

**STANDARDS, POLICIES AND PROCEDURES FOR THE EVALUATION
OF ASSESSMENT INSTRUMENTS USED
IN THE CALIFORNIA COMMUNITY COLLEGES
(4TH EDITION, REVISED MARCH 2001)**

PREFACE

Matriculation in the community colleges is a process that promotes and sustains students' efforts to achieve their educational goals. Matriculation affords the individual access to educational opportunities, and then takes steps to increase the likelihood of the individual's success. One aspect of matriculation is to provide assessments that are useful for assisting the student's selection of an educational program. That assessment instruments are fallible is axiomatic, and thus the goal is to select and use instruments that provide the most accurate and useful information. This document represents an up-dated version of the standards, policies and procedures that are to guide the choice and use of placement assessment instruments used in the California Community Colleges.

As the requirements and process for approving the use of assessments for placement into California Community College courses have evolved over the past several years, experience with this process and professional expectations regarding tests and testing have necessitated revisions in the standards, policies and procedures that are to govern California Community Colleges' use of assessment to assist course placement in the years to come. All colleges now have a placement assessment process implemented and operating, with most students following the advice given. The current revision of the **Standards** is based on the premise that these placement systems and the assessments being used as part of these systems should be examined and evaluated on a continual basis. Initial evidence on the validity of assessments was needed to support their use as placement devices. Ongoing evidence is needed to assure that the systems and assessments within the systems are maintaining their effectiveness and appropriateness.

Feedback from the field and a Chancellor's Office Task Force review of the current policies and requirements suggested a desire to streamline the process and lessen the burden on the local college. Attempts to incorporate this feedback into the revisions were made while still maintaining elements of the old Standards that were judged to be absolutes to maintain the integrity of the standards and the review process.

Little has been changed in Sections One, Two and Seven from the prior editions; mainly minor editorial changes to maintain consistency and bring the document up to date. As with previous revisions, the major changes occur in Section Three

that then carry over into current Sections Five and Six and include elaboration and clarification of material in previous editions of the **Standards** and addition of new information. Of particular importance is the addition of a section on Specific Criteria for Computer Testing (Section Four). All of Sections Three, Four, Five and Six should be reviewed in detail, but the following highlight the major changes in old material or the addition of new material are found.

- A) Clarification of the minimum number of colleges for which supportive validity publishers of second party must provide data tests;
- B) Both empirical and logical approaches addressing a test's freedom from bias, insensitivity and offensiveness are now required as one of the minimum pieces of acceptable evidence that must be satisfied before a second-party test can attain initial status in any of the approval categories;
- C) The burden on the college to collect either criterion-related or consequential related validity evidence has been lessened. Evidence of this type is only required if an empirical design is chosen as the procedure for establishing or validating cut-scores;
- D) Appropriate evidence documenting the reliability of test scores has been expanded for local colleges who are managing second-party tests, developing their own tests or involved in a critical mass submission. The colleges now have an option to provide either test-retest or internal consistency estimates of reliability;
- E) A minimum sample size of 50 cases has been set as the required number in studies to determine the reliability estimates;
- F) An entire section on criteria for computer testing has been added;
- G) The expectation for conducting disproportionate impact studies is that such data are to be collected and evaluated at least once every three years; and
- H) A matrix summarizing the responsibilities of local colleges and test publishers in the approval process replaces the summary tables in the old edition. This matrix was developed by the Standards Review Task Force and is intended to facilitate communication of the requirements.

Personnel are available to assist understanding and implementation of the assessment Standards by contacting the Dean of Students at the Community College Office.

SECTION ONE: ASSESSMENT IN MATRICULATION

Regulations to implement the California legislative mandate known as "matriculation" (AB 3) define matriculation as

a process that brings a college and a student who enrolls for credit into an agreement for the purpose of realizing the student's educational objectives through the college's established programs, policies, and requirements.

The intent of AB 3 is to establish a matriculation system that describes minimum standards in California's community colleges and provides guidelines for implementing these standards to make certain that students pursuing post secondary education have equal access to programs and services and opportunities for success.

One major component of the matriculation process is assessment. AB 3 and its Title 5 regulations clearly indicate that the primary function of such assessment is to assist the student in making decisions about appropriate course level enrollment, major area of study and vocational program choice. Assessment's primary role in matriculation is viewed as providing descriptive and predictive information about students and their "fit" to courses and programs, thus facilitating their potential for success at the community college.

The intent of the legislation, where assessment is concerned, is to establish guidelines, procedures and standards for ensuring that assessment instruments and procedures implemented in the community college system are appropriate and in line with intended use as defined and described by AB 3 and Title 5 regulations. As such, the Act specifically requires that an advisory committee to the Chancellor of the California Community Colleges review and make evaluative recommendations concerning all assessment instruments used for placement by the colleges. Based on these recommendations, the Chancellor shall establish and update, at least annually, a list of approved assessment instruments and guidelines for their use by community college districts. In line with this requirement, the purpose of this document is to: (1) specify the procedures for the Chancellor's advisory committee to follow in arriving at these recommendations, (2) delineate the standards that the committee should consider in the review and evaluation of assessment instruments, and (3) define the information that shall be provided to the Chancellor by the advisory committee concerning the assessment instruments.

The recommendations that are formed are to be established based on those professional standards that guide educational and psychological testing and those conditions called for in AB 3 and Title 5 regulations. While the review will focus on specific instruments (tests), the final scrutiny will be on the suitability and appropriateness of the use, i.e. the test scores' interpretation(s) and

resulting recommendation(s) must be evaluated and ultimately judged. The final responsibility for the proper use of assessment instruments and procedures and resulting scores remains with local colleges. An affirmative recommendation by the Chancellor regarding a test only provides the opportunity for a district to consider its use. An affirmative recommendation does not automatically endorse the local college's use of the test score(s) as proper.

Keeping this stipulation in mind, a seven-step process is followed when making recommendations to the Chancellor about specific assessment instruments. These seven steps are detailed in a later section, but include the following:

- Step 1. Compile Information on Assessment Instruments**
- Step 2. Develop Psychometric Expert Review**
- Step 3. Develop Content Expert Review (for second-party tests at the time of first review)**
- Step 4. Develop MAC Assessment Work Group Review**
- Step 5. Generate Review Recommendations**
- Step 6. Disseminate Chancellor's Decision**
- Step 7. Allow for an Appeals Process**

For assessment instruments approved by the Chancellor, each community college district must develop documentation to demonstrate that it is using the test appropriately. The necessary documentation by the local college or district is discussed in later sections. Specific criteria and standards have been identified for test publishers and colleges to meet. This information will be reviewed during matriculation technical assistance visits by a team representing the Chancellor's Office.

When the Chancellor has not approved an assessment instrument, no community college district may use the test except on an experimental or pilot basis. The purpose for experimental use is to collect research information pertinent to a reconsideration of the instrument in an attempt to obtain future approval for its use from the Chancellor's Office. In these instances, local users would place emphasis in their documentation efforts on areas defined as deficient in the report filed at the time a non-approval decision was made concerning a test's use. A re-review of an instrument would involve all steps in the process.

The remainder of this document delineates the procedural steps defining the assessment instrument review process and identifies the reviewers' criteria (standards) used when examining an instrument. These standards, while defining the criteria by which judgments will be made about the recommended acceptability of an instrument's use, also provide guidelines and criteria for the local community college districts when selecting or developing an assessment

instrument, in implementing the instrument within a local assessment system, and in collecting local documentation.

Assessment in Matriculation

Where assessment by instrument is concerned, the matriculation Title 5 regulations place responsibility for the approval of assessment instruments with the Chancellor of the California Community Colleges. This responsibility has necessitated the development of review procedures and criteria for making the determination. Specifically, the regulations state that:

The Chancellor shall establish and update, at least annually, a list of approved assessment instruments and guidelines for their use by community college districts. These guidelines shall identify modifications of an assessment instrument or the procedures for its use which may be made in order to provide special accommodations required by Section 55522 without separate approval by the Chancellor. Such guidelines shall also describe the procedure by which districts may seek to have assessment instruments approved and added to the list. The Chancellor shall ensure that all assessment instruments included on the list minimize or eliminate cultural or linguistic bias, are normed on the appropriate populations, yield valid and reliable information, identify the learning needs of students, make efficient use of student and staff time, and are otherwise consistent with the educational and psychological testing standards of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (Section 55524)

Within the matriculation regulations, the broader implications of this directive are presented through the statements on what "assessment instruments, methods and procedures" encompass. In the regulations, assessment is defined as:

the process of gathering information about individual students to facilitate student success. Assessment may include, but is not limited to, information regarding the students' study skills, English language proficiency, computational skills, aptitudes, goals, learning skills, career aspirations, academic performance, and need for special services. Assessment involves the collection of such information at any time before or after enrollment, except that the process of assigning a grade by an instructor shall not be considered part of the assessment process. Once a grade has been assigned and recorded in a student's transcript it can be used in the assessment process. (Section 55502[b])

The intent of the definition is to be inclusive rather than exclusive in terms of what constitutes assessment in matriculation and therefore what must undergo review. The review focus is on the mechanism for gathering information, i.e., the instruments, methods and procedures that are employed. The mechanisms of assessment include but are not limited to:

interviews, standardized tests, holistic scoring processes, attitude surveys, vocational or career aptitude and interest inventories, high school or college transcripts, specialized certificates or licenses, educational histories and other measures of performance. The term "assessment instruments, methods or procedures" also includes assessment procedures such as the identification of test scores which measure particular skill levels, the administrative process by which students are referred for assessment, the manner in which assessment sessions are conducted, the manner in which assessment results are made available, and the length of time required before such results are available. (Section 55502[c])

Furthermore, the regulations provide guidance in that specific practices are prohibited. In implementing matriculation services, community college districts shall not do any of the following:

- (a) use an assessment instrument which has not been approved by the Chancellor pursuant to Section 55524, except that the Chancellor may permit limited field-testing, under specified conditions, of new or alternative assessment instruments, where such instruments are not used for placement and are evaluated only in order to determine whether they should be added to the list of approved instruments (Section 55521[a]);**
- (b) use any assessment instrument in a manner or for a purpose other than that for which it was developed or has been otherwise validated (Section 55521[a]);**
- (c) use any single assessment instrument, method or procedure, by itself, for placement, required referral to appropriate services, or subsequent evaluation of any student; provided however that, in the case of assessment instruments, the use of two or more highly correlated instruments does not satisfy the requirement for use of multiple measures (Section 55521[a]);**
- (d) use any assessment instrument, method or procedure to exclude any person from admission to a community college (Section 55521[a]);**

- (e) **use any assessment instrument, method or procedure for mandatory placement of a student in or exclusion from any particular course or educational program, except that districts may establish appropriate prerequisites pursuant to Sections 55002, 55201, 55202 and 58106 (Section 55521[a]); or**
- (f) **use any matriculation practice which has the purpose or effect of subjecting any person to unlawful discrimination prohibited by Chapter 5 (commencing with Section 59300) of Division 10 of this Part. (Section 55521[a]).**

These regulations provide the context for establishing the procedures and standards for review. The implication is that any information gathered about an individual student which is subsequently used in the matriculation process by the student or by others to make decisions about the student other than for the assignment of a grade falls under the definition of assessment and, thus, must be reviewed.

Very importantly, some prohibited assessment practices involve consequences resulting from the use of a test. Evidence that these negative consequences do not result from the use of the assessment mechanism in question must be provided. As a specific example, the Chancellor is charged with ensuring "... that all assessment instruments included on the list minimize or eliminate cultural or linguistic bias..." Disproportionate impact resulting from the use of assessment is the issue being addressed. From the regulations, disproportionate impact is defined to occur when:

the percentage of persons from a particular racial, ethnic, gender, age or disability group who are directed to a particular service or placement based on an assessment instrument, method or procedure is significantly different than the representation of that group in the population of persons being assessed and that discrepancy is not justified by empirical evidence demonstrating that the assessment instrument, method or procedure is a valid and reliable predictor of performance in the relevant educational setting.

The regulations require documentation and evidence addressing this issue:

- (a) **Each community college district shall establish a program of institutional research for ongoing evaluation of its matriculation process to ensure compliance with the requirements of this chapter. (Section 55512[a])**
- (b) **As part of the evaluation required under subsection (a), all assessment instruments, methods or procedures shall be evaluated to ensure that they minimize or eliminate cultural**

or linguistic bias and are being used in a valid manner. Based on this evaluation, districts shall determine whether any assessment instrument, method or procedure has a disproportionate impact on particular groups of students described in terms of ethnicity, gender, age or disability, as defined by the Chancellor. When there is a disproportionate impact on any such group of students, the district shall, in consultation with the Chancellor, develop and implement a plan setting forth the steps the district will take to correct the disproportionate impact. Community college districts shall also evaluate the impact of assessment policies on particular courses or programs. (Section 55512[a])

Evidence on disproportionate impact, therefore, must become part of the criteria used at some level when reviewing assessment instruments, methods and procedures. Information used to make decisions that affect individual students, the instruments, methods and procedures for gathering and the consequential use of the information must be reviewed except for individual course assessments used as part of the course instructional design or for the assignment of course grades.

Similarly, another regulation that needs to be incorporated into the review and subsequent judgments on the adequacy of an assessment process is the stipulation that multiple pieces of information must be used for placement, required referral to appropriate services or subsequent evaluation of any student. The required use of more than a single assessment instrument, method or procedure for making these decisions is found in Section 55521(a), under the list of prohibited practices identified previously. This regulation again identifies a situation in which an instrument might be reviewed favorably at one level of review, but because it is not to be used as the only source of input for placement decisions at the local level, the total assessment process (of which the information from a single instrument is but one part) will need to be reviewed.

The Assessment Review Focus

To focus this initial review, the use of assessment as a placement tool in California Community Colleges has been selected. Selecting for review those instruments, methods and procedures that serve the placement function is consistent with the spirit of AB 3, the Title 5 regulations, and sound educational practice. Assessment resulting in appropriate, i.e. valid, placement of students in courses and programs will serve the desired outcome of matriculation, which is to "facilitate student success in college." Use of assessment for placement purposes is also the dominant practice in California Community Colleges as indicated by surveys of assessment practices.

To select instruments in the review, two criteria were used to define the placement function in matriculation. Any assessment instrument, battery, or device used in one of the following manners was intended for inclusion:

- a. The instrument, battery or device is used to assist/help with the appropriate placement of students into different levels of instruction (e.g., reading, writing, mathematics), classes or programs.
- b. The instrument, battery, or device is used to advise students on course selection, career choice/path, or personal guidance/counseling.

