	.54	7	
53	.53	9	.56	3	.55	3	
16	.53	9	.53	8	.51	12	
40	.52	11	.52	12	.51	12	
13	.52	11	.50	17	.50	16	
06	.52	11	.57	1	.55	3	
17	.52	11	.51	15	.51	12	
28	.51	15	.48	19	.48	19	
26	.51	15	.51	15	.50	16	
45	.51	15	.47	21	.47	28	
08	.51	15	.52	12	.51	12	
04	.51	15	.55	4	.52	11	
36	.50	20	—	—	—	—	Problem solving with chart of survey data using ratio and proportion
27	.50	20	.48	19	.48	19	
10	.50	20	.47	21	.47	28	
41	.49	23	.47	21	.50	16	
49	.49	23	.46	29	.47	28	
19	.49	23	.46	29	.48	19	
09	.49	23	.47	21	.48	19	
29	.48	27	—	—	—	—	Reasoning about a graph—selecting a conclusion
56	.48	27	.47	21	.48	19	
46	.47	29	—	—	—	—	Computation—figuring sales tax
58	.47	29	.47	21	.48	19	
05	.47	29	.47	21	.47	28	
57	—	—	.49	18	.48	19	Computation—column addition
50	—	—	.47	21	—	—	Computation—division miles per gallon
55	—	—	.46	29	.48	19	Computation—column addition
03	—	—	—	—	.48	19	Computation—average 4 numbers

[a]Grade 10 item analysis; Spring 1997; base form mathematics.

suggested that there was something wrong with the test because the first factor accounted for substantially less than 50% of the variance and that the factor structure of the test differed across ethnic subgroups. Plaintiffs stated that the

> objectives that were attempted to be measured by the test are factorially inconsistent, … scattered throughout the factors. [T]he lack of clustering suggests that five of the [math] objectives were not measured in a factorially "clean" way for *any* of the three ethnic groups studied.[43]

Defendants' Views

The defendants responded with the following points:

1. The use of factor analyses as described by plaintiffs was consistent with models used to develop tests of psychological constructs but not educational achievement tests.
2. An appropriate use of factor analyses for educational achievement tests is to provide evidence relevant to the unidimensionality assumption of IRT analyses used in scaling and equating.
3. When factor analyses are run on achievement test data, a principal axis analysis with oblique rotation (assuming correlated factors) is more appropriate.
4. If the plaintiffs' data were rescaled based on the first eight factors identified, the first factor would account for more than 50% of the variance as expected for an achievement test.
5. Comparisons for ethnic subgroups were run incorrectly. An initial run of the total data should have been completed and used to determine the appropriate number of factors; then separate runs with the number of factors fixed to the predetermined number should have been obtained for the ethnic subgroups.
6. The claim that the factor analyses demonstrated a different factor structure for each subgroup was not substantiated by similarity analyses or analyses of item content. Given that the identified factors each accounted for a very small percentage of the total variance, their ordering within the subgroups was not significant; content analysis of the item groupings identified by the plaintiffs demonstrated that the content of each major cluster was unique and that the same four major clusters (integers, fractions and decimals, concepts, money problems) were identified in each ethnic group; overall, 87% of the 60 items loaded on the same factor for two or more groups; of the 8 remaining items, 6 could have been solved using algebra or mathematical reasoning.

The Court's View

The court did not specifically address the use of factor analysis in test development but found "on the basis of the evidence presented at trial, that the disparities in test

scores do not result from flaws in the test or in the way it is administered" (p. 30). The court also held that the graduation test was valid as constructed.

DROPOUT AND RETENTION RATES

Society benefits when students stay in school and earn a high school diploma because high school dropouts typically hold lower paying jobs and have limited opportunities for advancement. Nationally, there is a growing concern about differential dropout rates among ethnic groups, particularly for Hispanic students who have the highest rates for leaving school without a high school diploma.[44]

Plaintiffs' Views

The plaintiffs argued that implementation of the Texas graduation test in 1990 caused increases in dropout rates among minority students and that schools were retaining 9th-grade students likely to fail the graduation test in 10th grade to bolster their accountability ratings. "The [graduation test] has led to dramatic increases in retention of Black and Hispanic students in grade 9, and this retention is certainly contributing to increases in students dropping out of high school before completion."[45] In support of their views, plaintiffs calculated progression ratios (number of students enrolled in one grade divided by the number of students enrolled in the preceding grade a year earlier) and success ratios (number of high school graduates divided by the number of students enrolled in the ninth grade 3 years earlier) for each ethnic group.

