DOCUMENT RESUME

ED 349 824 FL 020 643

AUTHOR Geisinger, Kurt F.
TITLE Testing Limited English Proficient Students for Minimum Competency and High School Graduation.
PUB DATE Aug 92
NOTE 55p.; In: Focus on Evaluation and Measurement. Volumes 1 and 2. Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd, Washington, DC, September 4-6, 1991); see FL 020 630.
PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Graduation; High Schools; High School Students; *Limited English Speaking; *Minimum Competency Testing; Student Evaluation; Student Needs; Test Bias; *Testing; Test Validity

ABSTRACT

At present, states have no consistent manner in which limited-English-proficient (LEP) students are assessed on statewide or district-level minimum competency examinations. In some states, LEP students need to take the same minimum competency examinations under the same rules as other students to graduate or be promoted. Competency tests have the capacity to improve the education of the students in our country's schools. To be effective, however, they need to be linked closely to instruction. That is, they need to have high instructional and curricular validity. Furthermore, the curriculum needs to drive the content of the examinations rather than vice versa. One must question whether a minimum competency test can possibly be equally valid from the perspective of curricular and instructional validity and not biased for LEP students, on the very basis of their differential needs and educational programs. For competency tests to be most useful for improving the education of LEP students, it is imperative that the tests be closely tied to the curriculum, be thoroughly integrated with the curriculum, aim toward providing diagnostic instructional and remedial feedback, provide scores that are readily interpretable by educational professionals, and become less threatening than they appear to have become. Responses to the paper by Michelle Hewlett-Gomez and Lawrence Rudner are appended. (VWL)

Testing Limited English Proficient Students for Minimum Competency and High School Graduation[1]

Kurt F. Geisinger
Fordham University

The Current Status of Minimum Competency Testing

Twenty years ago, only a handful of states in our union required students to pass a statewide examination to receive their high school diploma. Today, statewide high school competency tests are in use, in one form or another, in at least 40 states (Jaeger, 1989; Roeber, 1990). From an historical perspective, the widespread use of such examinations and assessments probably grew out of the "back to basics" movement which emerged in response to charges that many of the graduates of our educational system lacked the fundamental academic skills of reading, writing, and mathematics necessary to succeed in adult life, to hold useful and meaningful jobs, and to serve as responsible citizens. From the more limited psychometric or educational testing perspective, such tests probably developed out of the "criterion-referenced testing" movement which occurred in the period from approximately the mid-1960s through the early 1980s. The purpose of such tests was to integrate educational tests more meaningfully into the instructional process by reflecting exactly what knowledge, skills, and other educational behaviors students "mastered" and on which they therefore needed no further instruction. Criterion-referenced tests (CRTs) emphasize scores relevant to the knowledge domain and strongly de-emphasize comparisons of individual students with other children composing their norm group.

To combat charges that students were graduating from high school without being able to read, states imposed tests that students would need to pass to earn their high school diplomas, regardless of how well they had performed (e.g., in terms of grades) in their educational course work. In some cases, the tests were mandated by a state's department of education, or a similar body responsible for monitoring education within its jurisdiction. In other instances, the state legislature imposed the testing program on the educational community. Such tests then serve as a guarantee to society at large, parents, schoolchildren, potential employers, and others that high school graduates possess at least those minimal skills usually deemed necessary for successful survival in the modern world. Clearly, the responsibility for ensuring that those educated within a

given school district or state falls on both those charged with monitoring education within the jurisdiction and those with overall responsibility for governing the region.

Although the use of minimum competency tests for scrutinizing whether students have acceptable levels of the skills measured by the tests to be promoted or graduated is the most visible use of these tests, it should be noted that there are other uses to which the scores may be put as well. For example, Roeber (1990) provides a number of additional uses: appraising the general mastery by students of the state curriculum (for program evaluation uses); providing general information to policy makers, educators, or the public; system accountability; system planning and resource allocation; and system improvement. In addition, when administered prior to the terminal year of a student's education, the test may be used to help to direct a student to a particular school stream (e.g., vocational or academic system) (Roeber, 1990. p. 7-8).

For those states in which passing the minimum competency test was required for high school graduation, states frequently phased in their offering of minimum competency examinations by one of several methods: by administering them to the first one or more classes on a trial basis, by permitting students to take the tests on several occasions in order to pass, by gradually raising an initially lower standard of performance until the appropriate, planned passing score is achieved, or some combination of these techniques. Students need only pass the test once; there is no requirement that they demonstrate continued competency once they have passed the examination.

Most states that began offering such tests did so by administering them at least a year or more before the expected date of graduation, so that students who failed to pass them would be able to retake the tests on one or more future occasions in order to graduate from high school on schedule. The Director of Testing for one state, which employs a high school graduation test, recently stated that students who repeatedly fail the test and take it each time it is offered could conceivably take it as many as 11 times. He further informed me that some students had indeed taken the test this number of times. In other states, of course, the number of possible retestings is considerably reduced. In North Carolina, for example, students must pass both reading and mathematics tests; they are given a maximum of four trials to pass each (Jaeger, 1989). Students who never pass but choose to leave high school typically receive a certificate of completion or some similar acknowledgment that the student completed his or her studies but did not graduate.

In many states, simply passing the competency tests does not ensure that a student will receive a high school diploma. Rather, it is one requirement among several, such as satisfaction of attendance

policy, curricular breadth, and quality of academic course work requirements.

Many states that administer minimum competency tests also offer other preliminary tests at earlier points in the movement of students through the educational system. The purposes of these examinations is to identify those students who have fallen behind and who are likely to have problems when they eventually take the statewide graduation test. Students who perform poorly at these earlier grades may receive additional instruction, embellished instruction, and other special remedial services or be held in grade until they pass this preliminary examination.

How different states use minimum competency tests varies widely (Jaeger, 1989). The majority of states set standards which all students throughout the state must pass to earn their high school diploma. Others, however, permit each school district to set individual standards specific to their own school district. Still others do not require a passing score for high school graduation.

The passing of P.L. 94-142, the Education of All Handicapped Children Act, in 1975 mandated that all children be provided with a free and appropriate education, regardless of handicap. After this law went into effect, most states had to accommodate students with handicapping conditions and other disadvantaged students differentially. The law required the development of Individualized Educational Programs (IEPs) for all students with handicapping conditions. Description of the following items were mandated for inclusion in IEPs: a statement of the present levels of educational performance, short-term and long-term goals and objectives, specific educational services to be provided, the extent to which the student should participate in regular educational programs, the projected date for initiation of remedial services, the duration of the remedial services, and "appropriate objective criteria and evaluation procedures and schedules for determining, on at least an annual basis, whether instructional objectives are being achieved" (Willig & Ortiz, 1991, p. 282). Students with handicapping conditions, on the basis of their IEPs, can be rightfully exempted from the requirements to take and to pass the minimum competency examination in order to graduate from high school. However, unless LEP students also fit the criteria for handicapped status, no IEPs are developed for them. Because of this exclusion, "educators frequently fail to consider cultural/linguistic learner characteristics and their effects on the teaching-learning process" (Willig & Ortiz, 1991, p. 282). Thus, although some LEP children may be considered exceptional and others reside in states which provide special status to LEP students, most still need to pass statewide competency tests.

The Typical Content of State Minimum Competency Tests

A 1979 review (Gorth & Perkins, 1979) summarized the content of statewide competency examinations. (This information is summarized by Jaeger, 1989.) In general, two overlapping types of content are called for by these examinations: the basics of education (reading, writing, and arithmetic) and what are sometimes called "survival" skills for adults in our society. There is, of course, much overlap between the two content areas. Gorth and Perkins reported that over one-half of the states then using such examinations employed tests composed of multiple-choice questions of reading, writing, and arithmetic. In general, these examinations called for the students to demonstrate "nothing more than recognition of basic subject matter mechanics or the application of basic mechanics to so-called 'life skills' situations" (Jaeger, 1989, p. 510). Indeed, the tests were seen as measuring skills learned primarily at the elementary school level rather than either those drawing upon the high school curriculum or higher-order thinking processes.

Increasingly, states and district-level minimum competency examinations are including performance assessment components as parts of their competency testing program in addition to the traditional objective, multiple-choice test components. Such assessments are seen as differing from multiple-choice testing in that (1) students create responses rather than selecting them, (2) performance assessments emphasize problem solving and other higher-level integrative cognitive skills, and (3) performance assessments need to be scored by expert judges rather than machines (Finch, 1991). Because the skills that students use in generating their responses and the products that result from their responses are sometimes seen as more like those skills and products found in the classroom, performance assessment has occasionally been called authentic assessment. Among the types of performance assessment that are used are essays, sometimes with prompts provided; actual student writing samples; prepared portfolios which document the accumulated work of a student; problem solutions such as lab reports in the sciences; and reviews of productions in the realm of art and music. The most commonly used performance assessment component is the writing sample or essay as a measure of student writing ability (Roeber, 1990). These are sometimes administered as part of the examination process and in other settings students may write their essays during a time period of several weeks. A number of states are currently making efforts to increase their utilization of this form of performance assessment, especially in math and the sciences. The development and scoring of performance assessment measures is an extremely expensive undertaking. Therefore, performance assessment is likely to remain a component of minimum competency testing in conjunction with objective

measurement (e.g., multiple-choice tests) and/or as an alternative assessment device for those individuals who fail the objective test on one or more occasions.

The American Achievement Tests called for by the Department of Education in the AMERICA 2000 report (U. S. Department of Education, 1991) would seem to draw upon similar skills, although they would appear to be both heightened in terms of difficulty and level of cognitive processing and broadened in scope. Five subject matter areas will be addressed (English, mathematics, science, history and geography), although when the tests are first introduced, they may be limited to an assessment of reading, writing and arithmetic. The tests would appear to be conceived as both tied to subject matter and to broader thinking skills as are more typically found in tests of cognitive abilities than are subject-area tests. Like other competency examinations, preliminary competency tests will be administered at earlier grades. Thus, "American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy" (U.S. Department of Education, 1991, p. 9). Frankly, one might legitimately question whether modern psychometrics, educational testing, and psychology have advanced to the stage of being able to identify those skills necessary for responsible citizenship, much less to measure them. It may be noted, however, that the American Achievement Tests called for in the AMERICA 2000 report are not necessarily conceived of as minimum competency tests that would be used for high school graduation. It is also suggested that they be used for college admissions and employment decision making, for example. How a single test could meet these varying purposes is not clear and difficult to imagine from the current status of test construction and theory.

Assessing LEP Students with Minimum Competency Examinations

At present, states have no consistent manner in which LEP students are assessed on statewide or district-level minimum competency examinations. In some states, LEP students need to take the same minimum competency examinations under the same rules as other students to graduate or be promoted. In other jurisdictions, however, exceptional LEP students and those residing in locales that require individualized educational programs for LEP students may be exempted from the examination if their IEPs do not require them to take the examination to graduate from high school. (Such a plan

is similar to a common approach for waiving this requirement for special education students.) In yet other locations, LEP students may be permitted to take the examination in their native language, or at least in some common languages, if they enter the American educational system late in their formal education. In some schools, when students fail one competency test, they are given the option to take an alternative measure, perhaps a performance assessment or a test in their native language. In still other settings, they may only take the examination if they have first failed the examination in English. Furthermore, these options simply sample some of the possibilities. Thus, there is a wide variety of choices from which the educational community may select in deciding how LEP students should be tested with minimum competency examinations.

A few examples may demonstrate the diversity of options available. In Connecticut, LEP students are required to be tested unless a planning and placement team decision rules otherwise. In Florida, LEP students are exempt from taking the graduation test during their first two years in an English-speaking school, but are still required to pass the graduation test to qualify for a regular diploma. Similarly, in Michigan, non-English speaking students enrolled in schools in the United States less than two years may be excluded from taking the tests. In Ohio, students may defer taking the test, but may not earn a diploma without passing it. In Georgia, English-as-Second Language (ESL) students take the tests unless the school and parent(s) or guardian agree it is not in the best interest of the student to take it in its current administration. In Maryland, LEP students must pass all four of the examinations that they offer to earn a high school diploma.[2]

Little guidance is available from the educational research literature regarding which of the possible approaches to testing LEP students would be preferred. In fact, a computerized literature review of the ERIC database using either competency and minimum competency examinations and limited English proficiency students as key words or descriptors yielded no references. Certainly, survey efforts similar to those routinely performed by national organizations (e.g., Roeber, 1990) to document the policies each state follows with regard to LEP students would be a most helpful first step. Full-blown comparative evaluation studies such as those performed on the Headstart evaluation contrasting the effectiveness of different strategies in working with LEP students are needed.