In reviewing instruments that assist the placement function, the distinction between the potential for an instrument to validly serve its function and meet the mandate of AB 3 versus the actual use of information from the instrument and the consequences of its use at the local level needs to be noted. A requirement of AB 3 is that accurate (reliable and valid) information be used in forming the placement recommendations.

SECTION TWO: **STANDARDS FOR THE EVALUATION OF ASSESSMENT MEASURES**

The Identification of Standards

The Standards for Educational and Psychological Testing (1999) is a document specifying guidelines for the development and use of tests. The testing **Standards'** purpose is "to provide criteria for the evaluation of tests, testing practices, and the effects of test use" (Standards, p. 2). Prepared jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, the document represents a considerable effort on the part of many psychometric experts and test users to develop criteria for the development and evaluation of assessment instruments. The testing **Standards** is intended to provide a basis for evaluating the quality of testing practices as they affect the parties involved. Accordingly, the standards have served as the primary reference for the evaluation of measures in use by the California Community Colleges.

Since the philosophy underlying this document is similar to the one underlying the testing **Standards**, reviewers of tests for community colleges should become familiar with the information in the testing **Standards**. In general, the testing **Standards** state that the evaluation of tests should ultimately involve judgment and not some mechanical process such as a checklist. This judgment is to be based on a general knowledge of the behavioral sciences, a specific knowledge of the professional field to which the test applies, and alternative measures that are available for the same purpose. Also, knowledge of psychometrics along with a keen sense of practical issues such as feasibility of test use should play a role in the evaluation of any measure.

Introduction to the Standards for the Evaluation of Measures in Use by the California Community Colleges

The purpose of the testing **Standards** is broader than the purpose of this document: the establishment of criteria for the evaluation of assessment measures for the California Community Colleges. Accordingly, the following is an abbreviated version of the testing **Standards**. This abbreviated version reorganizes the testing **Standards** into five sections:

1. Validity;
2. Reliability and Errors of Measurement;
3. Scaling, Norming, Score Comparability and Equating;
4. Standards for Administration, Scoring and Interpretation; and,
5. Testing Special Groups.

Each section contains a number of criteria, with each criterion grouping together a number of similar principles from the testing **Standards** that have been paraphrased. These criteria have not been revised to conform to the wording in

the new edition of the testing **Standards** (1999). Rather, for continuity the older usage of terms is used. We believe the criteria and concepts presented are still appropriate and relate to the purpose of the present document. This abbreviated version of the testing **Standards** is presented for the convenience of the test reviewers for the California Community Colleges. The standards that have been abstracted are those that are expected to relate frequently to the community college reviewer's needs. However, test reviewers should be familiar with the testing **Standards** in its entirety and rely on its content as the primary source.

A related document that affords considerable guidance in the review of assessment devices is the Code of Fair Testing Practices in Education (1988) prepared by the Joint Committee on Testing Practices and distributed by the National Council on Measurement in Education. Unlike the testing **Standards**, which is designed to address the full range of testing applications, the Code is intended to be consistent with the testing **Standards**, yet differs in both audience and purpose. While based on the testing **Standards**, the Code has been prepared for the public at large. The Code narrows its scope to educational testing and focuses on those criteria that affect the proper use of tests. Given the Code's specific attention to educational testing, this document has been relied on as well to guide the identification of those criteria presented in this section.

A third primary source for detailing test standards judged particularly relevant to California Community Colleges was the Equal Employment Opportunity Commission's (EEOC) Uniform Guidelines for Employee Selection Procedures (1978). While prepared to provide guidance in matters of the use of tests for the selection of employees, the EEOC Guidelines do identify numerous criteria that can lead to the fair and equitable use of testing. As such, when specific criteria were evaluated for inclusion in this monograph, compliance with practices suggested in the Guidelines was monitored.

In abstracting and summarizing the testing **Standards**, the Code and the EEOC Guidelines, the criteria that are presented are compatible with these sources, but were identified specifically for evaluating assessment measures used by California Community Colleges. These criteria are offered for a number of related reasons. The testing **Standards** are intended to offer guidelines for test developers or test users who may be working with one of a number of types of tests used in one of a variety of settings. Therefore, the testing **Standards** are very broad and allow for multiple exceptions. In contrast, the criteria for the community colleges can be more specific. For example, since the California Community Colleges serve students from diverse populations, the criteria can identify as a requirement that test users must document how they plan to meet the special needs of these populations. Second, by having more specific criteria, the test reviewers should make more reliable judgments. Accordingly, the review system is more likely to be viewed as fair in that a test should be evaluated from the same basis regardless of the judges.

Finally, evidence for some criteria may likely not be available for review by the external test reviewers in their evaluations or are not restricted as required documentation by the test developer. However, local college personnel can expect that such evidence will be evaluated by the Chancellor's office during on-site visits to determine that such criteria are met. These criteria are specific to a particular college, such as criteria dealing with Cut Scores.

The sections that immediately follow address the criteria, standards and considerations necessary to judge tests in general as well as direct assessments of writing samples.

Select Criteria for the Evaluation of Measures in Use by the California Community Colleges

Validity

Validity is the most important concern when evaluating a test. Evidence of validity speaks to the suitability of the specific interpretation attached to or actions taken based on test scores. Seen in this light, the test itself is not validated. Rather, it is the use made of the test information that is to be validated. If use or practice differs across applications, then each specific application needs validation. Validity criteria that are especially germane to the California Community Colleges' use of tests follow. Unless otherwise noted, the test developer ordinarily has responsibility to provide the information called for by the criterion.

Criterion 1. Validity: General. Evidence should be provided by the test developer, test researchers, or users supporting the particular interpretation and use(s) of the test scores. Choice of types of evidence provided should be specified as well as a rationale for the mix of evidence. The specified purpose determines whether the reviewer should selectively weigh construct-, content-, or criterion-related evidence in the evaluative process.

Situations may occur in which, in a user's judgment, decisions may be based in part on tests for which little evidence of validity for the intended purpose is available. In these circumstances, the user should take great care not to imply that the test has established validity or place considerable reliance on the measure in decision making.

Necessary cautions concerning lack of evidence for the use of a measure, subscore, item, difference score, or profile interpretation are required within the testing materials.

If a user alters a test in some fashion, the user needs to revalidate the test under the changed condition or offer a rationale for why revalidation is unnecessary. Any substantial change in a test's format, alteration of how the test is administered, its language, instructions or content requires that additional

validation evidence be assembled. The extent of the revalidation will likely vary with the nature of the change(s).

Criterion 2. Content-Related Evidence A clear definition of the universe represented, its relevance to the test's use, and the procedures used to relate the item content to the universe should be described fully and accurately. Sufficient detail is expected so that users can evaluate the range of content in the measurement as appropriate for a domain. Even if the content-related evidence is fully described, a negative judgment might be rendered if the information does not support the choice of test content. If expert judges are used to make content evaluations, their qualifications must be specified.

Criterion 3. Construct-Related Evidence. If a test is proposed to measure a construct, the construct should be specified and well defined. In addition, evidence to support the inference from the test to the construct should be presented. Evidence is necessary showing that the instrument relates to what the test should measure, and that it does not relate to what it should not measure.

Criterion 4. Criterion-Related Evidence: General. All criterion-related studies should be completely described, including a specification of the sample, statistical analyses, criterion measures, and the time intended between predictor and criterion measurements. In addition, any factors that affect the results of the study should be reported. The criterion-related evidence should be judged to determine its support for and relevance to the intended use of the test. Further, criterion measures need to be described and the rationale for their selection provided. The generalization of the validity study to the intended use of the instrument must be described.

Criterion 5. Criterion-Related Evidence: Differential Prediction. Differential prediction should be investigated under two conditions: (1) when feasible, and (2) when prior research has established a substantial likelihood for differential prediction to occur with a particular type of test. That is, when groups differ in their demographics, past experiences, or instructional treatment and such factors are related to performance on the test, then investigations must be undertaken to determine if decisions are systematically different for the members of a given group than for all groups combined. Any differential prediction studies should involve proper statistical and methodological considerations.

Criterion 6. Cut Scores. In some applications, users make one decision if a test taker scores at or below one score (the cut score) and a different decision if the test taker scores above that value. The test user is required to document the method, rationale, and appropriateness for setting that cut score. Evidence of the appropriateness of the cut score for the decision must be documented. When Cut Scores are based on professional judgment, the qualifications of the judges should be documented.

Reliability and Errors of Measurement

Error in measurement is inevitable. Recognition of this fallibility serves as its own caution and requires that the likely extent of error associated with test scores be documented. While many (potential) sources of error exist, an essential requirement is that the approach to documenting the extent of measurement reliability takes into account errors of greatest concern for a particular test use and interpretation. Criteria associated with the reliability of tests most germane to California Community Colleges use are presented below.

Criterion 1. Estimates of Reliability and Standard Errors: General. For each reported score, estimates of reliability and standard errors of measurement should be provided to determine whether scores are sufficiently accurate for the intended use. The sample characteristics, statistics, and, more generally, the methodology employed to document a test score's reliability should be described completely. If theoretical or empirical reasons exist to suggest that estimates differ by population, estimates should be provided for each major population. Performance assessments that typically rely on human judgment must be monitored to insure adequate levels of scorer consistency.

Even if appropriate estimates are provided and the employed methodology is acceptable, the measure should be evaluated negatively if the estimates of reliability and standard error indicate low accuracy for the test's intended use.

Situations may occur in which, in a user's professional judgment, decisions should be based in part on tests for which little evidence of reliability is available. In these circumstances, the user should take great care not to imply that the test has established reliability.

Criterion 2. Estimates of Reliability and Standard Errors: Type of Estimates Provided. The type of reliability estimate should be carefully selected, appropriate for the expected use, and properly interpreted. For example, coefficients of internal consistency should not be interpreted as estimates of stability over time. Coefficients that yield spuriously high estimates of reliability for speeded tests should not be used if total performance is dependent on test takers' inability to complete a test due to time constraints.

When corrected coefficients are reported, uncorrected indices must be presented as well. If a test is scored using a judgmental process, the degree of error introduced by scoring should be documented.

Criterion 3. Estimates of Reliability and Standard Errors: Specific Applications. For tests whose use relies on Cut Scores, reliabilities need to be reported at the cut score or for score intervals. Also, decision consistency reliability information needs to be reported for selected score points. For adaptive tests, reliabilities must be reported for repeated administrations using different items.

Scaling, Norming, Score Comparability and Equating

The metric used to report test scores is chosen to enhance the interpretability of the information shared with the user or test taker. In the same vein, access to norms provides a means of referencing performance to a defined population of persons or groups. The utility of having multiple forms of a test available is realized only when performance can be reported on a single common scale. (Otherwise there will be an advantage or disadvantage associated with the particular form taken). Regardless of the utility achieved by establishing derived scales, resultant transformed scores can introduce error to the measurement process due to the procedure itself or the sampling methodology used. Important criteria for evaluating the appropriateness of test score transformations used in the community colleges are presented below. As these criteria suggest, benefits realized by transforming scores must be carefully weighed.

Criterion 1. Choice of Scales. The method used to compute the transformed (derived) scale or raw score should be clearly delineated. In addition, the rationale should address the relationship between the scaling methodology and the test's purpose. The measure is strengthened by the transformation of scores to the degree that the choice of scores is appropriate for and consistent with the intended purpose.

Criterion 2. Norms. The group of persons used for establishing the norm should be well described and appropriate for the test's intended use. The norm group must be a group to whom users will wish to compare the individuals tested. The methodology for constructing norms, including sampling plan, participation rates, and descriptive statistics, should be exhaustively specified. In addition, the year(s) that the data were collected should be reported. Out-of-date norms should be avoided.

Criterion 3. Comparability of Scores. When scores based on different test forms or different response formats are intended to be interchangeable, data that support the equivalence of the forms must be documented. If the content of the instrument changes across years, but the scale is intended to be comparable, the method for maintaining comparability should be described and be adequate for its intended use. When the test specifications change across years, the change should be fully described, the rationale for the change given, and the test user informed of the extent to which scores remain interchangeable. When the changes to a test are considerable, it becomes necessary to re-establish its psychometric characteristics including validity, reliability, etc.

Standards for Administration, Scoring and Interpretation

Information that documents the properties of a test must be available to users. In this respect, manuals and technical reports are particularly important for detailing how to appropriately use a particular measurement device. Completeness, accuracy, and clarity remain key to the selection and proper use

of a test. Criteria that speak to the adequacy of communication and the use of information for community college staff are presented in this section.

Criterion 1. Administration and Scoring. The standardized procedure(s) for the administration and scoring of a measure should be fully described in the test's manual. A test manual should identify the qualifications necessary to administer the test appropriately. The standards for test administration and scoring specified in the manual should match the characteristics of the test. Any modification of standard test administration procedures or scoring should be fully described in the manual with appropriate cautions noted. A college or its representative that develops a test has the same obligation to supply manuals and technical reports as does a commercial test publisher. Standardized scoring instructions and rubrics are essential when the measure is a performance assessment.

Criterion 2. Interpretation: General. A test manual should identify qualifications necessary to interpret test results. Tests should not be interpreted as ability tests without considering alternative explanations. Screening measures should be used only for identifying individuals for further evaluation. Test users should not use interpretations of test results unless they have documentation that indicates the validity of the interpretations for the intended use and on the samples on which they were based.

A test user that makes educational decisions based on differences in scores, such as aptitude and achievement scores, should take into account the overlap between the constructs and the reliability or standard error of the difference score.

Criterion 3. Interpretation: Test for Certification. If a test is used to certify completion of a given education level or grade level, both the test domain and the instructional domain at the given grade or education level should be described with sufficient detail so that the agreement between the test domain and test content can be assessed. Also, the test should not cover materials that a student has not had an opportunity to learn. Students should have multiple opportunities to take such a measure.

Criterion 4. Test Materials. The testing materials should be readable and understandable. Materials should limit claims of test properties and characteristics to those conditions for which data exist to support the claim. Such support data should be documented.

Criterion 5. Decision Making. A decision or characterization that will have a major impact on a test taker must not be made solely or automatically on the basis of a single score. Decisions are to be made in conjunction with other test information, previous classroom performance, and opinions by advisors from discussions with students, etc.

Testing Special Groups

Recognition of the difficulties associated with using tests appropriately with groups with linguistic differences merit careful evaluation and study if assessment results are to be valid. Also, providing accommodations for testing persons with disabilities will improve the appropriateness of the measurement. Criteria that enhance measurement and subsequent evaluation for diverse populations are presented below.

Criterion 1. Testing Special Groups: General. Test administrators and users should not attempt to evaluate test takers whose special characteristics, ages, disabilities, or linguistic, generational, or cultural backgrounds are outside their range of experience. A test user should seek consultation regarding test selection, necessary modifications of testing procedures, and score interpretation.