Defendants' Views

According to a study by TEA, since the implementation of the graduation test in 1990, dropout rates for all students have steadily declined and the gap between minority and majority students has shrunk from about 3.5 points in 1990 to about 1 point in 1997. There was no evidence in the TEA study that introduction of the graduation test affected the dropout rate for any group.

Both sides acknowledged the difficulty in obtaining accurate information about school-level decisions. The TEA study, based on information supplied by 58% of schools, provided the best available data on dropouts and retention. Acknowledging that definitions of dropouts had been adjusted periodically to increase the accuracy of reporting, these data provided support for a trend toward decreased percentages of dropouts for all ethnic groups across the last decade and a narrowing of the differences between groups. For 1997, the TEA data indicated annual dropout rates of 2.3%, 2.0%, and 1.0% for Hispanics, African-Americans, and Whites, respectively.

Dropouts were linked primarily to overage students; special education, LEP, and other at-risk groups were not overrepresented. Students who were overage and not on grade level constituted more than 80% of the dropouts, almost 50% greater than their percentage of the total enrollment. This suggested that the majority of student dropouts were having academic difficulties in school. These data were consistent with information on reasons for dropping out provided by the schools, which indicated that the majority of students from all groups dropped out due to academic difficulties and only 2% dropped out due to failing the graduation test or other graduation requirements.

Students first attempted the graduation test in 10th grade. If anticipated or actual failure on this test caused substantial numbers of minorities to drop out of school, one would expect a spike in the number of dropouts in the 10th and 11th grades. The data collected in the TEA study indicated no such spike. Dropout rates for all groups were relatively flat in 10th and 11th grade. The largest percentage of dropouts occurred in 12th grade for African-Americans and in 9th grade for Hispanics, well after and well before their first attempt to pass the graduation test, respectively.

TEA data also indicated higher retention rates in Grade 9 than at any other grade. However, a plausible explanation for the higher number of ninth-grade retentions was failure to complete the number of credits required to be classified as a sophomore. Defendants noted that such students could take sophomore classes in subjects for which the required freshman classes had been completed and were required to repeat or complete only those required freshman classes for which they had not yet received credit. Some students were able to make up the missed work and become classified as juniors the following year. In any case, retention in ninth grade clearly did not mean the same thing as repeating an entire grade in elementary school.

Calculation Methods

Much debate in this area centered around the appropriate method for calculating dropout and retention rates. Plaintiffs preferred to calculate completion rates by dividing the number of high school graduates in a given year by the number of students enrolled in ninth grade 3 years earlier, and grade progression ratios by dividing enrollment for a given grade by the enrollment for the previous grade 1 year earlier. Defendants observed that if such statistics were calculated using seventh grade as the base year, retention rates were negative. However, this debate was moot because the plaintiffs provided no data supporting their allegation that the graduation test caused minority students to drop out or be retained. Equally plausible reasons included academic difficulties, family obligations, jobs outside school, or more demanding courses.

The Court's View

> Plaintiffs presented sufficient evidence to support a finding that Texas students, particularly minority students, drop out of school in significant numbers and are retained at their current grade level in numbers that give cause for concern. … However, Plaintiffs have failed to make a causal connection between the implementation of the [graduation test] and these phenomena, beyond mere conjecture. In other words, Plaintiffs were only able to point to the problem and ask the Court to draw an inference that the problem exists because of the implementation of the [graduation test]. That inference is not, in light of the evidence, inevitable. The Defendants hypothesize, just as plausibly, for example, that the ninth grade increase in dropouts is due to the cessation of automatic grade promotion at the beginning of high school in Texas. (pp. 17–18)

CONCLUSION

In its decision, the *GI Forum* court provided guidance for graduation testing programs in several important areas. The *GI Forum* court:

- Upheld the *Debra P.* requirements of notice and curricular validity implicated by a property interest in a high school diploma.
- Credited the professional judgment of psychometric experts who had extensive, direct experience with large-scale achievement testing.
- Used reasonable commonsense interpretations of professional standards for evaluating test quality.
- Did not require perfection or mandate secondary or conditional professional standards.
- Supported the use of cumulative versus initial passing rates and the 80% rule for adverse impact analyses.
- Recognized that there was no evidence of a causal link between the graduation test and differential minority performance and that a variety of nontest factors may have contributed to the observed differences.
- Found that the graduation test was not the sole criterion for receipt of a high school diploma.
- Indicated that graduation test developers are not required to minimize differential performance among racial and ethnic groups or to validate the test against criteria (e.g., subjective teacher grades) that measure different student attributes than those measured by the test.
- Found evidence of successful remediation of minority students convincing and compelling.
- Upheld passing standards for a graduation test based on all the facts and circumstances, including multiple retakes but absent evidence of research-based methodologies.

- Found a graduation test developed using content and OTL procedures and criteria valid and with reported total score KR_{20} reliabilities by subgroup in the upper 80s and low 90s sufficiently reliable.
- Found a graduation test administered only in English to be valid for all students, including those whose native language is not English.
- Found notice of the graduation test 3 years prior to initial testing adequate.
- Found adequate OTL on all the facts and circumstances, including successful remediation but absent formal surveys of teachers and students.
- Noted that high dropout and retention rates among minority students were cause for concern but not shown to have been caused by the graduation test.
- Indicated that decisions of whether and what to test for high school graduation are the province of the legislature, not the courts.

In sum, the *GI Forum* court held that:

> While the [graduation test] does adversely affect minority students in significant numbers, the TEA has demonstrated an educational necessity for the test, and the Plaintiffs have failed to identify equally effective alternatives. … The TEA has provided adequate notice of the consequences of the exam and has ensured that the exam is strongly correlated to material actually taught in the classroom. In addition, the test is valid and in keeping with current educational norms. Finally, the test does not perpetuate prior educational discrimination. … Instead, the test seeks to identify inequities and to address them. (pp. 31–32)

In its finding of adverse impact, the court relied on statistical tests applied to population differences. If other courts adopt this strategy, virtually all differences will qualify as adverse impact because differences of less than 1 percentage point will be 2 or 3 standard errors apart given the typically large statewide populations of tested students. Rather than using inferential statistics incorrectly, courts could justify a finding of adverse impact by creating their own tests of practical significance based on judgmental criteria. Examples of such judgmental criteria might include a more stringent version of the 80% rule (i.e., a "90% rule") or a fixed percentage difference (i.e., 5% or 10%).

In this case, although the plaintiffs tried hard to discredit the psychometric quality of the test as a graduation requirement for individual students, they supported the continuation of the test as part of a system of school accountability. The defendants asserted that to ensure a common standard of achievement for high school diploma recipients, the responsibility for achieving the state curriculum must be shared jointly and concurrently by students and schools. The court found that:

> [T]he TEA has shown that the high-stakes use of the [test] as a graduation requirement guarantees that students will be motivated to learn the curriculum tested. … In addi-

> tion, the evidence was unrefuted that the State had an interest in setting standards as a basis for the awarding of diplomas. The use of a standardized test to determine whether those standards are met and as a basis for the awarding of a diploma has a manifest relationship to that goal. (p. 26)

This view was echoed by the director of the Washington-based Education Trust who commented:

> It's a little overly simplistic to say schools can be held accountable without high stakes for the kids. At the high school level, if there are no incentives for students to work hard and meet standards and do well on a test, they won't.[46]

One high school sophomore observed: "[The graduation test] is easy, and everybody should be able to pass it. In your classes you can cheat on everything. You can't cheat on [the graduation test]. It's the only true measure of what you know."[47]

In the *GI Forum* case, the plaintiffs challenged the psychometric quality of the current graduation test although their arguments suggested that they objected to any policy of graduation testing. Thus, even when a state has worked hard to follow psychometric standards, a graduation test may be a visible and accessible target for those whose political and social beliefs lead to different conclusions about appropriate test use. One important lesson learned from this lawsuit is that state testing programs must not only follow legal and professional standards but must also create detailed documentation of those efforts and be prepared to defend all their testing program decisions in court.