Methodological Issues in Minimum Competency Testing

Validity

Validation has been thoroughly described by Messick (1989), Anastasi (1988), and others and need receive only a cursory treatment here. Validation refers to the process of documenting that a test is being used in a justifiable fashion, typically as determined by research studies providing documented evidence supportive of its planned use. Cronbach (1971) has sometimes been credited with the notion that we do not validate tests, rather we validate the accuracy of inferences that we make from test scores. There are generally three acknowledged models or approaches to test validation: criterion-related, construct-related, and content-related. With regard to construct validation, Anastasi (1988) has written:

> The construct-related validity of a test is the extent to which the test may be said to measure a theoretical construct or trait....It derives from established interrelationships among behavioral measures. Construct-related validation requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations represent appropriate evidence for this validation (p. 153; also cited in Geisinger, in press b).

"Criterion-related validity is based on the degree of empirical relationship, usually in terms of correlations or regressions, between the test scores and criterion scores" (Messick, 1989, p. 17). What differentiates orthodox criterion-related validation from the other typically empirical method of validation, (i.e., construct validation), is that criterion-related validation focuses upon "selected relationships with measures that are critical for a particular applied purpose in a specific applied setting" (Messick, 1989, p. 17). The basis for making assessments regarding content validation is "professional judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain" (Messick, 1989, p. 17).

In 1974, Standards for Educational and Psychological Measures recommended for the first time that social consequences of testing such as adverse impact and test bias, should be considered in evaluating a test (Geisinger, in press b). Indeed, Messick (e.g., 1975, 1980, 1989) has argued that values should guide both test use and test evaluation and, hence, such factors need to be considered in evaluat-

ing the use of tests and other measurement procedures. The use of tests with groups underrepresented in many settings within our society, such as LEP students, clearly invokes the values component of the evaluation of these measures.

Since minimum competency tests are the present focus and as educational tests represent a domain (of basic educational skills such as reading, writing, and arithmetic), content validation is the strategy most commonly associated with the corroboration of their use.

Content validation. The basis for content validation is "professional judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain" (Messick, 1989, p. 17). Educational measures developed and validated using content validation involve carefully developed domain specifications based upon curricula, studies of actual instruction provided and educational goals. Content validation comprises both the relevance of the content called for by the test domain (or plan) as well as judgments regarding how well the test ultimately represents the test plan or domain.

Standard 8.4 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985) deals with competency tests and is found below.

> When a test is to be used to certify the successful completion of a given level of education, either grade-to-grade promotion or high school graduation, both the test domain and the instructional domain at the given level of education should be described in sufficient detail, without compromising test security, so that the agreement between the test domain and the content domain can be evaluated (p. 54).

Because a test could meet the traditional standards of content validity yet fail to meet the criteria specified in Standard 8.4 above, demonstrates that traditional approaches to content validation do not provide the specificity called for by Standard 8.4. New concepts were required to gauge meaningfully the value of a minimum competency examination. These terms are curricular validity and instructional validity.

Curricular validity. The notion of curricular validity was introduced by McClung (1978, 1979). It is not a traditional type of validation called for by professional standards (such as content validation), but it has nevertheless, become an important principle in the evaluation of minimum competency examinations, especially in court cases (e.g., *Debra P. v. Turlington*, 1979, 1981, 1983, 1984). "Curricular validity is a measure of how well test items represent the objectives

of the curriculum. An analysis of curricular validity would require comparison of the test objectives with the school's course objectives" (McClung, 1979, p. 682).

Instructional validity. A second characteristic that should be present in competency tests has been called instructional validity. "Instructional validity is an actual measure of whether the schools are providing students with instruction in the knowledge and skills required by the test" (McClung, 1979, p. 683). An assessment of instructional validity would require proof that students are actually exposed pedagogically to the content covered on the examination. Assuming that a state's minimum competency examination is valid for the majority of students in a state, an important question when considering the testing of LEP students is whether their instruction parallels that of the majority students. The concept of differential validation impacts this judgment.

Differential validity or population validity. Differential validity (sometimes referred to as population validity) is a concept closely aligned with that of test bias. It has traditionally been used in criterion-related validation studies, primarily with regard to admission to higher education and employment decisions. A test may be said to be differentially valid if its validity differs across subgroups of test takers. Predictive tests are differentially valid if the empirical relationships between the predictive test and a measure of criterion performance differ systematically across groups. To assess whether a test is differentially valid across groups, one must perform regressions between test and criterion variables for each population. Then, the slopes of the varying regression lines, their respective intercepts, and the degree to which the relationships are free from statistical error are compared. (See Anastasi, 1988, pp. 193-199 for an elaboration of this concept.) "Validity coefficients, regression weights, and cutoff scores may vary as a function of differences in the test takers' experiential backgrounds" (Anastasi, 1988, p. 194).

It is not clear how or why differential validation studies would be performed using minimum competency examinations since there is no criterion representing the kind of behavior that minimum competency tests attempt to measure or predict. In the present instance, however, there is no obvious, relevant criterion for a test of minimum competency. Possible criteria include teacher judgments, high school grade-point average (GPA), subsequent college GPA, etc., but all of these criteria have methodological problems and, more importantly, they lack relevancy in that none of these criteria bear on the meaning and purpose of such tests.

The concept of differential validity can nevertheless be generalized to the testing of LEP students. A competency test might be differentially valid in terms of *instructional validity* if the material cov-

ered on the examination is not equivalently presented to the majority students and LEP students. That is, if the material composing the examination is more to be found in the classroom of traditional students than it is in bilingual classrooms, a case could be made that the test is differentially valid and, in this instance, biased against the LEP students.

Reliability

The study of test reliability has largely been the study of the consistency of the test scores that individuals achieve across different administrations of the same test, across different test forms, across different test administrators (especially for individually administered examinations), and across the individual questions composing a single test. Each of these kinds of reliability indicates a somewhat different generalization about which we may have a degree of confidence when we talk about an individual's test score. Such depictions of test reliability do indeed have relevance for competency testing, but the relevance needs to be reformulated to a degree. The stability and consistency of an individual's score are important, but the degree to which the decisions made with the examination do not change is even more critical. These approaches are known as the decision-consistency approaches to test reliability. Excellent reviews of the literature on the reliability of tests scored in a pass-fail manner may be found in Berk (1984), Brennan (1984) and Subkoviak (1984).

The notion of stability over different testings must be clarified with regard to competency testing. Classical reliability theory, upon which the notion of both the test-retest and alternate-forms approaches to test reliability, are based, assumes no change in the underlying variable being measured. In the case of a competency test, however, it is certainly hoped and expected that instruction -- remedial and traditional -- increases the competency of the student between the first and second testings. Thus, the assumptions of the classical reliability model are clearly violated in the case of minimum competency testing. Indices lower than those that might otherwise be acceptable may be tolerated due to these expected changes over time. That is, since students are engaged in learning in their educational activities, their performance on the minimum competency tests changes from the first administration until the second. When this learning process occurs, it appears as though the test is less reliable (stable) than it really is.

Standards of Performance

Techniques used in setting standards. All certification testing requires that the performance of individual students be compared with or evaluated against a predetermined standard of performance.

A decision is made regarding each student in terms of whether that student is competent. The degree of competence is not critical as it is in most tests of individual differences or so-called norm-referenced tests. The only scoring that is critical is whether the student has met the minimal standard or not.

While psychometricians have developed theories for test reliability and validity, the development of such approaches for the setting of standards on examinations is in its infancy. A number of techniques or strategies for setting standards on educational and other psychological measures have been proposed, but it is agreed that there is no way to prove that one technique is better than any other. "While there is no agreement on a best method, ...some procedures are far more popular than others" (Jaeger, 1991, p. 491). Standard-setting procedures are based on pragmatics, not science. All of the techniques require that those setting the test standard impose their professional judgment to the task. To some (e.g., Glass, 1978), the judgments involved in these tasks are intrinsically arbitrary and therefore of questionable value. One reason that some testing professionals exhort caution in the standard setting process is that in choosing among the available standard setting techniques, one influences the standards to some degree. Similarly, the judges one uses in establishing test standards also impact the standards to a substantial degree (Jaeger, 1991).

The techniques employed in setting standards have been presented by Livingston and Ziecky (1982) and reviewed completely by Jaeger (1989); such detail is certainly beyond the scope of the present discussion. However, most graduation tests set standards by holding panels which review the test item-by-item to determine what the appropriate passing score should be. Such approaches are what Hambleton and Eignor (1980) call judgment models since they rely on the judgments of the panel members. The most common of these models is what has become called the Angoff procedure (after Angoff, 1971). In this procedure, a panel of judges is convened and each member of the panel reviews each question on the test and estimates the probability (a proportion from 0.00 to 1.00) that a minimally competent student would answer each correctly. These estimates are summed for each judge and then the individual judges' estimates are averaged. The resulting value becomes the passing score. The advantage of this procedure is that the passing score that is set is specific to the test in question and is based on judgments of those presumably knowledgeable to make such judgments. Among the disadvantages are the difficulties in determining what a "minimally competent" student would be, much less how he or she would perform on the test.

A few variations to the standard Angoff procedure may be employed. For example, one can have the judges themselves take the

examination prior to their making judgments about the test questions. One can provide the judges with item analysis data so that they can see how test takers actually performed on each test question. One can also iterate the Angoff procedure several times with the same or different panels and provide each successive panel with the results of the preceding judgments. Another modification is to permit the judges to select a probability that a minimally competent test taker would answer a question correctly from a shortened list of the possible values from 0.00 to 1.00.

To be able to make ratings on specific items, as in an Angoff panel, a clear understanding of what minimum competence means is needed. Mills, Melican and Ahluwalia (1991) have addressed techniques to use with Angoff panelists to help them understand the multiplicity of different interpretations of the minimum competence concept. For example, the panel may begin by listing the levels of knowledge and skills that such an individual might possess. Those running the meeting need to keep the panelists focussed on the target individual -- a person graduating from high school with the least amount of knowledge and skills permissible. To be able to make such judgments, those making the ratings must be knowledgeable about the full range of skill levels of graduating seniors and about the curriculum and instruction that such students receive.

Jaeger (1991) has addressed the issue of who should compose a standard-setting panel. Standard 6.9 of the Standards for Educational and Psychological Testing (AERA et al., 1985) requires that the qualifications of judges composing the panel should be enumerated in an appropriate publication -- an implication that such factors have relevance. The task of serving on a standard-setting panel is a complex one involving the reading, understanding, and evaluating of a vast amount of detailed information typically in a confined time period. Judges must possess "substantial knowledge...that is rapidly accessible and readily integrated" (Jaeger, 1991, p. 3). Jaeger defines the ideal judges as experts.

> In the case of a high school graduation test, such individuals know the knowledge requirements of entry-level, post-high-school jobs or freshman courses in colleges and universities, assuming the purpose of high school graduation tests is to ensure that high school graduates possess knowledge sufficient to enter the labor force or enter a post-secondary education program. Judges most likely to possess this kind of expertise include directors of apprenticeship programs for craft unions, personnel directors of service-oriented companies that hire large numbers of recent high school graduates, college and university admission officers, and college and university faculty members who teach freshman courses (Jaeger, 1991, p. 4).

Mills et al., (1991) supplement Jaeger's recommendations by suggesting that "panelists in standard-setting studies should be chosen to represent all appropriate groups in the profession relevant to establishing the cutoff scores for the test. These panelists, therefore, will bring a diversity of knowledge, training, and opinions about the test and testing situation to the rating session" (Mills et al., p. 9). In the instance of setting standards that affect LEP students, such panels should probably include ESL instructors and others knowledgeable about the performance of LEP students.

It may be recalled that only one procedure for setting a standard has been described.

> A large number of empirical studies have addressed the question of whether different standard-setting procedures, when applied to the same competency test, provide similar results. Most research has answered this question negatively. Different standard-setting procedures generally produce markedly different test standards when applied to the same test, either by the same judges or by randomly parallel samples of judges (Jaeger, 1989, p. 497).