Criterion 2. Test Design for Non-native English Speakers. Tests, their items and their administration instructions should be designed to minimize threats to validity and reliability that may arise from language differences. Any recommended linguistic modification should be described in detail. When a test is to be used with linguistically diverse test takers, information should be provided for appropriate use and interpretation. Translated tests should be evaluated for reliability, validity, and comparability with the English version.

Criterion 3. English Language Proficiency Tests for Non-native English Speakers. English language proficiency should not be determined solely with a test that demands only a single linguistic skill. The caution here is not limited to the test that is dependent on a single format (for example, a multiple-choice, paper and pencil device). The attention of this standard is focused on the breadth and depth of the construct being appraised. Users need to be aware of the needs of students in an academic environment and demand a complete range of language skills, that is for written as well as reading, oral and listening proficiency.

Criterion 4. Test Design for People with Disabilities. Expertise, both psychometric and training or experience with populations with disabilities, is a prerequisite to any modification of a test for an individual or group with disabilities. Knowledge of the effects of various disabilities on test performance is essential. Until validity data are obtained for scores secured from non-standardized testing conditions, documentation must be available and the results must be interpreted cautiously. Pilot testing of the modified measure(s) is strongly advised with persons having the same or similar disability. When feasible, time limits should be modified for the person with disabilities based on previously conducted reliability and validity studies.

Criterion 5. User Responsibility. Knowledge of alternative measures is a precondition to test selection for use with linguistic minorities or persons with

disabilities. When modified forms are available, they are to be used with these persons. Proper selection of appropriate norms to facilitate score interpretation is essential. Using personnel for test administration who have been specifically trained for the group or person to be tested is strongly encouraged.

Criteria Associated with Direct Performance Assessment

The preceding discussion of validity, reliability, scaling and equating, administration, and testing special groups offers guidelines for the evaluation of tests in general. Additional criteria are useful in the evaluation of the direct performance assessment, e.g., writing or oral interviews. With these tests, individuals are asked to write responses to a question, prompt or task which are scored following some well-defined procedure. Additional criteria are presented for these tests as they are scored using subjective rather than objective methods.

Validity

Three aspects unique to direct performance assessment need special attention where the validity standards are concerned:

1. question, prompt or task development procedures,
2. scoring procedures, and
3. multiple questions, prompts or tasks or the repetitive use of any one question, prompt or task over time.

For 1 and 2, the key focus will be on evidence documenting the standardized implementation of procedures. The quality of evidence on these two procedures will directly affect judgments about the appropriateness of using the same or different (but assumed parallel) questions, prompts or tasks across students on the same assessment occasion or across occasions, e.g., semester to semester (#3 above). The degree to which evidence suggests that the question, prompt or task development and scoring guidelines result in a standardized (consistent and well-defined) implementation of procedures will be critical in the overall validity review. As such, the emphasis in the review of evidence on validity for direct assessment of performance procedures will shift toward construct validity interpretations. For example, in the direct assessment of writing, content validity evidence as defined and reflected in the development and the subsequent fit of scoring procedures to actual course content will take on greater importance since the procedures must provide assurances that a process is in place that will result in student scores reflecting the identical writing skill construct across writing sample stimulus prompts that become the focus of instruction.

Reliability and Errors of Measurement

Reliability considerations remain in force as for multiple-choice examinations, but the focus of the reliability evidence changes. For example, if scoring systems for a measurement instrument are subjective, such that different scorers have opportunity to evaluate the same performance, but arrive at different scores (e.g., direct measures of writing), evidence of interrater reliability should be provided. Similarly, if multiple questions, prompts or tasks are being used to place different students within the same assessment period, equivalent forms of reliability evidence would be expected. Additionally, when appropriate, procedures or evidence should be presented addressing intrarater reliability issues and concerns to insure that rater drift does not occur over the scoring period. While stability coefficients are appropriate, the inter/intra-rater and equivalent-forms coefficients are viewed as most important.

Scaling, Norming, Score Comparability and Equating

As many of the direct performance measures are expected to be locally developed and used, there will be little need to attend to normative

transformations. However, two particular points from the **Standards** in this section need increased emphasis. First, the method used to compute the transformed (derived) scale or raw score needs to be clearly delineated. The rationale should address the relationship between the scaling methodology and the test's purpose. Similarly, when analytic and primary trait scoring methods are used, the total score is a composite of component scores. A rationale for inclusion of components, their scaling and any weighting of component scores in comprising the total score index should be provided.

A second primary consideration in this section for direct performance assessment is the documentation of the equivalence of forms when scores based on different questions, prompts or tasks are intended to be interchangeable. If the content of the instrumentation or the scoring procedures or criteria change across occasions, but the scale is intended to be comparable, the method for maintaining comparability should be described and be adequate for its intended use.

Standards for Administration, Scoring and Interpretation

Those standards stressing the need for a clear description of the standardized procedures for the scoring of a measure need emphasis from this section, given the concern over the subjectivity of the scoring models. Of particular note should be an expected description of procedures implemented to train potential scorers. Test security issues relative to the unintentional, but potential dissemination of question, prompt or task wording or content is a concern to be addressed under this standard.

Testing Special Groups

Special care needs to be taken when considering the appropriateness of direct performance assessments for special student groups due to the performance-based response format required. For example, the literature provides some evidence that scorers are potentially influenced by the physical appearance (neatness, writing legibility) of a writing sample, an outcome characteristic that may put some individuals with disabilities at a disadvantage while having nothing to do with their writing ability. Alternative means of responding may need to be incorporated into the assessment process for some groups (e. g., the visually impaired) with attention to any resultant influence on scoring.

SECTION THREE: SPECIFIC CRITERIA FOR TEST USAGE FOR THE CALIFORNIA COMMUNITY COLLEGES

The preceding section summarizes and abstracts portions of the Standards for Educational and Psychological Testing (APA, 1999) that are most relevant to the evaluation of measures used by the California Community Colleges. The **Standards** were written to help individuals assess the quality of all types of instruments and to describe the obligations of all parties involved in the testing process, including test developers and users. Because it offers standards for diverse measures and all interested parties, the **Standards** are written very broadly and often lack specificity for a particular application. For example, although the **Standards** clearly emphasize that validity evidence is required for the use of a test, the **Standards** do not suggest what type of validity evidence is needed for a test with some particular use or the required strength of the evidence.

Most tests used by California's community colleges are intended to help students select appropriate courses. The tests serve a placement purpose, offering students guidance as to whether they should enroll in a course at the beginning of the sequence of courses in mathematics, for example, or somewhat later in the sequence. Because of the specific nature of these measures and their common use within California's community colleges, more explicit criteria can be written for both the test developer and the test user than what are or could be presented in the **Standards**. It should be noted that a test user may be a college, a district, or a consortium of colleges that choose to work together to establish the suitability of a test.

More explicit criteria for the test user and test producer, referred to below as specific criteria, are needed and are desirable for a number of reasons. First, the diverse audiences involved in the California Community Colleges should understand the evaluation system used in the review process. In this sense, the **Standards** by themselves are too vague for establishing an evaluation policy. More specific criteria will allow all parties involved to understand what is required for the Chancellor to approve the use of an instrument. Second, with more specific criteria, test developers and users will be better able to estimate whether their tests need further refinements and documentation before being submitted for review. Accordingly, fewer tests will be reviewed and the tests that are reviewed are more likely to receive a favorable evaluation. Third, the decisions that are reached in the review process should be more consistent from year to year. Without explicit specific criteria, the likelihood is higher that the same instruments judged acceptable one year might be judged unacceptable another year because the interpretation of the **Standards** would vary across years. Fourth, the evaluation system is likely to be less arbitrary with explicit criteria. Finally, defining specific test criteria at the state level indicates a clear commitment on the part of the Chancellor's Office that tests be used to aid

students in making important academic decisions rather than preventing them from reaching their educational goals.

The expectation is that all specific criteria will be met. However, a test developer or user occasionally might not fulfill one or more of the specific criteria, but could still justify the use of the test. A test developer or user may indicate in writing why a specific criterion has not been met. Whether an exception is granted will be considered on a case-by-case basis.

In all cases there are two criteria which the test's technical report(s) must address and present evidence in order to attain even the lowest level of approval status. In the absence of such information, the test will not be approved for use in California Community Colleges. **There must be evidence of the test's validity for the intended purpose, and there must be evidence that the test minimizes cultural/linguistic bias, insensitivity and offensiveness.**

Assessment instruments presented for approval by second party publishers (Section I or III below) must provide acceptable criterion-related validity evidence; colleges locally managing an instrument (Section III) can meet the minimum validity condition by presenting acceptable content-, criterion- or consequential-related evidence. Second party publishers must at a minimum provide acceptable documentation addressing the freedom from bias, insensitivity and offensiveness based on both empirical and logical analysis findings (I.A.1.c or III.A.1.c below). It is noted, however, that meeting the minimum requirements is not sufficient to attain Full Approval status.

Specific criteria are presented below for three different types of tests:

- I. Objectively scored measures developed and maintained by a second party external to the California Community Colleges,
- II. Objectively scored measures developed or managed by a California Community College or district, and
- III. Performance Assessments.

The primary distinction between the two types of objectively scored tests is that in the second instance a California Community College or group of colleges has assumed the responsibilities of both the developer and the user. More specifically, objectively scored measures developed or managed by a California Community College fall into two categories:

1. those developed and used by a California Community College or district, or
2. those developed by an independent vendor and not approved by the Chancellor but whose use is deemed appropriate by a college or district, and that institution assumes responsibility for bringing the test into compliance with the **Standards** as a locally managed or controlled test.

Appendix A in this document provides a copy of the request form that a college must complete and include whenever it submits documentation to gain approval for the use of an assessment instrument.

The third type of tests, performance assessments, are those that require students to create a response or product as a demonstration of their knowledge or skill. These assessments often are referred to as a direct measure of knowledges or skills and require the active participation of the student in constructing a response to a question, prompt or task requirement. The student's response (behavior or product) is then typically scored following some well-defined procedure. Assessments requiring the production of a writing sample by students or oral responses to a set of interview questions are examples of performance assessments. The scores resulting from the evaluation of the quality of these writing samples or oral responses might then be used to advise students in making course placement decisions. For direct performance assessment measures, a different set of specific criteria are used in that these measures are scored using subjective rather than objective methods.

In general, the appropriate use of a test needs to be considered in light of all applicable standards described in the Standards for Educational and Psychological Tests (1999), the Code of Fair Testing Practices (NCME, 1988), and the Guidelines on Employee Selection Procedures (EEOC, 1978). Consequently, meeting the specific criteria described below is a necessary, but not sufficient, condition to receiving a favorable recommendation for a test's use. It should be noted that evaluating the appropriateness, suitability and usefulness of a particular test is an ongoing, continuous activity. Accordingly, as a test undergoes changes over time, it will be necessary for a test developer to re-evaluate it and to submit the results of studies for review and a recommendation by the Chancellor.

Finally, colleges have a responsibility for documenting and maintaining evidence of the appropriateness of those assessments they employ. As such, colleges are advised to prepare a "validity and fairness portfolio" for each sequence of courses in which results from assessment testing are used in generating student placement recommendations. The portfolio would constitute a body of evidence addressing the accuracy, utility and fairness of the assessment process, assessments and practices at the local college. The portfolio would include information on score validity and cut score validation, test bias, reliability, standard error of measurement, disproportionate impact and special testing accommodations.

The portfolio may include both fundamental and supplemental evidence of validity. It is educationally sound to base evidence of validity on a variety of sources of information. Beyond the evidence called for in this **Standards** guide, other relevant information may include student and faculty surveys reflecting on

satisfaction with placement recommendations, anecdotal information addressing assessment utility, independent investigation carried out by local users or others, etc. Refer to the Local Research Options-Assessment Validation Project, February 1991, for research designs in collecting supplemental data. Additionally, a monograph has been prepared that depicts exemplary research already completed by California Community Colleges (November 1993); this document is available through the Chancellor's Office.

A narrative should be included within the portfolio which summarizes the body of evidence supporting the use of the assessment instrument for placement into the sequence of courses. If findings are included which reflect negatively or questionably on the validity, utility and fairness of the instrument for any of the courses in the sequence, an explanation should be incorporated that clarifies why these outcomes are not of consequence for those courses.

The remainder of this Section presents standards for test developers, colleges and test users. Please note when considering the following explicit specifications of assessment standards, methods and criteria that validation should be viewed as an ongoing process. A college or publisher should continue to assemble information that may lead to changes and improvements in a test or will provide assurance that a test and its resultant scores continue to be suitable for California Community College students and the decisions based on these scores. To support this belief and expectation, the prior version (1998) of these standards introduced the concept of *Consequential-Related Validity* evidence which by its very nature needs to be ongoing and is generated after the initial evidence is provided for approval of an instrument. Further, the ongoing data collection and documentation efforts associated with a test can be used to explore and extend evidence addressing multiple measurements associated with the test instrument and student performance. In the case of multiple measures, the goal over time for a college would be to identify companion measures that support an existing test or a separate assessment that provides for a more complete or comprehensive assessment of the criterion behavior.

A Summary Matrix of the specific requirements detailed in the following sections is provided on pages 43 to 47. This matrix summarizes the requirements and expectations for each standard that publishers and colleges need to meet for each level of approval, i.e., to attain probationary, provisional or full approval status.

I. Specific Criteria for Objectively Scored Measures Developed and Maintained by a Second Party External to the California Community Colleges

A. Primary Responsibility of the Test Developer

1. Validity and Fairness

- a. Content-Related Validity Evidence. The test developer is to describe the content of the test's items with sufficient and clear specificity. Explicit statements of test objectives and tables of specification need to be available to inform college users. Based on these sources of content description, users at community colleges are able to consider the test's appropriateness for making placement recommendations for a sequence of courses at their colleges.

So that local colleges can conduct required content-related validity, test publishers must provide test booklets or a sufficient representative sample of test items (in the case of computerized tests) such that local colleges may conduct an item-by-item review. If a sample of items is provided, the number should be such that they represent what is a psychometrically sound single form of a traditional paper-and-pencil administered form of the test.

- b. Criterion-Related or Consequential-Related Validity Evidence. Data must be presented to indicate that the test is useful for making placement decisions for California Community College student populations and courses in these colleges. Empirical evidence should support the following conclusion: test takers who achieve scores within some specified range should take a different course or set of courses in comparison with test takers who score outside that range. Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

When submitting evidence to meet this standard, the following design criteria should be met: the course content should bear a close logical relationship with courses offered by the California Community Colleges; the students should be similar to the students enrolled in the California Community Colleges; when used as the primary index, the correlation between the test and a student's readiness to assimilate course content or performance (e.g., mid-term grade, student or instructor evaluation, end-of-course grade) should be substantial; and, across all data sets presented by a test publisher, the average correlation should be greater than or equal to .35 (or a comparable value if an alternative statistical analysis was performed). The magnitude of the correlation may vary as a function of the degree to which a test was used to place students in the course under investigation and/or the variation in grading standards across classrooms. Coefficients corrected for restriction of range are acceptable.