Other lessons learned from the *GI Forum* case included these:

- The *Debra P.* requirements as implemented in the *Test Standards* remain in effect.
- Creation of a technical manual is a valuable method for collecting and memorializing important graduation test procedures, decisions, and psychometric data.
- A potential finding of adverse impact can trigger comprehensive scrutiny of all facets of a graduation test and of a statewide testing program in general.
- Courts may allow wide latitude to the plaintiffs to present any information potentially related to the graduation test, reserving determinations of relevance, quality, and credibility until all evidence has been presented. In response to objections by defendants, the court may admit any psychometric or statistical data with the caveat that the court will give it whatever weight it deems appropriate in the final decision.
- Absent any evidence that a graduation test caused other educational outcomes, courts may still be receptive to and troubled by evidence of outcomes, such as dropout and retention rates, that negatively affect historically disadvantaged subgroups.

• The time elapsed from the initiation of a lawsuit against a graduation test to a final court decision may be 2 years or more.

ENDNOTES

[1]This article is based on firsthand knowledge of the Texas graduation test, pretrial activities, and trial testimony. The author has served as a consultant for the Texas Student Assessment Program since the early 1980s and served as a consultant and expert witness for the Texas Education Agency (TEA) in the *GI Forum* case.

[2]Texas Assessment of Academic Skills (TAAS), hereinafter referred to as "the graduation test."

[3]42 U.S.C. § 2000d, 20 U.S.C. § 1703, 34 C.F.R. § 100.3, 42 U.S.C. § 1983, and United States v. Texas, 330 F. Supp. 235 (E.D. Tex. 1970), respectively.

[4]*See* Order Granting, in part, and Denying, in part, Defendants' Motion for Summary Judgment, GI Forum et al. v. TEA et al., CA No. SA–97–CA–1278–EP, U.S. District Court, Western District of Texas, San Antonio, TX, July 27, 1999, p. 23.

[5]GI Forum et al. v. TEA et al., F.Supp., 1 (W.D. Tex. 2000).

[6]Summary Judgment Order at 11–12; Debra P. v. Turlington, 644 F.2d 397 (5th Cir. 1981).

[7]*Id.* at 20.

[8]*GI Forum* at 2.

[9]The new graduation test replaced the old graduation test as a diploma requirement in 1992.

[10]Debra P. v. Turlington, 730 F.2d 1405, 1416–17 (11th Cir. 1984).

[11]Psychometric arguments and opinions advanced by litigants appeared in written expert witness reports, trial testimony of those experts, concurring testimony of affiliated experts, colloquies between counsel and the court regarding objections, representations to the court, and pleadings filed in the case. In this article, litigants' views are referred to as plaintiffs' or defendants' arguments or opinions with a citation only for direct quotes. Expert witness reports documenting most of the litigants' positions included: **Plaintiffs:** Haney, W. M., *Preliminary Report on Texas Assessment of Academic Skills Exit Test (TAAS–X)*, December 11, 1998; Haney, W. M., *Supplementary Report on Texas Assessment of Academic Skills Exit Test (TAAS–X)*, July 30, 1999; Shapiro, M. M., *Declaration of Martin M. Shapiro*, November 23, 1998; Bernal, E. M., *Report of Dr. Ernesto M. Bernal*, no date; Bernal, E. M., *Item-factor Analysis of the 1997 TAAS Exit-level Tests*, July 30, 1999; Fassold, M., *Written Report of Mark Fassold*, November 15, 1998; Fassold, M., *Revised Written Report*, August 2, 1999. **Defendants:** Phillips, S. E., *The Texas Assessment of Academic Skills Exit Level Test*, January 1999; Mehrens, W. A., *GI Forum, et al. v. Texas Education Agency, et al. Expert Report*, January 1999; Treisman, P. U., *Preliminary Expert Witness Report*, February 25, 1999; Porter, R. P., *Declaration of Rosalie Pedalino Porter*, January 7, 1999.

[12]American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). *Standards for Educational and Psychological Testing* [hereinafter *Test Standards*]. Washington, DC: American Psychological Association, 1985.