That different panels of judges and different procedures may elect to set varying standards has led some scholars (e.g., Jaeger, 1989; Shepard, 1980) to suggest using several methods in combination and then "consider all of the results, together with extrastatistical factors when determining a final cutoff score" (Jaeger, 1989, p. 497).

Adjustments made to initial standards. Geisinger (1991) has provided a list of some of the kinds of information that may be used to adjust the proposed passing scores that emerge from standard setting panel meetings. With respect to high school graduation tests, this information includes: (a) what passing rates/failing rates are acceptable to relevant parties; (b) the relative costs of misclassification errors (e.g., failing someone who should have passed); (c) societal needs; (d) adverse or disparate impact data; (e) errors of measurement due to the test's unreliability; (f) errors of rating due to differences among raters within a standard-setting panel and across different panels; (g) anomalies in the rating process (e.g., judges who are found to lack the expertise required of them); (h) how frequently and how often students are able to re-take forms of the examination; and (i) results of other standard-setting procedures. One can imagine several of these adjustments that are relevant for the assessment of LEP students. Most obvious, of course, is (d) adverse impact data. If the proportion of Hispanics passing the test, for example, is sufficiently below that of other groups, test makers, educational leaders and other concerned parties should review the results as well as the education of the students involved to consider

what should be done. Perhaps some adjustment either to the overall passing point or the passing point for Hispanic test takers may be in order. A more subtle example concerns (e) test reliability. Passing scores are sometimes adjusted (typically in a downward direction) due to unreliability. Students who fall just below the passing score are seen as being strong contenders for passing the test, if it were only more reliable. The reliability coefficient and, more importantly, the standard error of measurement for LEP students taking the examination should be computed and compared to that of the majority students. If the reliability is lower and the standard error of measurement higher, an argument for a reduction in the passing score for LEP students would appear justifiable. As a final example, consider (c) societal needs. Paulson and Ball (1984) have argued that minorities were not as able to receive employment in the State of Florida after the high school graduation test was announced. Such information might argue that the test standard be reduced. On the other hand, if the results of testing are used to provide high quality remedial education to the LEP test takers who fail and this remedial education provides LEP students with improved academic skills without consequential personal, social, or academic costs (e.g., stigmas), then the competency test standard should be kept where it is or even increased.

There may be circumstances in the use of minimum competency examinations where it is appropriate to employ a different standard as the passing score than is used in the general population. In some instances, LEP students have already been identified for special test administration procedures such as being excluded from taking the examination altogether on the basis of an IEP or a similarly institutionalized policy, bypassing the first test administration for which they are eligible, having the test administered in their native or first language, or taking an alternative measure. Under such circumstances, it may also be appropriate to use a different passing score in the recognition that their more limited English skills inhibit their best performance. Padilla (1979) suggested a similar notion with regard to employment settings in noting that there are situations in which it is appropriate for job candidates to be essentially given "extra-credit" for being bilingual. "In job settings where such bilingualism is functionally related to job success, such credit is indeed appropriate, although it is rarely given in civil service settings, for example. Such bonuses, appropriately awarded because the language skills enhance job performance, should be clearly seen as additional to any other advantages provided to members of language minorities in the attempt to increase their representation in the work force, on campuses, in advanced instruction, etcetera. Credit for being bilingual (French/English) is appropriately provided to managers in the public service of Canada, for example" (Geisinger, in press, a).

Methodological Issues Specific to LEP Students

Test Bias

Test bias is intrinsically and closely tied to the concept of test validity because, like validity, it rests primarily upon inferences based on test scores. As in the case of all judgments of test validation, threats to validity threaten proper test score interpretation. Just as validation was dominated by the criterion-related approach until the last decade or two (see Geisinger, in press, b), so has the study of bias been dominated by the criterion-related approach. Many of the definitions of bias that have been traditionally provided are difficult to extract from the criterion-related validation paradigm.

One definition of bias (Cole & Moss, 1989) moves beyond the criterion-related model. This definition states:

> An inference is biased when it is not equally valid for different groups. Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications for the remainder of the test takers. Thus, *bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers* (p. 205).

Such a definition, as is shown below in this paper, has implications for the competency testing of LEP students. With respect to the content validation approach, problems in making valid inferences may be based upon differences across groups with regard to the appropriateness of a given domain of content or testing format or for how well specific questions cover the content domain. For example, when Spanish-speaking tenth grade students write responses on an essay final examination in History, the quality of their responses may be limited by their ability to write the answer in English. A source of test score variance becomes English writing ability and inferences which assume that the scores are solely due to knowledge of History are incorrect. (This information has been cited in Geisinger, in press, a.)

Bias detection techniques. Test bias has been scientifically studied for several decades. Typically, reviews of test bias research subdivide the procedures which have been developed into external and internal methods. External methods are those that evaluate whether the relationship between test scores and extra-test criteria is comparable across groups. There are two types of internal methods. The first attempts to identify those test questions which are differentially more difficult for a given group than other questions com-

posing the test. The second involves factor analyses of test items to identify dimensions of test performance for each of the groups under study. The attempt is to show that the test measures the same, similar or different characteristics across the varying groups. If similar factors are found across groups, we have some reason to suppose that the test measures comparable constructs in each group. This second approach is not discussed in the present paper but may be found in Geisinger (in press b), Reynolds (1982b), or Shepard (1982).

Reynolds (1982a) offered the following definition of test bias from the perspective of construct-related validation:

> Bias exists in regard to construct validity when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with differing degrees of accuracy (p. 194).

Reynolds (1982b) suggested a number of different empirical techniques in which construct-related test bias might be identified. These include differences across reliability coefficients; rank ordering of item difficulties; correlations with other variables, such as age; comparisons of multitrait-multimethod matrices across groups; and factor-analytic differences. To know definitively whether competency tests measure the same cognitive processes for LEP and majority-group students requires research of this type. However, the numbers of LEP students, especially when subdivided by cultural or ethnic group, would likely prohibit such efforts.

In any test of cognitive skill or ability, a logical and elementary check that the test is comparable across groups is a determination that the relative difficulty of test questions is similar. That is, those questions which are difficult for one group should also be challenging for the other. Should such a finding not hold, one must question population validity of the underlying construct. If test items are rank ordered from easy to difficult within each group, rank order correlations, such as rho, may be calculated to demonstrate parity. Reynolds (1982b) suggests that rho's of .90 be taken as indicative of consistency of construct-related validity.

Test bias against LEP students. In some instances, the study of test bias against LEP students differs from that of other groups, such as African Americans, females, or the handicapped. The study of test bias is more difficult with language minorities because there are at least two ways in which such test bias differs from that against females, African Americans, and to a lesser extent, the handicapped. The first of these relates purely to language differences, both in test administration and in the interpretation of test results. The second situation considers differences between LEP students and the other

groups in our society stemming from cultural factors, with these cultural differences including those related to language.

Two primary factors confound the interpretations of some tests with LEP students: language and culture. Neither of these problems is easy to deal with, but language has received more recent attention from the psychometric literature (e.g., Duran, 1988, 1989), although it is potentially less complex.

(1.) Language differences. The first issue related to language concerns the question, in which language should the test be administered (Geisinger, in press, a)? If the purpose of the test is to provide diagnostic information so that an IEP can be developed and instruction optimized, administering comparable competency tests in both languages may yield useful information.

When English-language tests are used with LEP students, the level of language applied on the test needs to parallel that used in the schools (as it does with English-speaking test takers). With written tests, various readability formulae may be used to estimate both the reading levels of the examination and of materials used in the classroom and in educational materials to ensure that the test does not require an artificially high reading level. Once again, the concept of differential instructional validity may be relevant if LEP students use educational materials which are generally easier to read than are the test materials. In the scoring of certain free-response measures, such as the essay examinations, the level of English language skill necessary for achieving passing scores on the examination should also be considered when constructing the questions, scoring the responses, and interpreting the results.

The 1985 Standards for Educational and Psychological Testing provide a section on the testing of linguistic minorities. Seven standards were enumerated to provide some guidance toward good testing practice. In general, these standards emphasize attempts to achieve valid inferences from test scores coming from members of linguistic minorities. They accent the notion that tests may be influenced by language skills (irrelevant to the construct purportedly being measured) to a greater or lesser extent when given to linguistic minorities. Thus, these standards attempt to assure valid test use and interpretation. Furthermore, they state that test constructors (in the case of minimum competency examinations), test publishers and state departments of education, developing assessment instruments recommended for use with linguistic minorities need tc inform test administrators and test users of proper procedures and interpretations with those groups. The seven standards follow.

13.1 For non-native English speakers or for speakers of some dialects of English, tests should be designed to minimize threats

to test reliability and validity that may arise from language differences.

13.2 Linguistic modifications recommended by test publishers should be described in detail in the test manual.

13.3 When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation.

13.4 When a test is translated from one language or dialect to another, its reliability and validity for the uses intended in the linguistic groups to be tested should be established.

13.5 In employment, licensing, and certification testing, the English language proficiency level of the test should not exceed that appropriate to the relevant occupation or profession.

13.6 When it is intended that the two versions of dual-language tests be comparable, evidence of test comparability should be reported.

13.7 English language proficiency should not be determined solely with tests that demand only a single linguistic skill. (AERA et al., 1985, pp. 74-75).

The first six standards are clearly relevant to the testing of LEP students. Based on Roeber's (1990) survey, it would appear that few if any states are performing the research needed to ensure that different language test forms and other alternative test forms are comparable and equally valid. Without such information, equivalent interpretations may not be made using the different forms. Unfortunately, it may not be in the best interests of advocates of LEP students to demand such research. If states find it impossible to perform such research for budgetary, manpower, or other reasons, they may simply discontinue offering alternative testings or testings in a variety of languages.

(2.) Cultural differences. Gordon (1991) has defined culture as a complex whole that includes knowledge, belief, art, morals, law, custom and any other capabilities and habits acquired as a member of society. The total pattern of human behavior and its products -- embodied in thought, speech, action, and artifacts -- are dependent on the capacity to learn and transmit knowledge to succeeding generations through the use of tools, language, and systems of abstract thought. As a descriptive concept, culture is a product of human action; as an explanatory concept, it is seen as influencing further action (p. 101).

Scores emerging from tests are indeed subject to cultural influences. Not to reflect culture, of course, would likely mean that the scores do not validly reflect the construct or behavior they have been intended to assess. From two quotes by Anne Anastasi, we may glean when such influences represent valid influences upon test scores and when they do not. First, let us consider the valid perspective.

> Every psychological test measures a sample of behavior. Insofar as culture affects behavior, its influence will and should be reflected in the test. Moreover, if we were to rule out cultural differentials from a test, we might thereby lower its validity against the criterion we are trying to predict. The same cultural differentials that impair an individual's test performance are likely to handicap him (sic) in school work, job performance, or whatever other subsequent achievement we are trying to predict (Anastasi, 1967, p. 299).

Nevertheless, a "test may be invalidated by the presence of uncontrollable cultural factors. But this would occur only when the given cultural factor affects the test without affecting the criterion" (Anastasi, 1950, p. 15; also cited in Geisinger, in press, a).

Proper test score interpretation for linguistic minorities involves consideration of acculturation. "Acculturation refers to complex processes that take place when diverse cultural groups come into contact with one another. It is an extremely important aspect of the experience of linguistic minorities in the United States. Acculturation is also related to testing issues because it involves the acquisition of language, values, customs, and cognitive styles of the majority culture -- all factors that may substantially affect performance on tests" (Olmeda, 1981, p. 1082). Since acculturation can presently be assessed with substantial reliability and validity (Olmeda, 1979, 1981), teams planning the IEPs for LEP students should include formal measures of acculturation when making assessments of these students.

Item selection including item bias detection techniques. Item bias techniques are also known as methods to determine differential item functioning across groups, or DIF, and have been used in the pre-testing phase of test development to identify and remove those questions from a test that are differentially more difficult for one group or another. The manner in which most of the available techniques work may be explained as follows. The only two factors employed by these techniques are the group-specific difficulty level of each of the questions composing the test and the overall level of ability or knowledge of each test taker. The test taker's level of ability or knowledge is generally designated by the individual's overall score on the examination or some mathematical derivation of this value.