As a guide, supportive data from at least six community colleges are required to attain full approval status; supportive data from at least four community colleges are required to attain provisional approval status; and supportive data from at least three community colleges are required to attain probationary approval status. Additionally, a majority of the colleges included must be California Community Colleges representing the diversity of courses and students in the California Community College system, i.e., 4 of 6, 3 of 4 or 2 of 3, to attain a specific approval level status.

- c. Evidence Addressing Test Bias. Evidence focusing on cultural/linguistic bias, insensitivity and offensiveness must be provided. This work should be done on student groups that are similar to those ordinarily served by the California Community Colleges. This evidence needs to be of two types: evaluations of test items by culturally and linguistically diverse panels and results from test data that address bias. The findings from these investigations should be used to eliminate or minimize test bias, insensitivity, and offensiveness.

2. Reliability and Errors of Measurement

- a. Reliability. The stability of a placement test should be evaluated. The stability may be assessed by either administering the same test on two occasions (test-retest approach) or administering one form of a test on one occasion and a second form of a test believed to be equivalent on a second occasion (equivalent-forms approach). In order to assess stability, the time between the two testings should be at least two weeks for either approach. The minimum sample size required for an acceptable study is 50 individuals. The resulting correlation coefficients between test scores on two occasions are to be .75 or higher. If subtest scores are used to make placement decisions, the correlations between these occasions for these subtest scores must also be .75 or higher.
- b. Standard Errors of Measurement. Standard errors of measurement are to be provided for intervals across the score scale or at likely cut points.

3. Testing Special Groups.

- a. Publishers who seek to have assessment instruments approved for use in the California Community Colleges must agree to provide the test and response forms in alternate media upon request of a college. Testing instruments must be available in a place and manner accessible to persons with disabilities or offer alternative accessible arrangements for such individuals (i.e., Braille, large print, audiotape, and electronic tests). We believe this is consistent with the intent of Section 36.309 of the

American with Disabilities Act and applies to all publishers of testing instruments. Such a commitment is required for a test to be placed on the Chancellor's list of approved tests.

B. Primary Responsibility of the Local College or District

Colleges using an approved second party instrument are to maintain information and documentation locally that addresses the following Standards. This material should be up-dated every five years and will be evaluated during each on-site review team visit. The evidence to be assembled is to be done by the local college and is not the duty of the second-party publisher to perform the data collection for the local college.

1. Validity and Fairness

- a. Content-Related Validity Evidence. The college is to provide a comprehensive description of the appropriateness of a test for placement into a course or courses in a sequence at the college based on the overlap of knowledges and skills measured by the test and those knowledges/skills required as prerequisites to the course(s). Content-related validity evidence for a test's scores needs to be grounded in statements of specific pre-course expectations (i.e., prerequisite skills) which can then be linked to the actual tested skills (i.e., test items). Documentation of specific prerequisite skills is critical to portraying a course along with its objectives which together define the skills needing to be evaluated on a given test. Information addressing the extent of the link between pre-course skill requirements and the specific content measured by the test provides strong rational evidence of the content representativeness of the test for the course.

Procedurally, local college faculty are to evaluate the content representativeness of the test by participating in an item-by-item evaluation of the test content with reference to the skills and/or body of knowledge deemed necessary at entrance to each course within the sequence into which placement recommendations would occur based on performance on the test. Summary information that describes how and the extent to which judges (i.e., instructors) evaluate the fit between the test item content and skills required by the course(s) is to be provided. From such a faculty review process, data and tabulations addressing the extent of the match between critical (or representative) pre-course skills and assessed content/skills are to be reported. As complementary information, the necessary prerequisites not evaluated by the instrument should be noted as well as those skills/content tested that are not relevant to a placement decision into a specific course.

Specifically, the minimum data required* from a college to address content-related validity evidence must respond to the following test characteristics based on faculty evaluation of an instrument:

- Are the specific pre-course skills and content knowledge that need to have been mastered by a student to be placed appropriately in a course measured by the instrument?
- To what extent are these pre-course skills adequately assessed by the candidate instrument's items? Then, to what extent do the

* When the test can be used to "test out" of a course or course sequence (e.g., ESL), then a content-related validity evaluation **must** also document and judge the appropriateness and representativeness of the test's items for the objectives of that course.

test's items assess skills and knowledge that are not pre-course expectations?

The higher the extent of the overlap judged by instructors between course prerequisite skills and those skills measured by the test, the stronger the evidence in support of content-related validity. The extent to which a test measures non-prerequisite skills is to be documented and considered when judging the appropriateness of the test.

NOTE: Direct evidence addressing the criterion-related validity or consequential-related validity Standards is not necessarily required. Such evidence or similar evidence need only be submitted if an empirical design is chosen as the procedure for establishing or validating cut-scores.

- b. Criterion-Related Validity Evidence. Evidence addressing criterion-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. (See Section I. B. 1. e. below)
- c. Consequential-Related Validity Evidence. Evidence addressing consequential-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. . (See Section I. B. 1. e. below)
- d. Evidence Addressing Test Bias. Local community colleges will review the evidence addressing test bias supplied by the test developer (see I.A.1.c) to ensure that the results are generalizable to their colleges.
- e. Evidence Addressing Adequacy of the Cut Score(s). It is the local community college's responsibility to validate its Cut Scores. Data are to be collected by the individual college to justify the selection of any Cut Scores or score ranges used for placement advice. The adequacy of any cut score may be demonstrated by either a judgmental or empirical approach. A judgmental approach typically focuses on setting the initial cut score. However, if such judgmental data are to be used as the only evidence to support the adequacy of the cut score, then a formal procedure documented in the literature on setting cut-scores needs to be implemented (See Berk, 1986 and Jaeger, 1989 for appropriate procedures). Not only do the process and the results need to be documented, but also a description of the persons involved and their credentials for having participated in the process need to be provided.

As related to empirical procedures, at a minimum the data should demonstrate that individuals who score above the cut score or within the

score range identified have a greater expectancy of success (e.g., appear to be more prepared for the course based on instructor ratings, or a mid-term grade, or obtaining a C grade or higher) in a specific course for which placement recommendations are made than those who score below the score or score range. . Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

If a consequential-related validity design is used as the primary source of cut-score validity data, the following are to be considered. Under any approach chosen by a college, at a minimum items (A) and (B) below must be formally addressed, and positive evaluations on questions such as these must be reported for the instrument to be fully approved. Other research questions are possible and other investigative orientations are encouraged. (Items C and D which follow are optional and illustrative, but could be extremely useful sources of information to colleges who may choose to pursue such lines of inquiry.)

- (A) After the first few weeks of a course, how do students whose test scores recommend placement into a particular class evaluate the appropriateness and/or usefulness of their placement in that course? (The Standard is at least 75% affirmative endorsement by students.)
- (B) After the first few weeks of a class, how do instructors evaluate the readiness of individual students (those who follow their test performance recommendations) to undertake the material of their class? (The Standard is at least 75% judgment of proper placement by instructors.)
- (C) For students who opt not to follow a test's recommendation, how do these students fare (in terms of material learned, suitability of the placement, and their likelihood of successful matriculation) in the classes into which they choose to enroll, and can such performance be justified/expected?

- (D) What do students and instructors identify as undesirable results of an “incorrect” course placement and what are the consequences (on students, instructors, academic units and the institution) of such decisions?

Methods and procedures for carrying out such inquiries are a local college decision and preference. Proper and reasonable investigative approaches are expected to be followed (e.g., double blind experimentation, sufficient sample sizes, maintaining an objective judgment process, etc.).

2. Reliability and Errors of Measurement.

- a. Local community colleges will review the evidence addressing reliability and errors of measurement supplied by the test developer (refer to I.A.2) to ensure that the results are generalizable to their colleges.

3. Impact of Testing on Various Groups

- a. Disproportionate Impact. Disproportionate impact must be monitored on an ongoing basis for various demographic groups (including gender, age, racial, ethnic, and disability groups [refer to page 6]). A record of these data is to be maintained and periodically evaluated. These studies are expected to be conducted at least every three years. When disproportionate impact is observed, the college/district shall, in consultation with the State Chancellor, develop and implement a plan describing the steps the college or district will take to correct the disproportionate impact, including studies of differential prediction (Title 5, Section 55512 [a]). Colleges may consult the EEOC guidelines (see pp. 11 and 24) and the Local Research Options Handbooks (November 1989 and February 1991) for clarification on the definition, identification and treatment of disproportionate impact. The notion that disproportionate impact is to be continuously monitored must not be overlooked.
- b. Standardization. If the instrument is revised for testing of individuals who cannot take the tests under standard conditions, then there must be documentation of all changes along with the basis for any change. The justification for changed or altered assessment instruments or procedures must be on file at the local college.

II. **Specific Criteria for Objectively Scored Tests Developed or Managed by a California Community College or District**

A. **Primary Responsibility of the Local College or District**

1. Validity and Fairness

- a. Content-Related Validity Evidence. The college is to provide a comprehensive description of the appropriateness of a test for placement into a course or courses in a sequence at the college based on the overlap of knowledges and skills measured by the test and those knowledges/skills required as prerequisites to the course(s). Content-related validity evidence for a test's scores needs to be grounded in statements of specific pre-course expectations (i.e., prerequisite skills) which can then be linked to the actual tested skills (i.e., test items). Documentation of specific prerequisite skills is critical to portraying a course along with its objectives which together define the skills needing to be evaluated on a given test. Information addressing the extent of the link between pre-course skill requirements and the specific content measured by the test provides strong rational evidence of the content representativeness of the test for the course.

Procedurally, local college faculty are to evaluate the content representativeness of the test by participating in an item-by-item evaluation of the test content with reference to the skills and/or body of knowledge deemed necessary at entrance to each course within the sequence into which placement recommendations would occur based on performance on the test. Summary information that describes how and the extent to which judges (i.e., instructors) evaluate the fit between the test item content and skills required by the course(s) is to be provided. From such a faculty review process, data and tabulations addressing the extent of the match between critical (or representative) pre-course skills and assessed content/skills are to be reported. As complementary information, the necessary prerequisites not evaluated by the instrument should be noted as well as those skills/content tested that are not relevant to a placement decision into a specific course.

Specifically, the minimum data required* from a college to address content-related validity evidence must respond to the following test characteristics based on faculty evaluation of an instrument:

* When the test can be used to "test out" of a course or course sequence (e.g., ESL), then a content-related validity evaluation **must** also document and judge the appropriateness and representativeness of the test's items for the objectives of that course.

- Are the specific pre-course skills and content knowledge that need to have been mastered by a student to be placed appropriately in a course measured by the instrument?
- To what extent are these pre-course skills adequately assessed by the candidate instrument's items? Then, to what extent do the test's items assess skills and knowledge that are not pre-course expectations?

The higher the extent of the overlap judged by instructors between course prerequisite skills and those skills measured by the test, the stronger the evidence in support of content-related validity. The extent to which a test measures non-prerequisite skills is to be documented and considered when judging the appropriateness of the test.

NOTE: Direct evidence addressing the criterion-related validity or consequential-related validity Standards is not necessarily required. Such evidence or similar evidence need only be submitted if an empirical design is chosen as the procedure for establishing or validating cut-scores.

- b. Criterion-Related Validity Evidence. Evidence addressing criterion-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. (See Section II. A. 1. e. below)
- c. Consequential-Related Validity Evidence. Evidence addressing consequential-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. (See Section II. A. 1. e. below)
- d. Evidence Addressing Test Bias. Evidence focusing on cultural/linguistic bias, insensitivity and offensiveness must be provided. This evidence needs to consist of evaluations of test items by diverse panels who reflect the college's student population or results from test data that address bias. A description of the panel members' appropriateness for conducting the review should be included if that method is used. The findings from these investigations should be used to eliminate or minimize test bias, insensitivity, and offensiveness. When a college is using a second-party instrument for which there is adequate evidence from the test publisher or from another college's study that the instrument minimizes bias, offensiveness and insensitivity, then additional data from the college is not required. The latter evidence may be cited. A college may also supplement such available evidence as needed.

- e. Evidence Addressing Adequacy of the Cut Score(s). It is the local community college's responsibility to validate its Cut Scores. Data are to be collected by the individual college to justify the selection of any Cut Scores or score ranges used for placement advice. The adequacy of any cut score may be demonstrated by either a judgmental or empirical approach. A judgmental approach typically focuses on setting the initial cut score. However, if such judgmental data are to be used as the only evidence to support the adequacy of the cut score, then a formal procedure documented in the literature on setting cut-scores needs to be implemented (See Berk, 1986 and Jaeger, 1989 for appropriate procedures). Not only do the process and the results need to be documented, but also a description of the persons involved and their credentials for having participated in the process need to be provided.

As related to empirical procedures, at a minimum the data should demonstrate that individuals who score above the cut score or within the score range identified have a greater expectancy of success (e.g., appear to be more prepared for the course based on instructor ratings, or a mid-term grade, or obtaining a C grade or higher) in a specific course for which placement recommendations are made than those who score below the score or score range. Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

If a consequential-related validity design is used as the primary source of cut-score validity data, the following are to be considered. Under any approach chosen by a college, at a minimum items (A) and (B) below must be formally addressed, and positive evaluations on questions such as these must be reported for the instrument to be fully approved. Other research questions are possible and other investigative orientations are encouraged. (Items C and D which follow are optional and illustrative, but could be extremely useful sources of information to colleges who may choose to pursue such lines of inquiry.)

- (A) After the first few weeks of a course, how do students whose test scores recommend placement into a particular class evaluate the appropriateness and/or usefulness of their placement in that course? (The Standard is at least 75% affirmative endorsement by students.)
- (B) After the first few weeks of a class, how do instructors evaluate the readiness of individual students (those who follow their test performance recommendations) to undertake the material of their class? (The Standard is at least 75% judgment of proper placement by instructors.)
- (C) For students who opt not to follow a test's recommendation, how do these students fare (in terms of material learned, suitability of the placement, and their likelihood of successful matriculation) in the classes into which they choose to enroll, and can such performance be justified/expected?
- (D) What do students and instructors identify as undesirable results of an "incorrect" course placement and what are the consequences (on students, instructors, academic units and the institution) of such decisions?