[13]Smisko, A., Twing, J., & Denny, T., The Texas Model for Content and Curricular Validity. *Applied Measurement in Education* (this issue).

[14]See the *1996–97 Technical Digest*, Appendix 5, for a description of item review considerations used by the educator committees.

[15]Bernal report at 3.

[16]Haney report at 31–32.

[17]Porter, R., Accountability is Overdue: Testing the Academic Achievement of Limited English Proficient (LEP) Students. *Applied Measurement in Education* (this issue).

[18]The following table shows the Texas graduation test reliabilities.

Texas Graduation Test Reliabilities

Ethnic Group	*Reading (48 Items)*	*Mathematics (60 Items)*	*Writing (40 Multiple-Choice Items)*
African-American	.88	.94	.83
Hispanic	.89	.94	.86
White	.86	.92	.81

Note. Reliability estimates reported for writing include only the multiple-choice portion of the test. Writing multiple-choice scores are combined with an essay score to produce a writing total score. Essays are scored holistically on a 4-point scale and scorer agreement (interrater reliability) after three readings is 98%.

[19]Haney supplementary report at 14–15.

[20]Thorndike, R. L., & Hagen, E. P., *Measurement and Evaluation in Psychology and Education*, 4th ed., New York: Wiley, 1977, p. 79.

[21]The data were: Math retest 8 months = .85, 24 months = .84, 48 months = .64, KR_{20} = .94, Equivalent forms = .88; Language retest 8 months = .88, 24 months = .89, 48 months = .75, KR_{20} = .96, Equivalent forms = .91. Thorndike, R. M., *Measurement and Evaluation in Psychology and Education*, Prentice Hall, Upper Saddle River, NJ, 1997, p. 115.

[22]*Id*. at 102, 100.

[23]Haney supplementary report at 35.

[24]Personal recollection of trial testimony of named plaintiffs, confirmed by defendants' trial attorneys.

[25]Debra P. v. Turlington, 564 F. Supp. 177, 186 (M.D. Fla. 1983).

[26]Kerlinger, F. N., *Foundations of Behavioral Research*, 2nd ed., Holt, 1973, p. 414.

[27]*Debra P.*, 730 F.2d at 1410–11.

[28]*Debra P.*, 654 F.2d at 1088.

[29]Summary judgment order at 6, ftn. 3.

[30]Haney supplementary report at 7.

[31]*Id.* at 17–21.

[32]Note: No indication was given as to what evidence would verify that a student who failed the graduation test was "qualified." Furthermore, this wording may have been confusing because the consequence for failing a single administration of the graduation test was remediation, not denial of a diploma.

[33]Haney supplementary report at 16.

[34]Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607 (1985).

[35]Note: Increases in initial passing rates for White students were somewhat limited by ceiling effects.

[36]Haney original report at 7.

[37]*P*(pass the graduation test in 8 attempts | true achievement at the passing standard & no remediation) = .996.

[38]Sharif v. New York State Educ. Dept., 709 F. Supp. 345 (S.D. N.Y. 1989).

[39]Wards Cove Packing Co. v. Antonio, 109 S.Ct. 2115, 2127 (1989); United States v. South Carolina, 445 F.Supp. 1094, 1116 (D.S.C. 1977), *aff'd mem.* 98 S.Ct. 756 (1978).

[40]Shapiro report at 2.

[41]*See* Phillips, S. E., *The Golden Rule Remedy for Disparate Impact of Standardized Testing: Progress or Regress?*, 63 Educ. Law Rep. 383 (Dec. 20, 1990) for a discussion of the *Golden Rule* settlement and quotations from measurement professionals.

[42]Shapiro report at 5.

[43]Bernal factor analysis report at 6.

[44]Headden, S., *The Hispanic Dropout Mystery*, U.S. NEWS & WORLD REPORT, October 20, 1997, p. 64.

[45]Haney supplementary report at 44.

[46]Zehr, M. A., *Hispanic Students "Left Out" by High-Stakes Tests, Panel Concludes*, EDUCATION WEEK, September 22, 1999, p. 5.

[47]Balli, C., *Student retaking TAAS exam*, San Antonio Express-News, October 27, 1999, p. 14.