The logic of the process is that both the content and thought processes called for by a question determine the difficulty level of that question and should be comparable across groups. Since groups may differ in terms of overall ability for one reason or another, these techniques adjust difficulty levels of individual test questions for these overall group differences. If the difference between difficulty levels for an item for two groups is disproportionally large, even given the groups' overall test differences, the particular test question is considered biased. In the pre-testing of an examination, such questions would likely be removed from the instrument under development.

Issues Relating to Standard Setting

It was reported previously in this paper that the most common technique for setting standards is to bring together a panel which reviews the test on a question-by-question basis and generates data used to set the passing standard. Representation by LEP parents, educators or other delegates should appear on these panels. This representation would permit discussion among the members of the panel of issues of relevance to LEP students. It should be openly questioned as to whether a single standard is appropriate or rather if differential standards should be applied for varying groups of students. In cases where a single standard setting was held at some previous time and the state now simply equates the passing score of new tests to this pre-existing standard, either new standard setting panels should be convened or adjustments to the standard should be considered.

Equating

Scores on one form of a state's minimum competency test are frequently equated to previously used forms so that test scores and passing scores retain their meaning over time. Equating generally involves a sample of students taking part in all of both forms of the examination. It is critical that LEP students be included in representative numbers in these equating samples. In fact, given their small numbers in many locations, they may need to be greatly oversampled.

Furthermore, it is unlikely that states would be able to equate special language versions of the test if they are available. Equating methodology would generally require either that randomly drawn, equivalent groups of individuals take both versions of the examination or that if the same group took both test forms, their language skills be equal in both languages. These assumptions will almost never be met. In addition, the score distributions emerging from English-language tests and foreign language tests are unlikely to parallel each other, with the distribution of foreign test scores well below that of the English-speaking test takers.

Instructional Feedback

Competency testing laws in many states require that those students who fail the minimum competency test receive remedial instruction so that they may succeed when they re-take the examination. Ultimately, the key to the successful competency testing of LEP students involves a proper diagnosis of their academic weaknesses and strengths as well as the development of a well formulated educational plan to remediate their shortcomings in an optimal manner. A successful diagnostic test must be sensitive to their instruction at a micro-level (see Duran, in press) and able to yield reliable information about exactly what the students can and cannot do. Although minimum competency tests are generally group tests, they need to be interpreted on an individual basis, especially for LEP students. That is, educational professionals need to consider the test data along with other indices of educational performance (e.g., work in class), academic skills (e.g., strengths and weaknesses in English, their native language, and other academic information), and knowledge of the setting, broadly defined, in which the child may be found. Furthermore, remedial programs are likely to be negatively tainted and to have adverse impact against ethnic and language minorities. The remedial program in which LEP students are placed must be successful not only educationally but also in terms of overcoming such a stigma. To the extent that the tests and their use are well integrated into the instructional program, they may prove to be successful.

Advantages and Disadvantages of Minimum Competency Testing

In that minimum competency programs have been in effect in one manner or another for more than 20 years, it is disappointing that no comprehensive evaluation studies of these programs have appeared in the literature. If they exist in the files of states, they need desperately to be shared. With the lack of formal evaluations, we must hypothesize and reflect upon the potential advantages and disadvantages of these testing programs from the armchair. Formal evaluations of these testing programs would be strongly recommended before the federal government moves to operationalize the idea of national examinations.

Societal Effects

One advantage to society if the ideal of minimum competency testing were realized would be that society would become filled with adults each of whom is able to read, write, and use basic mathematical skills. Based on the requirements of some states, graduates

would also be able to communicate orally (speak, listen, etcetera) effectively.

From a more negative perspective, the success of minority students on minimum competency tests, as on other examinations, is far below that desired (e.g., Hartigan & Wigdor, 1989). "In states such as North Carolina that maintain statistics on the characteristics of students who fail competency tests, the failure rates of racial minorities are typically found to be 5 to 10 times higher than those of the majority white students. The social and economic consequences of failing to earn a high school diploma are well-known, particularly for youths from minority groups (cf., Eckland, 1980)" (Jaeger, 1989, p. 491). In light of these performance differences, test bias procedures, as mentioned earlier in this paper, may need to be applied. Industrial testing, as opposed to educational testing, has been forced to study the impact of testing when rewards are assigned on the basis of test scores.

The Uniform Guidelines on Employee Selection (1978), issued jointly by the Equal Employment Opportunity Commission, the then Civil Service Commission, the Department of Labor, and the Department of Justice, after considerable input from professional organizations interested in testing practice, operationalized good testing practice in industrial settings in many ways. The Guidelines defined a model of proper test use in which a test need only be shown to be valid for the use to which it is being put after it has first been shown to have adverse impact upon a protected group (defined as Blacks; American Indians; Asians including Pacific Islanders; Hispanics including persons of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish origin or culture; women and other groups). Adverse impact has generally been defined by the "four-fifths rule." That is, "a selection rate for any race, sex, or ethnic group which is less that four-fifths (4/5) (or 80 percent) of the rate for the group with the highest rate will generally be regarded...as evidence of adverse impact" (3D).

The use of minimum competency tests has the potential for real test misuse. Consider the following comments that were distributed to users of the Graduate Record Examination (Graduate Record Examination Board, 1989) as the first of 10 recommended guidelines for proper test use. The guideline states:

> Regardless of the decision to be made, multiple sources of information should be used to ensure fairness and balance the limitations of any single measure of knowledge, skills, or abilities.... Scores should not be used in isolation. Use of multiple criteria is particularly important when using...scores to assess the abilities of educationally disadvantaged students, students whose primary language is not English, and students returning to school after

an extended absence. Score users are urged to become familiar with factors affecting score interpretation for these groups (Graduate Record Examination Board, 1989, p. 6).

Four of the other nine guidelines are also relevant to competency testing. The four specific guidelines suggest that validity studies need be performed, test content be reviewed by subject matter experts, decisions based on small score differences be avoided and test users recognize limitations of scores earned on tests taken under special administrative arrangements (e.g., in a language other than English).

Effects on Students

Blau (1980), a clinical psychologist, has considered what the psychological effects of institutionalized minimum competency testing programs are likely to be. It should be noted from the outset that his sample was small (around 35 students) and apparently gathered from the many adolescents that he saw in his clinical practice. In some cases he was seeing them specifically because of their educational difficulties. Relevant to the present discussion, he also does not report if any of the members of his sample are linguistic minorities. Nevertheless, he reports that the students were "distressed and disdainful about the whole testing business. They saw it as another burden developed by adults to make their progress through school more difficult" (p. 176). With regard to their performance on the test, in this case the Florida high school graduation test, he reported that "the majority of the students, including the very bright ones, simply do not care" (p. 176). The rationale for their apathy was described as differing depending upon how strong the students were. "The poor students saw the tests as an additional barrier to success and esteem and not a help, while the good students saw them as a barrier to using time effectively" (p. 177). One factor appeared to moderate the involvement of students: the immediacy of the feedback that they received. When such feedback was received quickly by students, they did see it as of educational value. In attempting to address how such overly negative attitudes toward the competency testing process might be addressed, Blau called for the involvement of (representatives of) students involved in every stage of the testing process.

One problem that may beset students relates to their moving from one school district or one state to another. Suddenly, different requirements or higher standards impact students. Such problems would be especially notable in situations where school districts set district-level standards for passing statewide examinations. A student might move only a few blocks but, on that account, fail to pass an examination that he or she had apparently already cleared.

Effects on Schools and the Educational Process

The effects of minimum competency testing on the educational process can be considered from a variety of perspectives. One relates to the ultimate goal of the examinations in general. According to Jaeger (1989, pp. 486-87), "Although some competency testing programs attempt to inform students about their academic strengths and weaknesses, the principal use of competency-test results is to serve institutional purposes such as student placement rather than individual purposes such as student guidance and counseling."

One of the biggest fears of those reluctant to endorse minimum competency testing concerns the notion that the low level or minimal skills frequently measured on these tests will become the maximal skills taught by educators. That is, it is feared that teachers will stop striving to teach higher-level thinking and problem-solving skills as long as their students master those basic, life skills called for by the minimum competency examinations.

Legal Issues

Reviews of court records have indicated that courts have continually upheld the rights of states to employ minimum competency tests to monitor the success of educational programs and the skill levels of potential high school graduates (Citron, 1982, 1983; Jaeger, 1989).

The most influential case regarding minimum competency testing brought to date was *Debra P. v. Turlington* (1979, 1981, 1983, 1984). This federal case received considerable attention as evidenced by George Madaus' (1982) book dedicated to the history, effects, and implications of the case. The case, which related to Florida's high school graduation test, was brought by 10 African-American students who had failed the examination and who challenged the adverse impact of the examination against the backdrop of a long history of segregated schools and other forms of discrimination. Florida's minimum competency examination was a test of functional literacy which had been mandated by a 1976 statute requiring demonstration of such skills in order to receive a high school diploma. Functional literacy was defined as skills in reading, writing, and arithmetic needed to face successfully problems encountered in everyday adult life. Reading and writing were combined as a test of communication skills. In its first administration in 1977, 36 percent of high school seniors failed one or both of the examinations, but 77 percent of African-American students failed against 24 percent of white students (Pullin, 1982). "After three attempts to pass the test, 1.9 percent of white students and 20 percent of black students...still failed" (Jaeger, 1989, p. 507) the test. Two sets of claims were made against the test. First, it was argued that the test was discriminatory on the basis of

its adverse impact; it breached the constitutional rights of equal protection under the laws as guaranteed by the Fourteenth Amendment and as enforced under Title VI of the Civil Rights Act of 1964, which prohibited discrimination on the basis of race, among other factors. Since disparate impact had been found in this instance, Florida needed to prove that the differences in passing rates had not been caused by the state's history of discrimination. Second, it was contended that the test was invalid and the requirement that it be passed to earn a high school diploma hastily conceived.

The trial iterated several times between the federal district court level and the appellate court over a period of five years. Initially, a time-period delimited moratorium was placed on the use of the test so that students in the schools were able to prepare for it more properly. The principles of curricular and instructional validity advanced by McClung (1978, 1979) were critical in that the state of Florida was ordered to document that students in Florida schools universally received instruction in the content represented by the tests. The courts initially ruled that punishing the victims of past discrimination "for deficits created by an inferior educational environment neither constitutes a remedy nor creates better educational opportunities" (474 F. Supp., at 257, *Debra P. v. Turlington*, 1979; also cited in Jaeger, 1989). They ultimately ruled, however, based on an overwhelming amount of data indicating that the content of the tests was covered both in curricula throughout the state as well as in actual instruction, that the tests were both fair and valid.

In summary, courts have upheld the rights of states to use competency tests appropriately but have placed limitations on the testing programs (1) when there is a history of discrimination, (2) when students have not been given adequate advance warning about the necessity to pass the tests, and (3) when the curriculum and the instruction provided do not cover the material on the test (Jaeger, 1989).

Using Minimum Competency Tests with LEP Students

Competency tests, like other educational tests, have the capacity to improve the education of the students in our country's schools. To be effective, however, they need to be linked closely to instruction. That is, they need to have high instructional and curricular validity. Furthermore, the curriculum needs to drive the content of the examinations rather than vice versa. One of the most damning indictments of all educational tests is that they determine what is taught in some instances. It may be noted that teaching to a test is not always bad. Providing high quality instruction on topics of high rel-

evance and importance, of course, will always be paramount. However, decisions as to what should be taught are curriculum-level decisions that should be made when developing the curriculum and instructional approach, not after determining what is on an examination.

Making minimum competency tests instructionally meaningful involves more than curricular and instructional validity of the examinations, however. It also entails using the scores to provide access to remedial instruction rather than "as a stick" to punish those who fail, perhaps by withholding diplomas or other valued rewards (see Serow, 1984).

It has been stated that about 2 percent of the students who take minimum competency tests do not pass them, even after repeated administrations of forms of the examinations. It has been further argued by others (Serow, 1984) that this small percentage is politically acceptable to the policy makers who in some cases recommend or require the examinations. Serow reminds us that small percentages of large bodies are still a large number of failures. One wonders if these small numbers would be acceptable if they were being added to welfare rolls rather than being refused a high school diploma.