Methods and procedures for carrying out such inquiries are a local college decision and preference. Proper and reasonable investigative approaches are expected to be followed (e.g., double blind experimentation, sufficient sample sizes, maintaining an objective judgment process, etc.).

2. Reliability and Errors of Measurement

- a. Reliability. Either the internal consistency or the stability of a placement test should be evaluated. Evidence documenting the internal consistency reliability should be based on appropriate procedures, e.g., split-half coefficients, Kuder-Richardson indices, alpha coefficients, etc. The minimum acceptable value from these internal consistency measures is set at .80. The stability may be assessed by either administering the same test on two occasions (test-retest approach) or administering one form of a test on one occasion and a second form of a test believed to be equivalent on a second occasion (equivalent-forms approach). In order to assess stability, the time between testings should be at least two weeks for either approach. The resulting correlation coefficients between test scores on two occasions are to be .75 or higher. If subtest scores are used to make placement decisions, the correlations between these occasions for these subtest scores must also be .75 or higher. A minimum sample size on which any reliability

index is to be based is set at 50 individuals. A college using a second-party test for which stability coefficients exist from either the publisher or another college, and that meet the California Community College assessment standards, may rely on this information to support evidence of reliability. If an internal consistency approach to addressing the reliability of a test's scores is chosen, that index needs to be based on a study involving the students at the college.

- b. Standard Errors of Measurement. Standard errors of measurement are to be provided for the entire test, for intervals across the score scale or for the cut point.

3. Impact of Testing on Various Groups

- a. Disproportionate Impact. Disproportionate impact must be monitored on an ongoing basis for various demographic groups (including gender, age, racial, ethnic, and disability groups [refer to page 7]). A record of these data is to be maintained and periodically evaluated. These studies are expected to be conducted at least every three years. When disproportionate impact is observed, the college/district shall, in consultation with the State Chancellor, develop and implement a plan describing the steps the college or district will take to correct the disproportionate impact, including studies of differential prediction (Title 5, Section 55512 [a]). Colleges may consult the EEOC guidelines (see pp. 11 and 24) and the Local Research Options Handbooks (November 1989 and February 1991) for clarification on the definition, identification and treatment of disproportionate impact. The notion that disproportionate impact is to be continuously monitored must not be overlooked.

For the initial submission for approval of a test the college is to submit their plan describing how data on disproportionate impact will be monitored and evaluated locally. Actual data and findings are not an expectation at the time of a first submission, although information is welcomed if available. When a test instrument is submitted for approval later under the "renewal" process, the college must submit findings from their disproportionate impact monitoring, and actions and results that may have followed based on findings. Full renewal approval will be possible only when disproportionate impact findings are reported and judged acceptable.

- b. Standardization. If the instrument is revised for testing of individuals who cannot take the tests under standard conditions, then there must be documentation of all changes along with the basis for any change. The justification for changed or altered assessment instruments or procedures must be on file at the local college.

III. Specific Criteria for Direct Performance Assessments

Note: Performance assessments as discussed in this section include direct writing samples, oral interviews and other performance-based tasks that are scored using a well-defined (i.e., operational) scoring rubric.

A. Primary Responsibilities of Second-Party Direct Performance Assessment Developers

1. Validity and Fairness

- a. Content-Related Validity Evidence. Detailed documentation must be provided describing the guidelines for the development of performance assessment questions, prompts or tasks and the related scoring rubrics. Documentation should also include a description of the development and scoring procedures for the questions, prompts or tasks. A description of how the raters are trained to yield standardization in the performance assessment process and outcomes should be provided.
- b. Criterion-Related or Consequential-Related Validity Evidence. Data must be presented to indicate that the test is useful for making placement decisions for California Community College student populations and courses in these colleges. Empirical evidence should support the following conclusion: test takers who achieve scores within some specified range should take a different course or set of courses in comparison with test takers who score outside that range. Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

When submitting evidence to meet this standard, the following design criteria should be met: the course content should bear a close logical relationship with courses offered by the California Community Colleges;

the students should be similar to the students enrolled in the California Community Colleges; when used as the primary index, the correlation between the test and a student's readiness to assimilate course content or performance (e.g., mid-term grade, student or instructor evaluation, end-of-course grade) should be substantial; and, across all data sets presented by a test publisher, the average correlation should be greater than or equal to .35 (or a comparable value if an alternative statistical analysis was performed). The magnitude of the correlation may vary as a function of the degree to which a test was used to place students in the course under investigation and/or the variation in grading standards across classrooms. Coefficients corrected for restriction of range are acceptable.

As a guide, supportive data from at least six community colleges are required to attain full approval status; supportive data from at least four community colleges are required to attain provisional approval status; and supportive data from at least three community colleges are required to attain probationary approval status. Additionally, a majority of the colleges included must be California Community Colleges representing the diversity of courses and students in the California Community College system, i.e., 4 of 6, 3 of 4 or 2 of 3, to attain a specific approval level status.

- c. Evidence Addressing Test Bias. Evidence focusing on cultural/linguistic bias, insensitivity and offensiveness must be provided. This evidence needs to be of two types: evaluations of questions, prompts or tasks by culturally and linguistically diverse panels and results from score data that address bias. These studies need to be done using groups that are similar to those student groups ordinarily served by the California Community Colleges. The findings from these investigations should be used to eliminate or minimize test bias, insensitivity, and offensiveness. When assessment procedures are in place to contend with possible bias, for example providing students a choice of questions, prompts or tasks or placing no restriction on question, prompt or task chosen, the methodology must be clearly stated and described so a determination of equivalence of scoring may be determined.

2. Reliability

- a. Reliability. Interscorer reliability coefficients should be provided. If correlation coefficients are provided, these coefficients should be greater than .70. If percent agreement indices are provided, they should yield at least 90 percent agreement between scores, where an agreement is within 1 scale point on a 6-point scale. Additionally, how inconsistencies between scorers are resolved should be described.

When multiple question sets, prompts, or tasks are in use, equivalent-forms reliability coefficients should be reported for a subsample of available questions, sets, prompts, or tasks in use. The resulting correlation coefficients between scores on different question sets, prompts, or tasks should be .75 or higher.

3. Testing Special Groups.

- a. Publishers who seek to have assessment instruments approved for use in the California Community Colleges must agree to provide the test and response forms in alternate media upon request of a college. Testing instruments must be available in a place and manner accessible to persons with disabilities or offer alternative accessible arrangements for such individuals (i.e., Braille, large print, audiotape, and electronic tests). We believe this is consistent with the intent of Section 36.309 of the American with Disabilities Act and applies to all publishers of testing instruments. Such a commitment is required for a test to be placed on the Chancellor's list of approved tests.

B. Primary Responsibilities of the Local College or District for a Second-Party Developed Performance Assessment

Colleges using an approved second party instrument are to maintain information and documentation locally that addresses the following Standards. This material should be up-dated every five years and will be evaluated during each on-site review team visit. The evidence to be assembled is to be done by the local college and is not the duty of the second-party publisher to perform the data collection for the local college.

1. Validity and Fairness

- a. Content-Related Validity Evidence. The college is to provide a comprehensive description of the appropriateness of a test for placement into a course or courses in a sequence at the college based on the overlap of knowledges and skills measured by the assessment and those knowledges/skills required as prerequisites to the course(s). Content-related validity evidence for a performance assessment's scores needs to be grounded in statements of specific pre-course expectations (i.e., prerequisite skills) which can then be linked to the actual tested skills (i.e., test items). Documentation of specific prerequisite skills is critical to portraying a course along with its objectives which together define the skills needing to be evaluated by a given assessment. Information addressing the extent of the link between pre-course skill requirements and the specific content measured by the assessment provides strong rational evidence of the content representativeness of the test for the course.

Procedurally, local college faculty are to evaluate the content representativeness of the performance assessment by participating in an evaluation of the question, prompt or task content with reference to the skills and/or body of knowledge deemed necessary at entrance to each course within the sequence into which placement recommendations would occur based on the scoring of responses to the performance assessment. Summary information that describes how and the extent to which judges (i.e., instructors) evaluate the fit between the content of the questions, prompts or tasks and skills required by the course(s) is to be provided. From such a faculty review process, data and tabulations addressing the extent of the match between critical (or representative) pre-course skills and assessed content/skills are to be reported. As complementary information, the necessary prerequisites not evaluated by the instrument should be noted as well as those skills/content tested that are not relevant to a placement decision into a specific course.

Specifically, the minimum data required* from a college to address content-related validity evidence must respond to the following test characteristics based on faculty evaluation of a performance assessment instrument:

- Are the specific pre-course skills and content knowledge that need to have been mastered by a student to be placed appropriately in a course measured by the performance assessment?
- To what extent are these pre-course skills adequately assessed by the performance assessment's rubric? Then, to what extent do the assessment's rubric evaluate skills and knowledge that are not pre-course expectations?

The higher the extent of the overlap judged by instructors between course prerequisite skills and those skills measured by the instrument, the stronger the evidence in support of content-related validity. The extent to which an instrument measures non-prerequisite skills is to be documented and considered when judging the appropriateness of the instrument.

NOTE: Direct evidence addressing the criterion-related validity or consequential-related validity Standards is not necessarily required. Such evidence or similar evidence need only be submitted if an empirical

* When the test can be used to "test out" of a course or course sequence (e.g., ESL), then a content-related validity evaluation **must** also document and judge the appropriateness and representativeness of the test's items for the objectives of that course.

design is chosen as the procedure for establishing or validating cut-scores.

- b. Criterion-Related Validity Evidence. Evidence addressing criterion-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. (See Section III. B. 1. e. below)
- c. Consequential-Related Validity Evidence. Evidence addressing consequential-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. . (See Section III. B. 1. e. below)
- d. Evidence Addressing Test Bias. Local community colleges will review the evidence addressing test bias supplied by the test developer (see III.A.1.c) to ensure that the results are generalizable to their colleges.
- e. Evidence Addressing Adequacy of the Cut Score(s). It is the local community college's responsibility to validate its Cut Scores. Data are to be collected by the individual college to justify the selection of any Cut Scores or score ranges used for placement advice. The adequacy of any cut score may be demonstrated by either a judgmental or empirical approach. A judgmental approach typically focuses on setting the initial cut score. However, if such judgmental data are to be used as the only evidence to support the adequacy of the cut score, then a formal procedure documented in the literature on setting cut-scores needs to be implemented (See Berk, 1986 and Jaeger, 1989 for appropriate procedures). Not only do the process and the results need to be documented, but also a description of the persons involved and their credentials for having participated in the process need to be provided.

As related to empirical procedures, at a minimum the data should demonstrate that individuals who score above the cut score or within the score range identified have a greater expectancy of success (e.g., appear to be more prepared for the course based on instructor ratings, or a mid-term grade, or obtaining a C grade or higher) in a specific course for which placement recommendations are made than those who score below the score or score range. . Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific

courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

If a consequential-related validity design is used as the primary source of cut-score validity data, the following are to be considered. Under any approach chosen by a college, at a minimum items (A) and (B) below must be formally addressed, and positive evaluations on questions such as these must be reported for the instrument to be fully approved. Other research questions are possible and other investigative orientations are encouraged. (Items C and D which follow are optional and illustrative, but could be extremely useful sources of information to colleges who may choose to pursue such lines of inquiry.)

- (A) After the first few weeks of a course, how do students whose test scores recommend placement into a particular class evaluate the appropriateness and/or usefulness of their placement in that course? (The Standard is at least 75% affirmative endorsement by students.)
- (B) After the first few weeks of a class, how do instructors evaluate the readiness of individual students (those who follow their test performance recommendations) to undertake the material of their class? (The Standard is at least 75% judgment of proper placement by instructors.)
- (C) For students who opt not to follow a test's recommendation, how do these students fare (in terms of material learned, suitability of the placement, and their likelihood of successful matriculation) in the classes into which they choose to enroll, and can such performance be justified/expected?
- (D) What do students and instructors identify as undesirable results of an "incorrect" course placement and what are the consequences (on students, instructors, academic units and the institution) of such decisions?

Methods and procedures for carrying out such inquiries are a local college decision and preference. Proper and reasonable investigative approaches are expected to be followed (e.g., double blind experimentation, sufficient sample sizes, maintaining an objective judgment process, etc.).

2. Reliability

- a. Local community colleges will review the evidence addressing reliability supplied by the test developer (see III.A.2) to insure that the results are generalizable to their colleges.

3. Impact of Testing on Various Groups

- a. Disproportionate Impact. Disproportionate impact must be monitored for various demographic groups (including gender, age, racial, linguistic and disability groups [refer to page 7]). A record of these data is to be maintained. These studies are expected to be conducted at least every three years. When there is a disproportionate impact, the district shall, in consultation with the State Chancellor, develop and implement a plan describing the steps the district will take to correct the disproportionate impact, including studies of differential prediction (Title 5, Section 55512 [a]). Colleges may consult the EEOC guidelines (see pp. 11 and 24) and the Local Research Options Handbooks (November 1989 and February 1991) for clarification on the definition, identification and treatment of disproportionate impact. The notion that disproportionate impact is to be continuously monitored must not be overlooked.
- b. Standardization. If the instrument is revised for testing of individuals who cannot take the tests under standard conditions, then there must be documentation of all changes along with the basis for any change. The justification for changed or altered assessment instruments or procedures must be on file at the local college.

C. Primary Responsibilities of the Local College or District Developed or Managed Direct Performance Assessment

1. Validity and Fairness

- a. Content-Related Validity Evidence. The college is to provide a comprehensive description of the appropriateness of a performance assessment for placement into a course or courses in a sequence at the college based on the overlap of knowledges and skills measured by the performance assessment and those knowledges/skills required as prerequisites to the course(s). Content-related validity evidence for a performance assessment's scores needs to be grounded in statements of specific pre-course expectations (i.e., prerequisite skills) which can then be linked to the actual tested skills (i.e., performance assessment questions, prompts or tasks). Documentation of specific prerequisite skills are critical to portraying a course along with its objectives which together define the skills needing to be evaluated on a given

performance assessment. Information addressing the extent of the link between pre-course skill requirements and the specific content measured by the performance assessment provides strong rational evidence of the content representativeness of the performance assessment for the course.

Procedurally, local college faculty are to evaluate the content representativeness of the performance assessment by participating in an evaluation of the question, prompt or task content with reference to the skills and/or body of knowledge deemed necessary at entrance to each course within the sequence into which placement recommendations would occur based on the scoring of responses to the performance assessment. Summary information that describes how and the extent to which judges (i.e., instructors) evaluate the fit between the content of the questions, prompts or tasks and skills required by the course(s) is to be provided. From such a faculty review process, data and tabulations addressing the extent of the match between critical (or representative) pre-course skills and assessed content/skills are to be reported. As complementary information, the necessary prerequisites not evaluated by the instrument should be noted as well as those skills/content tested that are not relevant to a placement decision into a specific course.