One issue in discussing the competency testing of LEP students is not a testing matter at all, but purely an educational one. Perhaps all LEP students should have IEPs just as other exceptional students and students with handicaps already do. LEP students are among the most disadvantaged students not so covered at the present time. Should they be provided with the planning and supportive remediation required by an IEP? With such attention, the success rate of LEP students would surely rise.

Summary

Deciding between withholding a diploma from a student who has spent 12 years in ineffective schooling and graduating a student who lacks basic academic and life skills is a no-win choice. The only acceptable solution to this decision is to use the test scores to identify students needing remedial instruction. The most useful such test would be one that is diagnostic rather than summative. However, most statewide competency tests are by their very nature summative tests that do not provide diagnostic information.

One must question whether a minimum competency test can possibly be equally valid from the perspective of curricular and instructional validity and not biased for LEP students, on the very basis of their differential needs and educational programs.

For competency tests to be most useful for improving the education of LEP students, it is imperative that the tests be closely tied to the curriculum, be thoroughly integrated with the curriculum, aim toward providing diagnostic instructional and remedial feedback, provide scores which are readily interpretable by educational professionals, and become less threatening than they appear to have become. Failing scores on competency examinations need to be attuned to the development of IEPs for those LEP students requiring them. The notion that all LEPs would benefit from IEPs has some merit and should be investigated.

The psychometric literature coupled with pragmatic realities of the situation have little to offer at the present time with regard to ways of determining (1) whether minimum competency tests are as valid for LEP students as for others and (2) what passing scores should be used for such students.

Interpretation of individual test scores is extremely demanding. Complex interactions of psychological, language, culture, and other background factors affect the test performance of linguistic minorities. Examiners and educational planners need to be specially trained to test such individuals and to consider language skills, acculturation, socioeconomic factors and other variables in any assessment of an individual's level of functioning.

Notes

[1] The author would like to thank Scott Cone for his help on this paper, Janet F. Carlson for her careful reading of the paper, and Michael Beck of Beck Evaluation Testing Associates, Chris Pipho of the Education Commission of the States and Ed Masonis of the New Jersey Department of Education for their helpful information. Any errors in this paper, of course, remain those of author.

[2] The information provided in this paragraph was taken from Roeber's (1990, pp. 17-18) survey of statewide testing practices.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. American Psychologist, 22, 297-306.

Anastasi, A. (1950). Some implications of cultural factors for test construction. In Proceedings of the 1949 ETS Invitational Conference (pp. 13-17). Princeton, NJ: Educational Testing Service.

Anastasi, A. (1988). Psychological testing (6th ed.) New York: Macmillan.

Anastasi, A. (1990). What is test misuse: Perspectives of a measurement expert. In The uses of standardized tests in American education: Proceedings of the 1989 ETS Invitational Conference (pp. 15-25). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.; pp. 508-600). Washington, DC: American Council on Education.

Berk R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), A guide to criterion-reference test construction (pp. 231-266). Baltimore: John Hopkins University Press.

Blau, T. H. (1980). Minimum competency testing: Psychological implications for students. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 172-181). Berkeley, CA: McCutchan.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), A guide to criterion-reference test construction (pp. 292-334). Baltimore: John Hopkins University Press.

Citron, C. H. (1982). Competency testing: Emerging principles. Educational measurement: Issues and practice, 1, 10-11.

Citron, C. H. (1983). Courts provide insight on content validity requirements. Educational measurement: Issues and practice, 2, 6-7.

Cloud, N. (1991). Educational assessment. In E. V. Hamayan & J. S. Damico (Eds.), Limiting the bias in the assessment of bilingual students (pp. 219-246). Austin, TX: Pro-ed.

Cohen, D. K. & Haney, W. (1980). Minimum competency testing and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 15-22). Berkeley, CA: McCutchan.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 201-220). New York: American Council on Education & Macmillan.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

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

Debra P. v. Turlington, 633 F. 2nd 397 (5th Cir. 1981).

Debra P. v. Turlington, Case 78-792, Memorandum Opinion and Order (M.D. Fla., May 4, 1983).

Debra P. v. Turlington, Case 83-3326 (11th Cir. C. App., April 27, 1984).

Donlon, T. F. (In press). Legal issues in the testing of Hispanics. In K. F. Geisinger (Ed.), The psychological testing of Hispanics. Washington, DC: American Psychological Association.

Duran, R. P. (1983). Hispanic's education and background: Predictors of college achievement. New York: College Entrance Examination Board.

Duran, R. P. (1988). Validity and language skills assessment: Non-English background students. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 105-128). Hillsdale, NJ: Lawrence Erlbaum Associates.

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 573-588). New York: American Council on Education & Macmillan.

Duran, R. P. (In press). Clinical assessment of instructional performance in cooperative learning. In K. F. Geisinger (Ed.), The psychological testing of Hispanics. Washington, DC: American Psychological Association.

Eckland, B. K. (1980). Sociodemographic implications of minimum competency testing. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 124-135). Berkeley, CA: McCutchan.

Finch , F. L. (Ed.) (1991). Educational performance assessment. Chicago: Riverside.

Gallagher, J. J. (1980). Setting educational standards for minimum competency: A case study. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 239-257). Berkeley, CA: McCutchan.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. Educational Measurement: Issues and Practice, 10 (2), 17-22.

Geisinger, K. F. (In press, a). Fairness and selected psychological issues in the testing of Hispanics. In K. F. Geisinger (Ed.), The psychological testing of Hispanics. Washington, DC: American Psychological Association.

Geisinger, K. F. (In press, 1992). The metamorphosis in test validation. Educational Psychologist. Also published in The Educational Researcher, 1991, 29(1), 2-12.

Glass, G. V (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Gordon, E. W. (1991). Human diversity and pluralism. Educational Psychologist, 26, 99-108.

Gorth, W. P. & Perkins, M. R. (1979). A study of minimum competency testing programs. Amherst, MA: National Evaluation Systems.

Hamayan, E. V. & Damico, J. S. (1991). Limiting the bias in the assessment of bilingual students. Austin, TX: Pro-ed.

Hambleton, R. K & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396). Berkeley, CA: McCutchan.

Haney, W. (1982). Validity and competency tests: The Debra P. Case, Conceptions of validity, and strategies for the future. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing (p. 63-94). Boston: Kluwer-Nijhoff Publishing.

Hardy, R. A. (1984). Measuring instructional validity: A report of an instructional validity study for the Alabama High School Graduation Examination. Journal of Educational Measurement, 21, 291-304.

Hartigan, J. A. & Wigdor, A. K. (Eds). (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 485-514). New York: American Council on Education & Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10 (2), 3-6, 10, 14.

Jaeger, R. M., & Title, C. K. (Eds.). (1980). Minimum competency achievement testing: Motives, models, measures, and consequences. Berkeley: McCutchan.

Linn, R. L. (1982). Curricular validity: Convincing the court that it was taught without precluding the possibility of measuring it. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. (pp. 115-132). Boston: Kluwer-Nijhoff Publishing.

Linn, R. L., Madaus, G. F., & Pedulla, J. J. (1982). Minimum competency testing: Cautions on the state of the art. American Journal of Education, 91, 1-35.

Livingston, S. A. & Ziecky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Madaus, G. F. (1982a). Competency testing: State and local level responsibilities. Educational Measurement: Issues and practice, 1, 10.

Madaus, G. F. (Ed.) (1982b). The courts, validity, and minimum competency testing. Boston: Kluwer-Nijhoff Publishing.

Madaus, G. F. (1982c). Minimum competency testing for certification: The evolution and evaluation of test validity. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing (pp. 21-62). Boston: Kluwer-Nijhoff Publishing.

McClung, M. S. (1978). Are competency testing programs fair? Legal? Phi Delta Kappan, 57, 397-400.

McClung, M. S. (1979). Competency testing programs: Legal and educational issues. Fordham Law Review, 47, 651-712.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13-103). New York: American Council on Education & Macmillan.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimum competence. Educational Measurement: Issues and Practice, 10(2), 7-10.

32

Oller, J. W., Jr. & Damico, J. S. (1991). Theoretical considerations in the assessment of LEP students. In E. V. Hamayan & J. S. Damico (Eds.), Limiting the bias in the assessment of bilingual students (pp. 77-110). Austin, TX: Pro-ed

Olmeda, E. L. (1979). Acculturation: A psychometric perspective. American Psychologist, 34, 1061-1070.

Olmeda, E. L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Padilla, A. M. (1979). Critical factors in the testing of Hispanic Americans: A review and some suggestions for the future. In R. W. Tyler & S. H. White (Eds.), Testing, teaching and learning: Report of a conference on research on testing (pp. 219-233). Washington, DC: National Institute of Education.

Paulson, D., & Ball, D. (1984). Back to basics: Minimum competency testing and its impact on minorities. Urban Education, 19, 5-15.

Reynolds, C. R. (1982a). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), The handbook of school psychology. New York: Wiley.

Reynolds, C. R. (1982b). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting bias (pp. 199-227). Baltimore, MD: Johns Hopkins University Press.

Roeber, E. D. (1990). Survey of large-scale assessment program: Fall, 1990. Lansing, MI: Michigan Educational Assessment Program, Michigan Department of Education.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1982). Validity as a variable: Can the same certification test be valid for all students? In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. (pp. 133-150). Boston: Kluwer-Nijhoff Publishing.

Serow, R. C. (1984). Effects of minimum competency testing for minority students: A review of expectations and outcomes. The Urban Review, 16, 67-75.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting bias (pp. 9-30). Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A. (1980). Technical issues in minimum competency testing. Review of Research in Education, 8, 30-82.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A guide to criterion-reference test construction (pp. 267-291). Baltimore: John Hopkins University Press.

Tittle, C. K. (1982). Competency testing. In H. E. Mitzel (Ed.), Encyclopedia of educational research (fifth edition) (pp. 333-352.). New York: Macmillan and the Free Press.

Trent, C. C., & Hinton, M. W. (1984). Mainstreaming, minimum competency testing and the minority student. The Western Journal of Black Studies, 8, 167-171.

Uniform guidelines on employee selection. (1978). Federal Register, 43(166), 38296-38309.

U. S. Department of Education. (1991). America 2000: An educational strategy. Washington, DC: Author.

Walker, D. F. What constitutes curricular validity in a high-school-leaving examination? In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. (pp. 169-180). Boston: Kluwer-Nijhoff Publishing.

Willig, A. C. & Ortiz, A. A. (1991). The nonbiased individualized educational program: Linking assessment to instruction. In E. V. Hamayan & J. S. Damico (Eds.), Limiting the bias in the assessment of bilingual students (pp. 281-303). Austin, TX: Pro-ed.

Other Relevant Readings

Anastasi, A. (1958). Differential Psychology (3rd ed.). New York: Macmillan.

Baecher, R. E. (1982). The instruction of Hispanic American Students: Exploring their educational cognitive styles. In J. A. Fishman & G. D. Keller (Eds.), Bilingual education for Hispanic students (pp. 368-383). New York: Teachers College Press.

Barton, J. I. & Robinette, C. L. (1980). Minimum competency testing of pupils: Psychological implications for teachers. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 155-171). Berkeley, CA: McCutchan.

Bernstein, D. K. (1989). Assessing children with limited English proficiency: Current perspectives, Topics in Language Disorders, 9(3), 15-20.

Broudy, H. S. (1980). Impact of minimum competency testing on curriculum. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 108-117). Berkeley, CA: McCutchan.

Calfee, R. (1982). Establishing instructional validity for minimum competency programs. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing (pp. 95-114). Boston: Kluwer-Nijhoff Publishing.

Chamberlain, P., & Medeiros-Landurand, P. (1991). Practical considerations for the assessment of LEP students with special needs. In E. V. Hamayan & J. S. Damico (Eds.), Limiting the bias in the assessment of bilingual students (pp. 111-156). Austin, TX: Pro-ed.

Easton, L. B. (1991). Developing educational performance tests for a statewide program. In F. L. Finch (Ed.), Educational performance assessment. Chicago: Riverside.

Feldmesser, R. A. (1980). Minimum competency as an individual right. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 425-437). Berkeley, CA: McCutchan.

First, J. M. & Cardenas, J. (1986). A minority view on testing. Educational Measurement: Issues and Practice. 5(1), 6-11.