Specifically, the minimum data required* from a college to address content-related validity evidence must respond to the following test characteristics based on faculty evaluation of a performance assessment instrument:

- Are the specific pre-course skills and content knowledge that need to have been mastered by a student to be placed appropriately in a course measured by the performance assessment?
- To what extent are these pre-course skills adequately assessed by the performance assessment's rubric? Then, to what extent do the assessment's rubric evaluate skills and knowledge that are not pre-course expectations?

The higher the extent of the overlap judged by instructors between course prerequisite skills and those skills measured by the performance assessment, the stronger the evidence in support of content-related validity. The extent to which a performance assessment measures non-prerequisite skills is to be documented and considered when judging its appropriateness.

* When the test can be used to "test out" of a course or course sequence (e.g., ESL), then a content-related validity evaluation **must** also document and judge the appropriateness and representativeness of the test's items for the objectives of that course.

In addition, documentation must be supplied describing the development of performance assessment questions, prompts or tasks and the related scoring rubrics. Documentation should also include a description of the resulting products, i.e., the questions, prompts, or tasks and the rubric used in scoring. A description of how the raters are trained to yield standardization in the performance assessment process and outcomes should be provided as well.

NOTE: Direct evidence addressing the criterion-related validity or consequential-related validity Standards is not necessarily required. Such evidence or similar evidence need only be submitted if an empirical design is chosen as the procedure for establishing or validating cut-scores.

- b. Criterion-Related Validity Evidence. Evidence addressing criterion-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. (See Section III. C. 1. e. below)
- c. Consequential-Related Validity Evidence. Evidence addressing consequential-related validity need only be collected if such a design were implemented as the mechanism to provide the empirical validation of local cut-scores. . (See Section III. C. 1. e. below)
- d. Evidence Addressing Test Bias. Evidence focusing on cultural/linguistic bias, insensitivity and offensiveness must be provided. This evidence needs to consist of evaluations of the questions, prompts or tasks by diverse panels who reflect the college's student population or results from score data that address bias. A description of the panel members' appropriateness for conducting the review should be included. The findings from these investigations should be used to eliminate or minimize test bias, insensitivity, and offensiveness. . When a college is using a second-party instrument for which there is adequate evidence from the test publisher or from another college's study that the instrument minimizes bias, offensiveness and insensitivity, then additional data from the college is not required. The latter evidence may be cited. A college may also supplement such available evidence as needed.

When assessment procedures are in place to contend with possible bias, for example providing students a choice of questions, prompts or tasks or placing no restriction on question, prompt or task chosen, the

methodology must be clearly stated and described so a determination of equivalence of scoring may be determined.

- e. Evidence Addressing Adequacy of the Cut Score(s). It is the local community college's responsibility to validate its Cut Scores. Data are to be collected by the individual college to justify the selection of any Cut Scores or score ranges used for placement advice. The adequacy of any cut score may be demonstrated by either a judgmental or empirical approach. A judgmental approach typically focuses on setting the initial cut score. However, if such judgmental data are to be used as the only evidence to support the adequacy of the cut score, then a formal procedure documented in the literature on setting cut-scores needs to be implemented (See Berk, 1986 and Jaeger, 1989 for appropriate procedures). Not only do the process and the results need to be documented, but also a description of the persons involved and their credentials for having participated in the process need to be provided.

As related to empirical procedures, at a minimum the data should demonstrate that individuals who score above the cut score or within the score range identified have a greater expectancy of success (e.g., appear to be more prepared for the course based on instructor ratings, or a mid-term grade, or obtaining a C grade or higher) in a specific course for which placement recommendations are made than those who score below the score or score range. . Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

If a consequential-related validity design is used as the primary source of cut-score validity data, the following are to be considered. Under any approach chosen by a college, at a minimum items (A) and (B) below must be formally addressed, and positive evaluations on questions such as these must be reported for the instrument to be fully approved. Other research questions are possible and other investigative orientations are encouraged. (Items C and D which follow are optional and illustrative,

but could be extremely useful sources of information to colleges who may choose to pursue such lines of inquiry.)

- (A) After the first few weeks of a course, how do students whose test scores recommend placement into a particular class evaluate the appropriateness and/or usefulness of their placement in that course? (The Standard is at least 75% affirmative endorsement by students.)
- (B) After the first few weeks of a class, how do instructors evaluate the readiness of individual students (those who follow their test performance recommendations) to undertake the material of their class? (The Standard is at least 75% judgment of proper placement by instructors.)
- (C) For students who opt not to follow a test's recommendation, how do these students fare (in terms of material learned, suitability of the placement, and their likelihood of successful matriculation) in the classes into which they choose to enroll, and can such performance be justified/expected?
- (D) What do students and instructors identify as undesirable results of an "incorrect" course placement and what are the consequences (on students, instructors, academic units and the institution) of such decisions?

Methods and procedures for carrying out such inquiries are a local college decision and preference. Proper and reasonable investigative approaches are expected to be followed (e.g., double blind experimentation, sufficient sample sizes, maintaining an objective judgment process, etc.).

2. Reliability

- a. Reliability. Interscorer reliability coefficients should be provided. If correlation coefficients are provided, these coefficients should be greater than .70. If percent agreement indices are provided, they should yield at least 90 percent agreement between scores, where an agreement is within 1 scale point on a 6-point scale. Additionally, how inconsistencies between scorers are resolved should be described.

When multiple question sets, prompts or tasks are in use, equivalent-forms reliability coefficients should be reported for a subsample of available question sets, prompts or tasks in use. The resulting correlation coefficients between scores on different question sets, prompts or tasks should be .75 or higher.

3. Impact of Testing on Various Groups

- a. Disproportionate Impact. Disproportionate impact must be monitored on an ongoing basis for various demographic groups (including gender, age, racial, ethnic, and disability groups [refer to page 7]). A record of these data is to be maintained and periodically evaluated. These studies are expected to be conducted at least every three years. When disproportionate impact is observed, the college/district shall, in consultation with the State Chancellor, develop and implement a plan describing the steps the college or district will take to correct the disproportionate impact, including studies of differential prediction (Title 5, Section 55512 [a]). Colleges may consult the EEOC guidelines (see pp. 11 and 24) and the Local Research Options Handbooks (November 1989 and February 1991) for clarification on the definition, identification and treatment of disproportionate impact. The notion that disproportionate impact is to be continuously monitored must not be overlooked.

- b. Standardization. If the instrument is revised for testing of individuals who cannot take the test under standard conditions, then there must be documentation of all changes along with the basis for any change. The justification for changed or altered assessment instruments or procedures must be on file at the local college.

The matrix on pages 43 to 47 presents a summary of the requirements and expectations discussed above.

SECTION FOUR: SPECIFIC CRITERIA FOR COMPUTER TESTING

(The following applies to second party publishers, and locally controlled and managed assessments.)

A new generation of testing formats provides for assessments that present tests via electronic media, i.e., computer testing. In this framework, tests are typically computer administered or computer adaptive tests. In the case of the former, computer administered testing, the presentation of a fixed assessment is controlled by the computer wherein all examinees are administered the same test in a standardized format. Variation from the traditional paper and pencil format could include controlling the time examinees are given to respond to an item, not allowing examinees to return to items already presented, allowing for or supporting a “respond until correct” approach, etc. In the framework of computer adaptive testing, the items administered to each examinee are selected from an item pool uniquely intended for the person based on the pattern of correct and incorrect answers to previous questions. In computer adaptive testing, two examinees may take different questions yet both will receive a score on the same scale. Both approaches, computer adaptive and computer-administered tests are becoming commonplace in testing. The California community colleges allow the use of assessments in the computer testing or traditional paper and pencil formats.

Criteria associated with the review and evaluation of computer tests follow below. In most cases, the criteria and guidelines presented in Section II & III apply to electronically formatted testing. The criteria presented below enhance and in some cases modify those requirements and expectations. As this is a relatively new and evolving area, publishers are advised to contact the Chancellor Office to determine if there may have been changes to these specific criteria.

A. Primary Responsibility of the Test Developer

1. Validity and Fairness

- a. Content-Related Validity Evidence. The test developer is to describe the content of the test's items with sufficient and clear specificity. Explicit statements of test objectives and tables of specification need to be available to inform college users. Based on these sources of content description, users at community colleges are able to consider the test's appropriateness for making placement recommendations for a sequence of courses at their colleges.

So that local colleges can conduct required content-related validity, test publishers must provide test booklets or a sufficient representative sample of test items such that local colleges may conduct an item-by-

item review. If a sample of items is provided, the number should be such that they represent what is a psychometrically sound single form of a traditional paper-and-pencil administered form of the test.

Further, in the case of computer adaptive tests that rely on item banks (also know as item pools) from which items are chosen as testing is initiated or underway (i.e., a non standard collection of questions are administered to examinees), there must be documentation of the algorithms (rational or empirical) that are used to select items which lead to estimates of examinees' scores. Presentation and discussion of the conditions regarding the selection of the tested subset of items that assures the representative of and appropriateness for the content domain test specifications being evaluate needs to be offered.

Given the newness of computer testing, it is desirable for there to be sample and illustrative testing modules so as to inform and advise prospective users in their decision to adopt a device.

Publishers must provide documentation that item pools will be periodically (i.e., annually) reviewed to excise items that are no longer appropriate for the content domain(s) being evaluated. As item pools can be extremely large collections of items, unlike traditional paper and pencil tests whose structure is finite and fixed, some computer test items may become outdated and otherwise inappropriate for the pool and thus inappropriate for domain measurement.

- b. Criterion-Related or Consequential-Related Validity Evidence. Data must be presented to indicate that the test is useful for making placement decisions for California Community College student populations and courses in these colleges. Empirical evidence should support the following conclusion: test takers who achieve scores within some specified range should take a different course or set of courses in comparison with test takers who score outside that range. Several different approaches to data collection are allowable in securing evidence that support this conclusion. Either criterion-related or consequential-related evidence is permissible to meet this standard. As a general principle, criterion-related studies are most appropriate when the test being evaluated has not been used for placement into specific courses; consequential-related studies are most appropriate when scores on the test being evaluated have been used to assist student placement into specific courses. For criterion-related studies, use of a variety of designs and variables as criteria are permitted. This includes mean difference or correlational designs. Criterion variables might include student ratings of ability to meet course requirements, instructor

ratings of students' abilities to meet course requirements, midterm grades or test scores, final exam grades or test scores, etc.

When submitting evidence to meet this standard, the following design criteria should be met: the course content should bear a close logical relationship with courses offered by the California Community Colleges; the students should be similar to the students enrolled in the California Community Colleges; when used as the primary index, the correlation between the test and a student's readiness to assimilate course content or performance (e.g., mid-term grade, student or instructor evaluation, end-of-course grade) should be substantial; and, across all data sets presented by a test publisher, the average correlation should be greater than or equal to .35 (or a comparable value if an alternative statistical analysis was performed). The magnitude of the correlation may vary as a function of the degree to which a test was used to place students in the course under investigation and/or the variation in grading standards across classrooms. Coefficients corrected for restriction of range are acceptable.

As a guide, supportive data from at least six community colleges are required to attain full approval status; supportive data from at least four community colleges are required to attain provisional approval status; and supportive data from at least three community colleges are required to attain probationary approval status. Additionally, a majority of the colleges included must be California Community Colleges representing the diversity of courses and students in the California Community College system, i.e., 4 of 6, 3 of 4 or 2 of 3, to attain a specific approval level status.

- c. Evidence Addressing Test Bias. Evidence focusing on cultural/linguistic bias, insensitivity and offensiveness must be provided. This work should be done on student groups that are similar to those ordinarily served by the California Community Colleges. This evidence needs to be of two types: evaluations of test items by culturally and linguistically diverse panels and results from test data that address bias. The findings from these investigations should be used to eliminate or minimize test bias, insensitivity, and offensiveness.

Further, in the case of computer adaptive tests that rely on item banks (pools) from which items are chosen, it may be the case that not all items in the pool can realistically be empirically "pre-evaluated" for bias/differential item functioning. To permit data to be gathered by the publisher and allow the colleges to have access to promising and potentially acceptable testing tools, the following accommodations are permitted. First, irrespective of the live exposure rate for items, all items

contained in the test pool must have been subjected to logical review by impacted group members following acceptable procedures.

Given the above, not more than 20 percent of the items in the scoreable pool of items are to be items for which an empirical bias study has yet to be evaluated. It is expected that within three years of the inclusion of an item(s) in the pool that lacks empirical bias evaluation, that the item(s) will be statistically reviewed and a decision made to maintain or release the item from the pool. Gathering data and evaluation of such items is the responsibility of the publisher.

2. Reliability and Errors of Measurement

- a. Reliability. The stability of a placement test should be evaluated. The stability may be assessed by either administering the same test on two occasions (test-retest approach) or administering one form of a test on one occasion and a second form of a test believed to be equivalent on a second occasion (equivalent-forms approach). In order to assess stability, the time between the two testings should be at least two weeks for either approach. The minimum sample size required for an acceptable study is 50 individuals. The resulting correlation coefficients between test scores on two occasions are to be .75 or higher. If subtest scores are used to make placement decisions, the correlations between these occasions for these subtest scores must also be .75 or higher.

In the case of computer administered tests that previously were available in the paper and pencil mode, prior Approval does not automatically transfer to a computer version of that assessment. For a paper/pencil test, for prior Approval status to generalize to the computer version of that assessment, at a minimum an equivalence form reliability evaluation is required. The minimum acceptable coefficient to attain some level of Approval for the computer test is .80. Publishers are not limited to the equivalence forms design. Alternative data gathering and analysis models will be considered. Nevertheless, there must be documented a high level of consistency between scores obtained in the differing formats for Approval to generalize between the assessments. In effect, claims of test equivalence must be supported by evidence.

- b. Standard Errors of Measurement. Standard errors of measurement are to be provided for intervals across the score scale or at likely cut points.

3. Testing Special Groups.

- a. Publishers who seek to have assessment instruments approved for use in the California Community Colleges must agree to provide the test and response forms in alternate media upon request of a college. Testing instruments must be available in a place and manner accessible to

persons with disabilities or offer alternative accessible arrangements for such individuals (i.e., Braille, large print, audiotape, and electronic tests). We believe this is consistent with the intent of Section 36.309 of the American with Disabilities Act and applies to all publishers of testing instruments. Such a commitment is required for a test to be placed on the Chancellor's list of approved tests.