Flaugher, R. L. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Hunter, J. E. Schmidt, F. L. & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds, & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 41-100). New York: Plenum.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Katzenmeyer, W. G. & Steener, A. J. (1977). Estimation of the invariance of factor structures across race and sex with implications for hypothesis testing. Educational and Psychological Measurement, 37, 111-119.

Kretschmer, R. E. (1991). Exceptionality and the Limited English Proficient student: Historical and practical contexts. In E. V. Hamayan & J. S. Damico (Eds.), Limiting the bias in the assessment of bilingual students (pp. 1-38). Austin, TX: Pro-ed.

Leinhardt, G. (1982). Overlap: Testing what is taught. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing (pp. 151-169). Boston: Kluwer-Nijhoff Publishing.

Linn, R. L. (1982). Two weak spots in the practice of criterion-referenced measurement. Educational measurement: Issues and practice, 1, 12-13,25.

Nuttall, E. V. (1987). Survey of current practices in the psychological assessment of Limited-English-Proficiency Handicapped Children, Journal of School Psychology, 25, 53-61.

Pennock-Roman, M. (1990). Test validity and language background: A study of Hispanic American Students at Six Universities. New York: College Entrance Examination Board.

Reynolds, C. R., & Brown, R. T. (1984). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 1-40). New York: Plenum.

Samuda, R. J. (1975). Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead & Company.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 32, 257-281.

Steward, A. A., & Bernazza, H. (1982). Cloze procedure with Spanish, English and Bilingual adults. In J. A. Fishman & G. D. Keller (Eds.), Bilingual education for Hispanic students (pp. 326-332). New York: Teachers College Press.

Thorndike, R. L. (1971). Concepts of culture fairness. Journal of Educational Measurement, 8, 63-70.

Response to Kurt Geisinger's Presentation

Michele R. Hewlett-Gomez
Sam Houston State University, Texas

Response: Last July, Former Labor Secretary William Brock stated in a Time Magazine interview, "We are the only country in the industrial world that says to one out of every four of its young people, "We are going to let you drop out of sight. We are not going to give you the tools to be productive" [p.12, 1990].

Competency testing has been one answer for school districts in preparing students for the real world to become productive citizens. Competency testing began with the back to basic movement with the intentions to assess minimum skills as deemed necessary by some for a student to function in the real world. Does mastery of identified competency skills guarantee a society that high school graduates will possess the necessary skills to be productive citizens in the real world? This question ponders policy makers, educators, and parents in our efforts to determine a meaningful education for students and especially for limited English proficient (LEP) students within our public schools.

Today's topic thus drives the question of relevancy on evaluating students' learning outcomes and on a student's individual merit of success. Dr. Kurt Geisinger, in his paper, clearly presents the current testing issues facing policy makers, educators, parents, and test publishers in our public schools and, in particular, the issue of linking competency testing to a high school diploma for limited English proficient students. Dr. Kurt Geisinger brings to the forefront the urgency to reevaluate the purpose of minimum competency testing for students and for limited English proficient students, in light of our public education goals.

Dr. Kurt Geisinger precisely identified seven subtopics related to minimum competency tests and limited English proficient students. These thoroughly researched subtopics included minimum competency and its current status, methodological issues of the tests, methodological issues of the tests and LEPs, testing standards, advantages/disadvantages, legal issues, usage of tests with LEPs, and finally solutions. In essence of time, five subtopics are highlighted with the intention to arouse and stimulate further discussion.

1. **Current Status of Minimum Competency Testing** Dr. Kurt Geisinger used Roeder's national survey (1990) on the current status of minimum competency testing to survey states testing standards and found that the majority of states use a version of

competency testing to assess academic skills. Approximately 37 of the 50 states, responding to Roeber's survey, stated either criterion referenced, norm-referenced or performance tests were administered in a variety of disciplines ranging from reading, writing, and mathematics to citizenship, health, and fine arts. Students were generally tested at alternating grades every two to three years beginning at the third grade. In addition, Valdez-Pierce (1991) surveyed the southern states to find Mississippi and North Carolina administered minimum competency tests, totalling 39 states.

With the state's functions on competency testing varying due to diversities of populations, policies and educational philosophies, states tended to place their priorities on functions such as general information [34], followed by system accountability [28], and then curriculum mastery [20]. Interestingly, these functions lead one to ask, "why would a state demand schools to test students for general information?," "how does accountability correlate to a high school diploma?," "what are states teaching students when the testing program is not aligned to curriculum?," or "how many limited English proficient students do pass the minimum competency test?"

One function, system accountability, seemed to be a way for states to guarantee to society that districts will be accountable for student learning outcomes. Accountability issues generally stem from state governments that use the data on student performance to reward, punish, and/or assist schools. Obviously, these states take accountability seriously and attempt to improve student performance on the indicators stressed by the state government. The effect has been in many schools to deemphasize instructional quality, narrow the curriculum, and emphasize mastery of the test objectives.

For example, since 1981, Texas legislation has mandated testing minimum skills in mathematics, reading, and writing; first with the Texas Assessment of Basic Skills and then with the Texas Educational Assessment of Minimal Skills. In October 1990, a new criterion-referenced testing program, Texas Assessment of Academic Skills [TAAS], was implemented to provide a shift from an assessment of minimum skills to an assessment of academic skills with the intention to assess higher level thinking skills and problem solving abilities. What the Texas state government had discovered about its testing program over the past ten years was that students had not achieved sufficient mastery of objectives in mathematics, reading, or writing that address higher level thinking skills and problem solving abilities. For example, in the Grade 11 writing composition test, which uses analytic and holistic scoring, students were able to compose and sequence thoughts, yet lacked the ability to support and/or elaborate their thoughts in a composition. In the Grade 11 reading test, students scored below 70 percent mastery on objectives relating

to summarization and points of view. In the Grade 11 mathematics test, students' mastery of some conceptual development, operational, and problem-solving skills were also consistently below 70 percent and in some cases below 60 percent such as in problem-solving using solution strategies. In effect, the testing program had shifted schools' standards for student success and accountability by rewarding districts for student success and punishing student with failures, alias no high school diploma. The ownership for student failure then had been on the student rather than the school system.

A second function, curriculum mastery, was used by states as a vehicle to align curriculum to an instructional and testing program. The relevancy of students' mastering specific skills became important, especially in the efforts to equal differential learning opportunities. Illinois and Minnesota were two states that developed their competency tests comprised of subject matter teams of teachers and curriculum experts. Texas, also, mandates a core-curriculum from which the districts design instructional programs to pass the state testing program. In addition, Roeber's survey identified Alberta, Canada, as using a minimum competency test, developed by teacher committees and scored by teachers, as partial criteria to receive a high school diploma. The test results in all three cases differ in their usage for program evaluation and instructional improvement. Should teachers be able to design their own testing programs to align with the curriculum and instructional program that addresses their student population? The issue of a decentralized authority to grant teachers control of curriculum for their student's learning merits priority consideration.

National. Besides the test functions, Dr. Kurt Geisinger addressed the current status of competency testing at the national level. The issue of a national test, American Achievement Test, is being called for by the Department of Education in the AMERICA 2000 report [U.S. Department of Education, 1991]. Since the purposes of this test are not clear, it then becomes even more critical that experts in the field of first and second language learning and testing present the relevant issues to policy making committees to define the uses of such unrealistic and misguided testing program. On April 23, 1991, Edward De Avila, an expert in linguistics and psychometrics, addressed the House Subcommittee on Select Education to state "That the development of a national test today was clearly to put the cart before the horse" (CTB News, 1991). His rationale was based upon the unfairness of administering a test to groups of children who had not received the same instruction, which would dictate local curriculum and, in particular, which would not necessarily tell us anything new about how students perform. Yet more importantly, De Avila focuses on the problem of definition. The lack of a consistent definition for limited English proficiency, compounds the tendency to pile all LEP students in one group regardless

of the varying language and academic skills. De Avila states "that without a clear understanding [definition], we have no way of deciding who should be in one program or another, or who would be eligible to take a national test." [p. 2, 1991].

Future Trends. Dr. Kurt Geisinger position magnifies the trends among states toward performance-based testing, focusing on higher level skills and problem solving. States, that know the benefits to holistic and analytic scoring, have begun to explore assessment alternatives such as performance-based and portfolio assessment as an answer to the limitations of minimum competency testing for all disciplines and students. For example, Kentucky has discontinued norm-referenced testing in favor of a performance test for 1996. Louisiana has expressed an interest in performance-based assessment. Colorado and Connecticut have placed greater emphasis on performance and applications of student learning. Illinois expressed gradual increase in testing mathematics and higher level thinking skills. South Carolina is headed toward testing higher level thinking skills. Massachusetts will use proficiency scales in reporting an increase in the use of open-ended questions and will use performance assessment as supplements to the current program. Minnesota will add performance testing in science and health. Missouri will field test performance assessment in science.

Though these trends may reflect tests for some states on student competency through performance-based assessment alternatives, it is not necessarily as a function of curriculum mastery nor as partial criteria for a high school diploma. Schools are not headed toward deleting the mastery of a discipline as a criteria for a high school diploma.

Certainly, Texas is no exception with its testing program pressing "Onward Through the Testing Fog" in at least three directions. First, in October 1992 in grades 4, 6, 8, 10, a new norm-referenced test, Texas Test of Basic Skills [TTBS], will be administered in reading, writing, mathematics, science and social studies. The TTBS is being developed by Riverside Publishing for implementation to all students with the decisions on exemptions for limited English proficient students pending. Secondly, the TAAS will add disciplines of science and social studies by October 1993 in grades 5, 7, and 9. By October 1994, test publishers will have added grades 3 and 11 to these disciplines. And thirdly, a reanalysis of the entire testing program is proposed in 1993 by the Texas legislators for a report to Governor Ann Richards with hopes for a realistic answer to measure student learning outcomes.

2. **Assessing LEPs with Minimum Competency Testing** To test, when to test, or not to test LEP students? Dr. Kurt Geisinger found that states had varied testing practices for lim-

ited English proficient students. Table 1 identified this variation in at least 13 states using minimum competency testing as a criteria to issue a high diploma [Roeber, 1990; Valdez-Pierce, 1991].

The policy decisions used to determine eligibility exemptions for limited English proficient students can provide answers to the dilemma of "to test, when to test, or not to test." Among these 13 states decisions to exempt or not exempt limited English proficient students from a minimum competency test to receive a "standard" high school diploma, three patterns evolved.

Table 1

Exemptions or No Exemptions for LEP Students By States Who Link Minimum Competency Tests and a High School Diploma

States With Mandates	Exemptions for LEPs	High School Diploma/ Passage of Test	Certificate of Completion
Florida	Yes [1 to 2 years]	Yes	Yes
Georgia	Yes [Parent/District]	Yes	No
Louisiana	No	Yes	No
Maryland	No	Yes	Yes
Michigan	Yes [2 years]	Yes	No
Mississippi	Yes	Partial Exemption	No
Nevada	No	Yes	Yes
North Carolina	Yes	Yes	No
Ohio	Yes	Yes	Yes
Oklahoma	Yes [Parent/District]	Yes	Yes
South Carolina	No	Yes	Yes
Tennessee	No [1 year/District]	Yes	No
Texas	Yes [1 test/District]	Yes	No

First, a pattern called "sink or swim," seemed to be used by Louisiana which provided no exemptions and no optional certificates. Second, a "good neighbor" pattern seemed to be used by Maryland, Nevada, South Carolina, and Tennessee to provide no exemptions and offer a certificate of completion to recognize student differences. One difference became apparent with Tennessee offering an exemption only to the students who attended school in the United States for less than one year and not to students with limited English proficiency.

Then, a third pattern, "half-way," seemed to acknowledge individual differences based on language and academic abilities and offer eligibility criteria for students who can take the test with a degree of success. Nonmastery of the test could either mean no high school diploma such as in the Texas, North Carolina, Mississippi, Michigan, and Georgia or a certificate of completion such as in Florida and

Ohio. Curriculum alignment to the testing program and remediation is unclear for each state with the exception of Texas which does mandate a state core-curriculum and testing program.

For example, Texas mandates decentralized decisions on student exemptions from the Texas Assessment of Academic Skills, by having districts form a Language Proficiency Assessment Committee comprised of teachers, administrators, and/or parents, to determine eligibility for the first test administration. Texas offers unlimited test retakes until age 21 and requires remediation courses prior to subsequent test administrations. Now with a 9 percent increase of limited English proficient students in 1990 to 314,674 of which 13,000 were in grades 11 and 12, with Hispanics having the highest dropout rate, 44 percent in 1989, and with an increase in student enrollment to 60,000 from which the minority language groups represented the majority [i.e., Hispanics (74 percent); Asians 67,735 (6 percent); Native American 6,275 (3 percent)], policy makers are challenged to design assessment and curriculum alternatives to the limitations of state mandates (TEA, 1991a).

Mississippi provides a two year waiver and time for test retakes for the state's Functional Literacy Examination or the Subject Area Testing Program. Exemption is offered on the Basic Skills Assessment Program and the Stanford Achievement Test. A LEP Assessment Committee, consisting of teachers, testing coordinator, counselors, psychometric personnel, and principals, determines documentation for exemption. Guidelines with definitions on the different levels of English language proficiency are utilized with such assessment alternatives as reading inventories, writing samples, course grades, teacher observations, and tests [i.e., teacher-made, achievement, and language proficiency] to determine test eligibility.

North Carolina provides guidelines to differentiate language proficiency levels for test eligibility with consultation from an assessment committee. An exemption is offered when a student's English language level hinders test mastery.

A common linkage in the "half-way" pattern between Texas, Mississippi, and North Carolina seemed to be the recognition to define language differences, as suggested by Ed De Avila, and offering test retakes with the inference that remediation and time would ensure the student opportunities to master the test. The weight and penalty of one criteria as a decision factor for these students' success and productivity as citizens is still questionable at best. Though for some states, these three patterns offer answers to linking minimum competency testing and a high school diploma. They certainly present a narrow vision for student success and are not without penalty to the student.

One possible answer not addressed by Dr. Kurt Geisinger to demonstrate student learning outcomes and without penalizing the student may be found in the "individual pattern." Here, the district takes ownership to achieve student learning outcomes by concentrating on alternative assessment. California's Option 1 Alternatives uses this pattern by decentralizing the assessment of student learning outcomes and offering to provide districts with an opportunity to design an outcome-based assessment process to demonstrate educational results rather than one test to "clinch a high school diploma" as the indicator of competency (California Department of Education, 1991). Option 1 Alternatives, though not tied directly to a high school diploma, provides six alternatives to design an individual evaluation program to measure educational results of a district's student population. Table 2 delineates the option alternatives.

Table 2
California State Department of Education: Option 1 Alternatives

Alternative	Description	Group Tested
A: Comparable Achievement-Norm-Referenced Test	Employs a norm-referenced test to demonstrate performance at or above national average.	Reporting eligible LEP or Former LEP.
B: Comparable Achievement-CAP	Employs CAP scores to demonstrate performance at or above the state average.	Reporting eligible LEP at grade levels tested on CAP.
C: Gap Reduction-Norm-Referenced Test	Uses a norm-referenced test to establish that the gap is lessening between scores of LEP students and all students nationwide.	Reporting eligible LEP students.
D: Gap Reduction-CAP	Uses CAP scores to establish that the gap is lessening between scores of LEP student and all students statewide. [Actually successive cohorts CAP method].	Reporting eligible LEP students at grade levels tested on CAP.
E: Successive Cohorts GAP Reduction-Norm-Referenced Test	Uses norm-referenced tests to demonstrate improvement in academic achievement scores of successive cohorts of LEP students.	Successive cohorts of reporting-eligible LEP students at specified grade levels.
F: Design and Implementation of an Alternative Evaluation Method to Demonstrate Educational Results	Allows a district to design an alternative use of standardized tests or other assessment methods to establish that it is effectively serving its LEP students.	Variable-to-be-defined by approved study design.

From these option alternatives, Alternative F seems to provide more flexibility in designing assessment alternatives for a district's student population by suggesting to [1] exercise caution in attempting to the design of an alternative methodology, [2] to use only well-supported academic achievement data to document claims of academic parity, [3] to carefully document the validity and reliability of the selected evaluation design and instruments, [4] to base district evaluations on the broadest range of student achievement, [5] to set outcome standards high enough to ensure that LEP students really are academically successful, [6] to select achievement tests that match the district's curriculum and have appropriate difficulty levels, [7] to explain the educational principles on which the instructional program offered to LEP students is based, and [8] to analyze collected data using procedures that are appropriate for the hypotheses that are being tested and the research questions that are being asked (CDE, 1991).

Thus, individual patterns direct districts to become owners of their evaluation program, align curriculum to evaluation program, differentiate instructional program to diverse learners, and become responsible for student success to determine who graduates with a high school diploma. The state government then holds the districts accountable through reporting requirements for preestablished "real world" outcomes for their student population.

Limitations to each policy decision pattern are evident. Yet continued decisions on assessment alternatives, definitions on language and academic proficiency, alignment of curriculum to an assessment program can guide districts to make competent, consistent, and relevant decisions on the academic performance of limited English proficient students.

3. **Advantage or Disadvantage.** To extend Dr. Gesinger's view, the biggest disadvantage is expectations, respectively by teachers and districts. When state government holds districts accountable for student learning outcomes with one single measure, then districts reconsider their priorities in terms of the state government's expectations on educational outcomes. Teacher expectations are a disadvantage for the reasons that teachers expect less, teach to the test, teach less creatively, and differentiate learning opportunities. First, teachers with students in lower tracks generally receive less rigorous and lower quality instruction. Second, teaching to a test and to minimal skills often fragments concepts instead of treating topics in depth and involves rote memory instead of critical thinking. District expectations, in turn, reflect assignments of the more experienced and effective teachers disproportionately to higher tracks rather than to work with the lower achievers or students needing remediation.

Minimum competency testing creates, by its nature alone, minimum expectations which in turn affect student learning outcomes. In Texas, state test results still indicate after 10 years of minimum competency testing that students are not learning well and it certainly is not to say they cannot learn.

4. **Legal Issues.** Even though the courts have upheld the rights of states to use competency testing appropriately, the limitations placed by testing programs used in states certainly needs further investigation. First, Title VI of the Civil Rights Act of 1964 merits a reminder since it prohibits discrimination based on race, color, or national origin in programs receiving federal financial assistance. States and districts should be cognizant that this law pertains to the issue of equity and language minority students. States, which use minimum competency as an answer to determine school success, need to reevaluate their state standards and criteria in respects to discrimination of one's national origin. The importance is magnified when the development of testing standards do not coincide to student assignments and language abilities and thus place limitations on a student's academic success.

Second, the issue of tracking students on the basis of test results is also relevant. The landmark 1967 case of *Hobson v. Hansen* deserves investigation in which plaintiffs successfully changed the testing and tracking system that had been instituted in Washington, D.C. In this case, students were placed in three instructional tracks at the elementary level [i.e., special academic (retarded), general (average), and honors (gifted)] and added at fourth track at the high school level [above average (college bound)]. Each track provided a different curriculum commensurate with the students' tested abilities. The findings indicated that African-American students were placed disproportionately in lower tracks compared with the district's more affluent white students. The proposed national test certainly has chances to become another vehicle to "track" language minority students' successes and failures.

Third, another case, *Lau v. Nichols*, presented underpinning issues for school districts to take affirmative steps to rectify English language deficiencies. The case in Ann Arbor, Michigan, in 1979, has implications in which a federal judge ruled that Ann Arbor school must recognize that students who speak Black English may need special help in learning standard English. Black English may constitute a language barrier. If barriers do exist in attempting to teach standard English, then students may be feeling inferior. Teachers requiring students to switch from Black English to standard English impedes the learning of standard English. Thus, teachers would need to be trained on language differences and the impact in assessing standard English. Obviously, further investigation on

the legalities of how testing impacts minority language groups' rights is warranted.

5. **Using Minimum Competency Tests with Limited English Proficient Students.** Dr. Kurt Geisinger's suggestions present limitations and merit discussion in the areas of remediation, LEPs for limited English proficient students, and curriculum alignment. First, I disagree that remediation after minimum competency testing to assist in mastery of the test objectives so students can receive a high school diploma. Remediation will assist a small percentage retakers at the high school level. The higher possibility exists that the retakers will not master the test objectives and dropout of school. For example, the preliminary Texas Assessment of Basic Skills test results in Grade 11 for April 1991 indicate 62,328 students were tested from which only 39 percent passed all three sections [mathematics, reading, and writing]; 4,860 students were retakers of which only 15 percent passed. Remember, retakers should have had the required remediation courses and counseling services. In addition, 12,628 students were retested with the TEAMS for a 57 percent mastery. TEAMS mastery does not indicate the number of test administrations by the retakers nor does it indicate the number of subtests taken.

In October 1990, 174,869 students were also given the TAAS in Grade 11 from which 65 percent mastered all three sections: Whites 75 percent, Asian 74 percent, American Indian 64 percent, Hispanic 52 percent and African American 45 percent. From the 13,659 identified limited English proficient students in Grades 11 and 12, 5724 were tested with only 18 percent passing, inclusive of retakers. Specifically, the discipline needing most remediation was mathematics with 74 percent mastery. Limited English proficient students, for obvious reasons, need remediation in all three disciplines with mastery as follows: reading 43 percent; writing 38 percent; and mathematics 40 percent. So when does one remediate?

Second, I disagree with the appropriation of monies and teachers to design specific individualized educational programs [IEPs] for limited English proficient students. These students' problems stem primarily from a lack of English language proficiency rather than a lack of academic knowledge. The isolated costs to districts would be monumental in the form of lower teacher-student ratio, evaluation experts and instruments, instructional materials, teacher training, and facilities to implement this solution. If districts were to assign IEPs for these students, then all students who have not mastered the test objectives should qualify. Instead of IEPs, the recommendations for alternative assessments to improve individualize instruction is suggested.

Third, I support the concept of curriculum alignment to testing. To change the process, districts must establish educational outcomes, think broadly, and align an assessment process with the curriculum and mastery standards. Remediation would then be integrated into the curriculum for each discipline as deemed necessary to master. Choosing the assessment alternatives would be dependent on the curriculum design and student population.

Two additional suggestions to Dr. Kurt Geisinger's position would include [1] to integrate teacher and policy decision training and [2] to develop an educational outcomes process. The first suggestion indicates the need for preservice and in-service training to successfully implement the instructional programs for limited English proficient students. Institutions of higher education, public schools, and state governments should actively collaborate to facilitate an understanding of assessment alternatives for limited English proficient students as presented by California. Today, teachers and policy decision makers [i.e., school board members, administrators, and counselors] have limited knowledge on how to assess students' academic progress much less their language proficiency, or how to interpret the test results to design and implement an aligned curriculum with an appropriate instructional program. Proper training for administrators and teachers is needed for policy to align itself with the instructional needs of limited English proficient students.

The second suggestion is the development of an educational outcomes process for limited English proficient students to include the definition of eligibility, alignment of curriculum and evaluation, integration of teacher training, integration of policy-decision training, identification of assessment alternatives, selection of program alternatives, and evaluation of program outcomes. The eligibility criteria for assessing educational outcomes would be the major first step to determine who should take or not take a competency test. This criteria could include a definition of language proficiency [i.e., listening and speaking] in English and the student's first language, a definition of language proficiency [i.e., reading and writing] in English and the student's first language, definition of academic proficiency [i.e., mathematics, science, social studies, health, fine arts, citizenship] in English and the student's first language, abilities on learning strategies [i.e., cognitive, metacognitive, affective, social], and the length of time in the school system.

Summary

The solutions suggested merit further discussion. Consideration of an educational process to access student learning outcomes is an alternative to the linkage of minimum competency testing for a high school diploma. Distribution of the ownership for student success will then be on the school system Therefore, if educators were to de-

velop a consistent and equitable process to determine a limited English proficient student's educational outcomes, then that student will be guaranteed a quality education to be a successful and productive citizen in our society.

Note

[1] 269 F. Supp. 40 (D.D.C. 1967).

References

Bolte. G. (1990). Interview: Will Americans Work for $5 a Day? Time. 7- 23. (pp. 12-13).

California Department of Education. (1991). Option 1 alternatives: Technical standards and recommended practices for development of outcome based assessment of district services to limited English proficient students. Sacramento, CA: Author.

CTB Macmillan/McGraw-Hill. (1991). Defining limited English proficiency makes national test premature, expert tells subcommittee. CTB News Bulletin. (pp. 1-2).