Further, when speededness or amount of time allowed an examinee for testing is a factor associated with performance on a computer test, then the computer assessment system must provide for training and assistance for examinees to assure that the format for testing does not interfere with the achievement estimate for the examinee.

The Manuals made available to test administrators must clearly and explicitly detail instructions and expectations associated with novel formats associated with the assessment (for example, how to save data or re-administer an assessment when there is a power failure, machine lock-up, how to attend to examinees who evidence significant anxiety related to the testing format, etc.).

Administrative Manuals must address and detail the technical properties of computer adaptive tests. Considerations as to interpretation of scores and scoring algorithms, how initiating items are chosen, stopping rules, item exposure control, security and non-traditional administrations, etc. must be formally and completely described. Suitable accommodations for examinees with learning difficulties must be provided for.

SECTION FIVE: SPECIFIC CRITERIA FOR "CRITICAL MASS" APPROVAL OF AN INSTRUMENT

The concept of "critical mass" pertains to situations where evidence on a specific assessment instrument addressing the standards has independently accumulated or been produced by design across several California Community Colleges such that its approval status may be judged to be generalizable to other colleges, and therefore, should be available for their use as a second-party test. The principle of "critical mass" and criteria for its application in attaining approval for a specific instrument are as follows.

Definition of Critical Mass. Any instrument for which evidence has been submitted by a minimum of six colleges from six different Community College districts as a locally managed instrument with approval status (as defined below) on the Chancellor's list shall be available as an approved test on the Chancellor's list for use by other colleges. Such an instrument is to be viewed by other colleges in terms of local responsibilities as a second-party instrument on the Chancellor's list.

1. The concept of critical mass for use with the test instrument approval process allows California Community Colleges to generalize to their own college data compiled by their colleagues in validating the fairness and appropriateness of an assessment instrument without having to replicate all of the data collection normally required for locally managed test instruments.
2. Critical mass applies when there is a sufficient number of colleges who collaborate on their validation efforts and submit the required data as a group. If approval is granted, the instrument will be placed on the Chancellor's Office Approved List as a second-party test. In submitting evidence, the criteria for local management of a test instrument would be followed. Studies would need to be conducted by **each** college for content validity and cut score validity. Group studies or aggregated data could be submitted for test bias evidence, reliability data and standard error of measurement (SEM). SEM could be represented by a single college's data. In addition, empirical item bias evidence is required but may be waived. A consortium of colleges that wishes to follow the critical mass process needs to apply to the Chancellor's Office providing notification of its intent and justification for that approach. In that application the consortium may request a waiver of the empirical item bias requirement.
3. The colleges constituting a critical mass need to be a representative sample of community college student populations (e.g., gender, age, race/ethnicity, etc.). A "sufficient" number of colleges is defined as, at a minimum, six colleges from six different community college districts.

4. Test instruments that are outside of basic skills (i.e., other than reading, writing, math and ESL) can pursue approval from the Chancellor's Office only through the critical mass approach delineated under 2b above. Test instruments, outside of basic skills, which don't have the sufficient number of colleges to constitute a critical mass approval will not be reviewed by the Chancellor's Office and may not be used for placement by colleges (Title 5 §55521).
5. The approval period for the instrument as a second-party test will commence when the test is approved as meeting critical mass criteria.
6. After an instrument has been placed on the Chancellor's list via the critical mass process, it will be considered a second-party instrument and the local colleges' responsibilities will be those listed in Section III criteria I.B. or III.B. of this document.

Information in the matrix on pages 43 to 47 summarizes the specific requirements for each standard that need to be met for each level of "critical mass" approval.

SECTION SIX: SPECIFIC CRITERIA FOR THE "RENEWAL" OF A TEST INSTRUMENT'S APPROVAL STATUS

An assessment instrument is considered approved (initially or following these renewal guidelines) for a six-year period starting with the time at which status in any of the three approval categories is attained. In order for a measure to remain on the Chancellor's list of approved instruments as this time period expires, sufficient evidence addressing the Standards will need to be submitted in advance of the expiration date to allow for the timely renewal of the instrument for an additional six-year period of approved use. That is, instruments nearing completion of six years of "approval" status are required to re-submit information and documentation in advance so that continued use can be maintained by colleges. In an attempt to maintain continuity in the use of second-party instruments, second-party publishers are required to submit renewal evidence at least one year in advance of the approval expiration date for the instrument. While not encouraged to wait until the last moment, colleges may wait to submit renewal information at either of the two review dates during the sixth and final year of approval. Under the renewal provisions that follow, an assessment instrument under "renewal" review may be placed in any one of the three approval categories. If a renewal instrument is probationally or provisionally approved, the timelines for attaining full approval status are the same as for first-time approval requests.

The approval "renewal" process is viewed as a time when instruments, evidence and procedures are to be "re-examined" relative to their appropriateness and continued use for placement in California Community Colleges. This requirement derives from the notion that the collection of evidence and the evaluation of the appropriateness of an instrument's use as a placement tool should be an ongoing and continuous process. As such, each of the standards needs to be re-evaluated by the test publisher, the local college and the MAC Assessment Workgroup.

The extent to which standards (validity, reliability, Cut Scores, test bias and disproportionate impact) are to be addressed was detailed in Sections I, II and III of this document. Information in the matrix (pages 43 to 47) summarizes the specific requirements for each standard that publishers and colleges need to meet for "renewal" approval. Note that when a local college continues to use a second-party approved test, they still have a responsibility to up-date their evidence on content-related validity, cut-score validity and disproportionate impact during the renewal period. This up-dated evidence is to be kept on file for review at the time of the matriculation on-site review visit. Also note that all data supporting the renewal of an instrument must be collected within the last three (3) year period of the initial renewal submission date.

**Matrix of requirements for each standard
that publishers and colleges need to meet for each level of approval**

PROBATIONARY STATUS: First level of approval (good for 2 years from approval)

AREA	1a. PUBLISHER RESPONSIBILITY TO GET TEST ON CO APPROVED LIST	1b. CRITICAL MASS: RESPONSIBILITY of 6 COLLEGES TO GET TEST ON CO APPROVED LIST <i>Each college must do:</i>	3a. CAMPUS RESPONSIBILITY IF USING A 2nd PARTY TEST NOT ON THE CO APPROVED LIST (Locally Managed)	3b. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED OWN TEST (Locally Developed)	4. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED PERFORMANCE ASSESSMENT (e.g., writing sample)
CONTENT VALIDITY	Content-related description of test and provide enough test items so colleges can do content validity study	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Content: Content of Scoring Rubric for the writing prompt by Course Prerequisite Skills
CRITERION OR CONSEQUENTIAL VALIDITY	Supportive Criterion or Consequential Validity Evidence from at least 3 community colleges, the majority of which are from California.				
RELIABILITY					
TEST BIAS	Test bias evidence a. test bias panel <u>and</u> b. empirical	Test bias using panel to judge test items.	Test bias using panel to judge test items. OR	Test bias using panel to judge test items.	Test bias using panel to judge test items.

		<p>OR Cite one or more community colleges with similar demographics who have conducted test bias study</p> <p>OR Cite publisher's test bias studies on students with similar demographics.</p>	<p>Cite one or more community colleges with similar demographics who have conducted test bias study</p> <p>OR Cite publisher's test bias studies on students with similar demographics.</p>		
CUT SCORE VALIDATION	Not applicable (Local college responsibility)	Cut Score rationale - <u>initial</u> setting of cut scores by judgmental or empirical approach	Cut Score rationale – <u>initial</u> setting of cut scores by judgmental or empirical approach	Cut Score rationale - <u>initial</u> setting of cut scores by judgmental or empirical approach	Cut Score rationale - <u>initial</u> setting of cut scores by judgmental or empirical approach
DISPROPORTIONATE IMPACT	Not applicable (Local college responsibility)	Plan to monitor disproportionate impact for various demographic groups	Plan to monitor disproportionate impact for various demographic groups	Plan to monitor disproportionate impact for various demographic groups	Plan to monitor disproportionate impact for various demographic groups
ADA	Alternative versions available	Accommodations provided	Accommodations provided	Accommodations provided	Accommodations provided

PROVISIONAL STATUS: Many to most, but not all standards met (good for one year from approval)

FULL APPROVAL STATUS: All standards met (good for six years from first approval; must be reached by three years from first approval)

AREA	1. PUBLISHER RESPONSIBILITY TO GET TEST ON CO APPROVED LIST	CRITICAL MASS: RESPONSIBILITY of 6 COLLEGES TO GET TEST ON CO APPROVED LIST <i>Each college must do (unless noted)</i>	3a. CAMPUS RESPONSIBILITY IF USING A 2 nd PARTY TEST NOT ON THE CO APPROVED LIST (Locally Managed)	3b. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED OWN TEST (Locally Developed)	4. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED PERFORMANCE ASSESSMENT (e.g., writing sample)
CONTENT VALIDITY	Content-related description of test and provide enough test items so colleges can do content validity study	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Content: Content of Scoring Rubric for the writing prompt by Course Prerequisite Skills
CRITERION OR CONSEQUENTIAL VALIDITY	Supportive Criterion or Consequential Validity Evidence presented from at least 4 community colleges for provisional and 6 for full approval, the majority of which are from California.				
RELIABILITY	Test-Retest Reliability Standard Errors of Measurement (SEM)	May pool data from one or more colleges: Standard Errors of Measurement	Standard Errors of Measurement (SEM) Internal Consistency (.80 minimum) or	Standard Errors of Measurement (SEM) Internal Consistency (.80 minimum) or	Inter-Rater Reliability Minimum .70 correlation

		<p>(SEM)</p> <p>Internal Consistency at each college (.80 minimum) or</p> <p><i>May pool data from one or more colleges:</i> Test-Retest Reliability (.75 minimum) OR Cite test-retest reliability studies from publisher. Minimum sample size of 50</p>	<p>Test-Retest Reliability (.75 minimum) OR Cite test-retest reliability studies from publisher OR Cite the evidence from <u>1 or more</u> community colleges that have done <u>test-retest</u> reliability studies.</p> <p>Minimum sample size of 50</p>	<p>Test-Retest Reliability .75 minimum Minimum sample size of 50</p>	<p>coefficient or 90 percent agreement</p>
--	--	---	---	---	--

PROVISIONAL STATUS: Many to most, but not all standards met (good for one year from approval)
FULL APPROVAL STATUS: All standards met (good for six years from first approval; must be reached by three years from 1st approval)
 (Continued)

AREA	1. PUBLISHER RESPONSIBILITY TO GET TEST ON CO APPROVED LIST	CRITICAL MASS: RESPONSIBILITY of 6 COLLEGES TO GET TEST ON CO APPROVED LIST <i>Each college must do (unless noted)</i>	3a. CAMPUS RESPONSIBILITY IF USING A 2 nd PARTY TEST NOT ON THE CO APPROVED LIST (Locally Managed)	3b. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED OWN TEST (Locally Developed)	4. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED PERFORMANCE ASSESSMENT (e.g., writing sample)
TEST BIAS	Test bias evidence a. test bias panel <u>and</u> b. empirical	Test bias using panel to judge test items. OR Cite one or more community colleges with similar demographics who have conducted test bias study OR Cite publisher's test bias studies on students with similar demographics.	Test bias using panel to judge test items. OR Cite one or more community colleges with similar demographics who have conducted test bias study OR Cite publisher's test bias studies on students with similar demographics..	Test bias evidence a. test bias panel <u>or</u> b. empirical	Test bias evidence a. test bias panel <u>or</u> b. empirical
CUT SCORE VALIDATION	Not applicable	Cut Score rationale: <u>initial</u> setting of cut scores by either judgmental or	Cut Score rationale: <u>initial</u> setting of cut scores by either judgmental or	Cut Score rationale: <u>initial</u> setting of cut scores by either judgmental or	Cut Score rationale: <u>initial</u> setting of cut scores by either

		empirical approach.	empirical approach.	empirical approach.	judgmental or empirical approach.
DISPROPORTIONATE IMPACT	Not applicable	Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.
ADA ACCOMMODATIONS	Alternative versions available	Accommodations provided	Accommodations provided	Accommodations provided	Accommodations provided

RENEWAL: if content of test or course or demographics have not changed (must be renewed at 6 years from initial approval)

AREA	1. PUBLISHER RESPONSIBILITY TO GET TEST ON CO APPROVED LIST <i>(Must provide evidence that they are going to submit evidence for renewal one year prior to the renewal deadline)</i>	CRITICAL MASS: RESPONSIBILITY of 6 COLLEGES TO GET TEST ON CO APPROVED LIST <i>Each college must do (unless noted)</i>	3a. CAMPUS RESPONSIBILITY IF USING A 2nd PARTY TEST NOT ON THE CO APPROVED LIST (Locally Managed)	3b. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED OWN TEST (Locally Developed)	4. CAMPUS RESPONSIBILITY IF CAMPUS DEVELOPED PERFORMANCE ASSESSMENT (e.g., writing sample)
CONTENT VALIDITY	Content-related description of test and provide enough test items so colleges can do content validity	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Item by item analysis comparing each test item to each course prerequisite.	Content: Content of Scoring Rubric for the writing prompt by Course Prerequisite Skills
CRITERION OR CONSEQUENTIAL VALIDITY	Supportive Criterion or Consequential Validity Evidence presented from at least 6 California Community Colleges				
CUT SCORE VALIDATION		Cut score validation will need some empirical evidence (e.g., through criterion or consequential	Cut score validation will need some empirical evidence (e.g., through criterion or consequential	Cut score validation will need some empirical evidence (e.g., through criterion or consequential	Cut score validation will need some empirical evidence (e.g., through criterion or consequential

		validity).	validity).	validity).	validity).
DISPROPORTIONATE IMPACT		Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.	Disproportionate impact monitored for various demographic groups every 3 years.
ADA ACCOMMODATIONS	Alternative versions available	Accommodations provided	Accommodations provided	Accommodations provided	Accommodations provided

**CAMPUS RESPONSIBILITY IF USING TEST ON CCCCO APPROVED LIST
(these materials are to be kept on campus and will be reviewed during each
matriculation site visit)**

AREA	CAMPUS RESPONSIBILITY IF USING TEST ON ON CO APPROVED LIST
CONTENT VALIDITY	Item by item analysis comparing each test item to each course prerequisite: a. For each test item, how many course prerequisites does it measure? b. For each course prerequisite, how many test items measure it? (See II.A.1.a. for details)
CRITERION OR CONSEQUENTIAL VALIDITY	
RELIABILITY	
TEST BIAS	
CUT SCORE VALIDATION	Cut Score rationale – <u>initial</u> setting of cut scores by either judgmental or empirical approach.
DISPROPORTIONATE IMPACT	Disproportionate Impact monitored for various demographic groups every 3 years.
ADA ACCOMMODATIONS	Accommodations provided

SECTION SEVEN: THE PROCESS FOR REVIEWING ASSESSMENTS USED IN THE CALIFORNIA COMMUNITY COLLEGE SYSTEM

The review of instruments to be used on California Community College campuses is a responsibility shared among the Chancellor's Office, the colleges and districts, test developers, and users of the particular test, as well as agents or agencies contracted to provide specific review and evaluation services. In this section the steps are presented that identify the process by which assessment instruments are reviewed. The responsibilities of the various parties are discussed.