Roeber, E.D. (1990). Survey of large-scale assessment programs: Fall, 1990. Lansing, MI: Michigan Department of Education, Michigan Educational Assessment Program.

Texas Education Agency. (1991). SnapShots '90: 1989-90 School District Profiles. Austin, TX: Author.

U.S. Department of Education. (1991). AMERICA 2000: An educational strategy. Washington, DC: Author.

Valdez-Pierce, L. (1991, March 26). Assessment issues concerning non-English language background and LEP students. Paper presented at the International Teachers of English to Speakers of Other Languages Conference, New York.

Reponse to Kurt Geisinger's Presentation

Lawrence M. Rudner
ERIC Clearinghouse on Tests,
Measurement and Evaluation

Minimum competency and graduation tests should be welcome by all. In theory, students lacking basic skills are identified and social promotions are ended -- thereby providing better quality and more appropriate educational opportunity. The flip side, passing the exam, should also be welcome. Passing serves as a certification that the examinee has achieved some level of competence and is ready for the next level of instruction or is worthy of being called a high school graduate. Professional standards and legal precedents hopefully assure that the tests are appropriate and that people are being classified fairly.

Unfortunately legal precedents and professional standards are not always followed when tests are used with Limited English Proficient students. The Children's English Services Study, for example, used a pathetically small sample to determine the score that was used to define students as Limited English Proficient (LEP). Most recently the U.S. Department of Education authorized the use of several inappropriate tests in order to have "something" for LEP students. The Secondary Level English Proficiency test, designed and validated to indicate whether a student has enough English skills to be mainstreamed into an English speaking classroom, for example, was approved as an admission tests for post secondary education. The Spanish version of the P.A.R. Ability to Benefit Test, based on the old Adult Proficiency Level (APL) examination, was also approved -- even though there were absolutely no statistics or documentation available for that version. It was merely a translation.

In his excellent review of minimum competency testing as it pertains to students with Limited English Proficiency, Geisinger (1991) describes the status of minimum competency testing, the methodological issues of MCT and relates those issues to the assessment of LEP students. Throughout his paper, Geisinger describes what I view most competent measurement specialists would advocate given the presented issue. He points to professional standards and does an excellent job of describing how they apply.

If followed uniformly, the guidelines and suggestions outlined by Geisinger would help assure fair and equitable testing. Indeed, Geisinger has worked as an expert witness testifying against companies that have not adhered to professional standards. The quality of

assessing LEP students would be vastly improved if all tests publishers and users followed Geisinger's recommendations.

Throughout my comments, I will be reiterating many of the goals espoused by Geisinger. We don't disagree on the goals and we don't disagree too much on what we view as the responsibilities of the profession. We might even agree on the theme of this editorial -- that commonly accepted practices don't go far enough to assure fair and equitable test usage.

The better test developers can point to numerous activities that they typically undertake to make tests fair and appropriate for all students, e.g. review panels, representation in norming groups, bias analysis. I will argue that while this current state of practice has positive effects, it does not sufficiently protect LEP students from being inappropriately labeled and classified.

I start by identifying a few key points made by Geisinger that I feel warrant further emphasis. I then discuss the concept of validity. With perfect validity, many of the MCT issues raised by Geisinger become moot. The issues are issues, however, because MCT tests, like all tests, are not perfectly valid. I discuss some of the steps described by Geisinger to resolve those issues, discussing why I don't feel they are good enough.

Points Warranting Reiteration

- We have a set of carefully drafted standards -- APA, Uniform Guidelines, the GRE guidelines -- that should be followed.

  These are statement by the profession outlining steps to assure quality assessment and meaningful documentation. The standards are rigorous -- even major test publishers typically fail to adhere to major standards.

- Test scores should not be used in insolation.

  On key advantage and hope for the current move toward "authentic assessment" is that multiple observations are used. A single test score is easily influenced by sampling error as well as individual variation. Multiple measurement has the potential to reduce this type of error. It is only a potential and not a given because authentic assessment has yet to clearly identify the universe of skills to which scores are supposed to generalize.

- Need to know basis information about policies concerning the testing to LEP students.

There has been very little systematic or in-depth analysis of testing policy and practices, let alone practices as they apply to LEP students. Such analysis is needed to initiate discussion. (The problem is not as bad as painted by Geisinger in the first draft of his paper. Finding no relevant articles, Geisinger's graduate student was evidently not proficient in searching the ERIC database. Our search yielded 30 article on Minimum Competency Testing and Limited English Speaking students and over 180 articles on testing Limited English Speaking students.)

- Testing should be educationally relevant.

  Diagnostic testing to help teachers identify weaknesses is much more useful than summative testing. One encouraging aspect of the current interest in testing and this conference is that educators, rather than statisticians, are taking control of testing activities. I fear, however, that the educators are being co-opted by the politicians.

- The effects (and use) of testing should play a major role in tests validation studies.

  While self-evident to some, this is considered a radical idea in the measurement community. If tests are to be used to promote the common good, then the social consequences of tests must be examined.

Validity

Geisinger cites the literature to identify several relevant validity concepts:

(1) we do not validate tests, rather, we validate the accuracy of inferences that we make from test scores (Cronbach, 1971)

(2) the degree of empirical relationship between test scores and criterion scores (Meassick, 1989)

(3) the extent to which a test may be said to measure a theoretical construct or trait (Anatasia, 1988)

(4) relevance of the content to the content of a particular behavioral domain of interest and about the representatives with which item or task content covers that domain (Messick, 1989)

(5) a measure of how well tests items represent the objectives of the curriculum (McClung, 1979).

The inference we want to make in a minimum competency test is whether the student has mastered some set of skills. The questions then are:

- do we have the right set of skills? and
- have we measured those skills adequately?

Let us assume for a moment affirmative answers to those two questions. If the skills are defined as those needed for success at the next higher level, then validity can be demonstrated empirically (definition 2). On a perfectly valid test, we would expect each tested skills to be a prerequisite for some higher level skill.

On the other hand, the skills on the tests may represent a theoretical construct -- for example, the skills a minimally competent high school graduate should have mastered (definitions 3, 4, and 5). On a perfectly valid test of this type, skills which should be mastered by the minimally competent appear on the test; skills that are not necessarily mastered are not on the test.

This right set of skills could be enormous. All the skills up to the minimal level should be included. If an individual fails to master a skill in the set, then that individual is not minimally competent. Many of the skills may appear to be trivial. Even if no minimally incompetent individual fails to master it, it belongs in the universe of skills mastered by the minimally competent.

If we have the right set of skills and all the skills are properly measured then English language skills don't matter. Either a student demonstrates minimum competency or he doesn't. Either he is ready for the next grade or not. Or, for a graduation examination, either he meets the definition or not.

Should the set "right skills" differ across population groups? Clearly not for a graduation exam; second class, standards are not equitable. Hopefully yes for a promotion exam. Hopefully our special programs for LEP students make a difference and have different prerequisite skills. As Geisinger points out, the curriculum for LEP students needs to be carefully examined. A well articulated instructional and testing program can greatly aid education. If it is poorly articulated, or if the relationship has not been examined, then the tested skills lack relevance, i.e., are not valid.

Standards and Adverse Impact

Close to the issue of "right skills" is the issue of standards. Tests typically have a passing score -- above which you are said to be com-

petent, below which you are not. The need for passing scores is an admission that we may not have "right" skills. If you need to be minimally competent to pass an item, then 100 percent of the minimally competent would get the item right and the passing score would be 100 percent. Herein lies problem number 1, we are not very adept at defining domains. We include skills that minimally competent people get wrong. We admit as much when applying a standard setting technique such as the one attributed to Angoff and ask "What proportion of minimally competent people will get this item right?" Tests are not perfectly valid, and we don't have any ironclad standards.

It would be nice if we had a test which measured the right skills and had an incontrovertible standard above which everyone is competent and below which everyone is incompetent. Adverse impact would not be an issue.

Adverse impact occurs when members of one group are underepresented by the selection rule. With top-down selection, for example, individuals are selected based on their ranking, starting with the highest score and working downward until all available slots are filled. If the group means are different, then the members of the group with the higher mean will be selected first and will occupy most of the available slots. The de facto standard is well above the minimal competency level. While capable, members of the lower scoring group are systematically denied access. There are numerous court cases, *Griggs v Duke Power* being the most famous, where employers intentionally discriminated under the guise of an objective test. (While Geisinger did an excellent job of describing the *Debra P* case, which was an MCT case, he did not describe an entire body of legal precedents which I know he knows well.)

Tests cannot be used to exclude systematically if everyone is simply rated as competent or not. Those that are doing the selecting must choose from a pool of qualified applicants. If people are randomly selected, that is, given equal opportunity, then the proportion of group members that are selected would be the same as the proportion of group members that are qualified. Of course, MCTs are rarely used just to make a dichotomous classification. Scores are used and individuals are ranked. We don't have pure MCTs.

Standards and Adjustments

Geisinger provides a list of nine pieces of information that may be used to adjust the standards on the kinds of MCTs that are usually developed. In my own research on standards for teacher licensure examinations (a form of MCT), I found that the standards were lowered to ridiculous levels -- to the point that the tests were meaningless.

The usual argument is to adjust for false negatives -- failing people that are really competent. While false negatives may scream louder, they are rarely more serious of a consequences than are false positive -- certifying someone who is not really competent. Indeed, I will argue for an upward adjustment: I would rather leave someone back than to incorrectly promote them. Neither should be given preference.

The false negative argument is closely allied with the adverse impact argument -- "we lowered the standards so more LEP students would pass." While such an action may be politically astute, it is not educationally sensitive. Promoting people that are not qualified obviates the entire purpose of the testing program.

Geisinger argues for adjustments in the name of measurement error due to unreliability. He advocates lowering standards if the reliability is lower and the standard error of measurement is higher for LEP students. If the reliability is that different, then perhaps the test should not be used. Downward adjustments are not justified as measurement error can be in either direction. Finally, one would expect different reliability estimates simply due to variance differences. The act of making adjustments, however, is an admission that there is either a problem with the test or with the standards as they stand. They are not necessarily "minimum" standards.

Using Group Data

To help assure that we have measured skills adequately, the better test developers make sure groups are adequately represented in the item tryout and norming studies. Group data such as this, however, can easily mask real differences. Suppose we have a norming group for a mathematics test made up of 80 percent English skills, and 20 percent LEP students, a 5-option multiple-choice item; the item p-value is .60 for native speakers; the item p-value is .40 for LEP students with adequate English skills, and the item's requisite English skill is a problem for 25 percent of the LEP students.

The fact that the English load is a problem for 25 percent of the LEP students should raise some flags. When LEP students get the item wrong, we don't know if it is because they legitimately don't have the math skill or if their lack of English caused the problem.

If there were no English load, the p-value for this item from the norming study would be .56 (8*.6 + .2*.4). The inappropriate English load lowers the p-value to.55 (the LEP contributions is .75*.2*.4 for students with adequate skills plus .25*.2*.2 for students with inadequate skills since they can guess). The inclusion of LEP students would have no appreciable effect on the norms or item statistics.

We would expect a bias analysis to flag an item that presents an inappropriate problem for 25 percent of a population. Using the above example, the LEP p-value for LEP students is .35 (.75*.4 + .25*.2). We would compare this to the expected value, .40, and conclude that there is no bias. The problem stems from using a heterogeneous group, LEP, to look for problems that occur with only some members of the group.

Recommendations

Recognizing that we have to use tests and standards that are not infallible, I would rather see the same standard for everyone and measures of the goodness-of-fit (individual assessment accuracy, person-fit) calculated. A goodness-of-fit could simply be the correlation between an individual's response pattern and the item difficulties. We expect people to get the easy items right and the hard items wrong. If an individual's response pattern, regardless of English skill or race, doesn't make sense then the test data should not be used. Testing problems should be identified at the individual, not the group, level.

LEP students need to be included in norming studies; bias analysis needs to be conducted; standard setting studies need to be conducted. These steps outlined by Geisinger will improve norms, identify many flagrantly bad items, and help establish meaningful standards. Following these steps will clearly improve the quality if not the credibility of a testing program. If we are interested in developing assessments that are truly applicable to all children, LEP and non-LEP, then we need to do a better job of identifying the skills that we want to assess and a better job at identifying which students were properly assessed and which ones were not.