The chart below summarizes timelines and reporting/submission requirements for colleges and second party publishers.

Reporting Requirements and Timeline

<u>Type of Test</u>	<u>Initial Submission</u>	<u>Renewal Requirements</u>
Second-Party Publishers	As available, submit request and evidence to California Community Colleges Chancellor's Office.	Renewal request and evidence is to be submitted no later than five years from the time an instrument was initially placed on the Chancellor's List of approved tests.
Colleges using approved Second-Party instruments	Conduct studies and maintain files, documentation and reports at the local college. Materials reviewed during matriculation site visits.	Up-date evidence as needed for examination during matriculation site visit 6-year review cycle.
Locally managed or developed test	Submit required documentation to California Community Colleges Chancellor's Office.	Renewal request and evidence is to be submitted no later than six years from the time an instrument was initially placed on the Chancellor's List of approved tests.

March, 2001

The Chancellor of the California Community Colleges is vested with the responsibility to allow or disallow use of any instrument on the campuses. Procedurally, the Chancellor, in coming to a specific decision, seeks the advice of the Matriculation Advisory Committee (MAC) and this group relies on its Assessment Work Group to guide MAC's recommendations to the Chancellor. In addition, the Chancellor's Office assumes responsibility to communicate instrument evaluation results to the appropriate testing agency responsible for the production and distribution of the instrument.

The evaluation results are to be based on those professional standards that guide educational and psychological testing, and on the requirements identified in AB3 and the Title 5 matriculation regulations. While the instrument review steps are discussed in terms of activities leading to a recommendation to the Chancellor, the process can also be adopted by a local district for arriving at its own decisions regarding a particular assessment instrument or process. The approach and procedures for the review of instruments are as follows.

Step 1. Compile Information on Assessment Instruments.

The quality of the recommendation made by the MAC to the Chancellor depends upon the quality of the information available to it. Consequently, as much information as possible should be available to the committee, including but not limited to the test and its manuals, technical reports, and for second party tests reviews such as those found in the Mental Measurement Yearbooks, articles that review a test published in professional journals and books, and technical reports prepared by California Community College users. Assembling this information is the responsibility of the psychometric experts contracted by the Chancellor's Office to serve as consultants to this process. Test developers are to be contacted and asked to provide copies of tests, test documents and related technical reports.

Although much of the documentation on a measure is likely to be supplied by the instrument's developer and made available in a manual, in time the user has the responsibility for supplying information that indicates that a measure is being used appropriately.

Step 2. Perform Psychometric Expert Review.

The information gathered in Step 1 will be reviewed by at least two psychometric experts. The psychometric experts must have received doctorates in a measurement-related area or have had five or more years experience in an occupation requiring expertise in tests and measurement. They must have a broad understanding of both theoretical and applied issues associated with testing so that they can make informed recommendations to the MAC Assessment Work Group. These experts will evaluate an instrument based primarily on criteria presented in the Standards for Educational and Psychological Testing, as well as the evaluation criteria that have been prepared specifically for the evaluation of instruments for the California Community Colleges. Reviewers may also use other guidelines that are commonly accepted by the psychometric community such as the Code of Fair Testing Practices in Education. In addition, the measure must be reviewed to ensure its compliance with the matriculation regulations. Selecting and defining the workscope of the psychometric experts is a responsibility of the Chancellor's Office and the MAC Assessment Work Group.

Step 3. Perform Content Expert Review (for second-party tests at the time of initial review).

For Second-Party tests, the test documentation gathered in Step 1 will also be reviewed by at least two subject matter content experts. (Note: This External Review will be conducted for second-party tests at the time of "renewal" evaluation also.) For example, if a measure purports to assess "Preparedness for Calculus," individuals who understand the preparatory information for learning calculus and the information that is presented in a calculus course would be

solicited to review the information concerning the test as well as the test itself. These reviewers will be selected from lists of potential content experts recommended by members of the MAC Assessment Work Group as being knowledgeable about specific content area courses for which an assessment instrument is presumed to have value. To the extent possible, content expert reviewers will be chosen to reflect the diversity of experiences with a given instrument in the colleges. The content reviewers' primary focus is to evaluate: 1) the match among the rationale underlying the measure as stated in the manual or by the user, the items on the measure, and the suggested interpretation of scores relative to the test's intended use, and 2) the content appropriateness for the diverse populations served in the California Community Colleges. The content experts will file a written report with the psychometric experts so that they may have as complete an understanding of the quality of the measure as possible. A review form will be utilized to guide and standardize the content expert's review.

Step 4. Perform Matriculation Advisory Committee (MAC) Assessment Work Group Review.

The MAC Assessment Work Group consists of individuals who work within the California Community Colleges. Its charge is to serve in an advisory capacity to the psychometric experts, the Matriculation Advisory Committee and the Chancellor's staff. The members include a cross-section of individuals who have expertise in assessment, research and evaluation or testing, are responsible for administering, scoring and interpreting tests, are faculty members of community colleges, or are community college administrators. They also serve as liaisons with the field. Once the procedures for the instrument review have been finalized, the Work Group is consulted and informed regarding the actual evaluation of instruments. The psychometric experts present their findings and judgments to the Work Group. Members of the Work Group are asked to review the experts' evaluations and to offer their opinions of the evaluations. Since the members have an understanding of how assessment measures are used within the community colleges, they are able to give feedback about the evaluation of an assessment measure as it applies to decisions made at the community colleges. The Work Group may solicit additional information from test developers or test users, if necessary.

Step 5. Generate Recommendations.

For each test evaluated, a written report is to be filed by the psychometric experts with the Chancellor's Office. Prior to filing a recommendation with the Chancellor regarding the first review of a specific test, a Preliminary Test Evaluation will be shared with the test publisher. The preliminary filing with the test producer will give the developer the opportunity to be informed about the report and to respond to it. A period of fourteen days following the preliminary filing will be allowed for the test producer to provide additional information that might lead to modification of the recommendation report. The interim period is

not planned for the developer to assemble, analyze and report on "new" data gathered in response to the preliminary findings; rather, the interval provides an opportunity for the test developer to supply information that may be already available but not previously provided for the review.

The report filed by the psychometric experts includes a summary of the opinions of the content experts, the MAC Assessment Work Group, and available information gleaned from external test reviews, test publishers, and the psychometric experts for the measure. The report offers a specific recommendation concerning the use of the measure. The recommendation will be in one of four categories: (A1) Full Approval; (A2) Provisional Approval; (A3) Probationary Approval; or, (B) Not Approved. The first three categories are intended to communicate different levels of approval with different consequences attached to the recommendation. Only instruments placed in one of the three "A" categories will be available for use by all California Community Colleges (second-party tests) or by a specific college (locally managed or developed tests).

The length of time an instrument will be available for use by the colleges without submission of additional information varies by specific approval category. However, an instrument may only maintain standing in Provisional Approval (A2) and a Probationary Approval (A3) for a period of time not to exceed three (3) years. That is, an instrument will not be permitted to remain approved for use without attaining Full Approval (A1) within three years. Further, once any approval status is attained for an instrument, that instrument will be treated as "approved" for a period not to exceed six (6) years. After this six-year maximum tenure interval, unless new supporting materials or documentation have been submitted and favorably reviewed for continued use in the California Community Colleges (i.e., upon renewal re-evaluation the instrument attains one of the three "approval" ratings), the instrument will revert automatically to the "Not Approved" status (B). During the period of submission toward Full Approval, new evidence to support the appropriateness and suitability of an instrument may be submitted at any time to the Chancellor's office for review and consideration; however, materials submission is expected to provide sufficient information to allow for review and consideration into at least the next highest approval classification. Instruments nearing completion of six years of "approval" status are encouraged to re-submit information and documentation during the fifth year so that continued use can be maintained by colleges (that is, initiating the Renewal process). The prior section presents information and the needed documentation required of colleges and publishers desiring to renew the approval status of an instrument.

The intended implications for the use of an instrument when placed in a specific evaluative category for initial submission or renewal are as follows.

A1 Full Approval - Instruments in this category meet the **Standards'** criteria.

The available evidence indicates these instruments have potential value when used to serve a specific assessment function in California Community Colleges. These instruments have high probability of yielding test scores useful in assisting decision making for a particular community college student.

A2 Provisional Approval - Instruments in this category meet relevant standards and criteria, but lack sufficient or recent information to assign the unequivocal "Full Approval" rating. Some criteria were not met because (a) documentation was lacking that in all likelihood should be provided in a relatively short time period or (b) criteria were recently introduced and the user/developer should be allowed time to meet them. In this conditional category, the expectancy is that the test, in time, will achieve a "Full Approval" recommendation. As such, the necessary clarifying information to receive a "Full Approval" evaluation on instruments for this "Provisional" category is expected to be provided in due course. Recommendation in this category means that the test developer or user must supply within one academic year the specified additional clarifying information. Failure to submit the required clarification(s) within one year will result in reclassification into the "Probationary" category.

A3 Probationary Approval - Instruments in this category are missing critical information and thus a clear-cut recommendation cannot be established; or from the information that is available, deficiencies are noted. The intended purpose for use of these instruments is clearly stated and some positive information supporting its use is available, but the necessary evidence available for a final judgment is incomplete. For tests once classified as "Probationary," additional data collection must be provided for further evaluation. Such instruments can only be maintained in the "Probationary Approval" designation for a maximum of two academic years.

B Not Approved - Instruments in this category are those for which the evidence indicates that they have failed to meet one or more of the standards or criteria considered essential by the reviewing bodies or have failed to meet a condition of AB 3 or Title 5. What is considered an essential element is likely to vary among applications (that is, tests can fail to be approved for differing reasons), but the specific deficiency will be identified in the report of the MAC Assessment Work Group to the Chancellor.

March, 2001

Validation and documentation of instrument quality is considered an ongoing process and thus publishers and colleges are expected to engage in continuous evaluation and monitoring of their instruments. In this spirit, periodic reviews by the Chancellor's office of instruments previously approved at the request of managers or users of instruments is both acceptable and encouraged in anticipation of the need to maintain a continuous approval status for a device.

Step 6. Disseminate Chancellor's Decision.

The Chancellor will make a decision concerning the use of each measure reviewed. Tests that are not on the Chancellor's List of approved tests must not be used in the placement process in the California Community Colleges system. Approval/disapproval decisions will be communicated formally to both the community colleges and the specific test publishers involved.

Step 7. Allow for an Appeals Process.

A decision by the Chancellor may be appealed (i.e., subject to formal reconsideration) by any individual, college, district, agency or entity. Requests for an appeal are to be submitted to the Chancellor within 30 days of the Chancellor's decision. The request must clearly explain why the decision is being questioned. The Chancellor's prerogative is to determine the next course of action, although one might expect that an Appeals Committee will be called upon to reconsider the standing recommendation. Appeals are to be acted upon, and a recommendation forthcoming from the Chancellor, within six months of the Appeal request.

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.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1988). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978). Adoption by four agencies of the Uniform Federal Guidelines on Employee Selection Procedures. *Federal Register*, 43, 38290-38315.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
- Jaeger, R. M. (1989). Certification of student competence. In Linn, R. L. (Ed.) *Educational measurement*. New York: American Council on Education and Macmillan Publishing Company.

GLOSSARY

- Consequential-related validity** evidence addressing desired or undesired outcomes that follow from the use of test scores to advise placement of students into courses.
- Content-related validity** evidence addressing the extent to which course pre-requisite knowledges and skills are being measured by the items on a test for all courses into which the test scores are being used to place students
- Corrected validity coefficients** psychometric procedures that estimate the relationship between two sets of scores if the test scores were measured with perfect reliability (corrected for attenuation) or full variability (corrected for restriction of range).
- Correlation coefficient** a statistical index that summarizes the magnitude of the relationship between two sets of scores for the same group of individuals. This index takes on values ranging from -1.00 to 1.00 with values around zero ($.00$) representing no relationship.
- Criterion-related validity** evidence addressing the extent to which scores on the placement test are related to scores on an appropriate criterion measure of student ability to meet different course requirements into which the students are being placed or an appropriate measure of student success in different courses.
- Critical mass** the accumulation of evidence across a diverse set of colleges which can be used to gain approval for the use of a test instrument by all colleges in the system.
- Differential prediction** evidence addressing the extent to which scores on a placement test are equally predictive of an outcome measure for all subgroup classifications, e. g., gender, ethnicity, age, etc.
- Direct performance assessments** that require an open-ended response from the test taker to a task, set of tasks or set of defined stimulus conditions. Responses then are scored using a standardized scoring rubric that has defined scale values indicating the adequacy of performance at different levels of proficiency.
- Empirical approach to setting cut-scores** procedures to identify cut-score values based on differential test taker test performance under certain design conditions.

Internal consistency a method of estimating test score reliability based on the consistency or relationship of responses to test items across test takers for a single administration of the test. Examples of methods or indices include Kuder-Richardson formula 20 or 21, coefficient alpha and split-half procedures.

Interscorer reliability coefficient an index of reliability indicating the consistency of ratings assigned to test taker responses (usually from performance assessment data) by two or more raters.

Judgmental approach to setting cut-scores procedures to identify cut-score values based on expert panel review, evaluation and judgments about the appropriateness and difficulty of test and test item content, and expected performance for identified populations of test takers.

Norms reported score distributional characteristics for samples of test takers that are intended to represent a population of test takers with described characteristics such that the performance of the norm group can offer relative interpretation of a person's test score with reference to the performance of test takers in the norm group.

Reliability evidence addressing the degree of consistency of measurements when the procedures producing test scores are repeated on a population of individuals or groups.

Stability coefficient an estimate of the reliability of test scores using a procedure requiring that data be collected from the same group of individuals on two separate occasions with an intervening period of at least two weeks between administrations.

Standard error of measurement an index related to the reliability of test scores which provides information addressing the degree of inaccuracy for specific test score values.

Transformed scale scores scores that are reported on a scale other than that produced by raw scores, e.g., percentile ranks or scores reported on a scale with a different mean and standard deviation than those of the raw scores.

Validity evidence addressing the extent to which the interpretation of scores from a test is meaningful, appropriate and useful to serve the purpose of placement of students into different courses.