ED 230 589	TM 830 369
TITLE	Guidelines for Proficiency Tests.
INSTITUTION	California State Dept. of Education, Sacramento. Office of Program Evaluation and Research.
PUB DATE	82
NOTE	67p.
AVAILABLE FROM	Publication Sales, California State Department of Education, P.O. Box 271, Sacramento, CA 95802 ($2.00)
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Check Lists; Competency Based Education; *Criterion Referenced Tests; Educational Legislation; Graduation Requirements; Guidelines; High Schools; *Minimum Competency Testing; Research Methodology; *School Districts; *Test Construction; Test Results; Test Use; Test Validity
IDENTIFIERS	California State Department of Education

ABSTRACT

Guidelines are presented for use by school personnel in reviewing and improving locally developed proficiency tests used in meeting the requirements of the California Pupil Proficiency Law. The guide is organized in three main chapters on test construction, test validation, and test documentation. The test construction chapter focuses on issues that should be addressed in the construction of a high-quality proficiency test. The test validation chapter covers psychometric indexes of test quality. Other procedures and indexes related to technical quality are also covered. The chapter on test documentation deals with the administration of proficiency tests and the reporting of proficiency test information. Also included is information on describing the tests to students who will take the tests and to their parents. Throughout this document questions are posed in the page margins to stimulate reader inquiries regarding the completeness and quality of proficiency tests. These questions are repeated in checklist form at the end of the book for quick reference. In addition, materials for further reference are cited at the end of each section. (PN)

Guidelines for Proficiency Tests

CALIFORNIA STATE DEPARTMENT OF EDUCATION
Wilson Riles-Superintendent of Public Instruction
Sacramento, 1982

Guidelines for Proficiency Tests

Prepared under the direction of the
Office of Program Evaluation and Research

This publication was edited by Bob Klingensmith, Publications Consultant, prepared for photo-offset production by the staff of the Bureau of Publications, and published by the California State Department of Education, 721 Capitol Mall, Sacramento, CA 95814. The original artwork for this publication was prepared by Cheryl Shawver McDonald, and the type was set by Anna Boyd and Leatrice Shimabukuro.

Printed by the Office of State Printing
and distributed under provisions of the
Library Distribution Act

1982

Copies of this publication are available for $2 each, plus sales tax for California residents, from:

Publications Sales
California State Department of Education
P.O. Box 271
Sacramento, CA 95802

A partial list of publications on testing, evaluation, and proficiency available from the Department of Education appears below. A list of all publications available from the Department can be obtained by writing to Publications Sales at the address listed above.

Measuring School Quality (brochure) (1981)	NC
Proficiency Assessment in California: A Status Report (1980)	2.00
Proficiency Skill Development Kit (1980)	7.50
Report of Selected ECE Schools' Increasing, Decreasing Reading Scores (1980)	1.50
Resource Catalog for Proficiency Assessment (1978)	.85
Sample Assessment Exercises, Proficiency Assessment, Grades 4, 6 (2 vol.) (1978)	54.00
Sample Assessment Exercises, Proficiency Assessment, Grades 7, 9, 11 (2 vol.) (1978)	54.00
Statewide Summary of Pupil Performance on School District Proficiency Assessments (1982)	NC
Student Achievement in California Schools: 1981-82 Annual Report (1982)	2.00
Survey of Basic Skills: Grade 3—Rationale and Content (1980)	1.50
Survey of Basic Skills: Grade 6—Rationale and Content (1982)	2.00
Technical Assistance Guide for Proficiency Assessment (1979)	7.50
Test Content Specifications, Basic Skills, Grades 6, 12, Mathematics (1975)	.85
Test Content Specifications, Basic Skills, Grades 6, 12, Written Expression, Spelling	.85
Test Content Specifications, State Reading Tests, Grades 2, 3, 6, 12 (1975)	.85

Contents

Preface v
A Note to the Reader vii

Introduction 1

Test Construction 3
Developing Proficiency Standards 3
Developing Item Specifications 7
Writing Items 10
Pretesting and Revising Items 12
Addressing Other Factors in the Test Construction Process 14

Test Validation 20
Conducting Item Analyses 20
Setting Passing Scores 26
Reviewing the Test for Bias 30
Assessing the Validity of the Test 34
Assessing the Reliability of the Test 36

Test Documentation 40
Providing a Test Description for Lay Audiences 42
Documenting the Test Administration and Scoring Procedures 43
Documenting the Test Construction Process 44
Documenting the Test Validation Process 45

Glossary 48

Guidelines Checklist 49

User Questionnaire 59

Preface

The California Pupil Proficiency Law (Education Code sections 51215–51218) requires that all students demonstrate proficiency in the basic skills prior to graduation from high school. Although the law is quite flexible with regard to measurement options, common sense and fair practice dictate that proficiency tests be psychometrically sound. High quality proficiency tests are essential, given the importance of the test results in determining whether or not California students ultimately graduate from high school.

The testing technology for criterion-referenced tests is still in its infancy. While the theoretical and technical developments have neither the power nor the sophistication of classical test theory or item response theory, the current technology for criterion-referenced tests does have some indicators of test quality. Districts interested in reviewing and refining proficiency tests will want to employ this methodology to upgrade their tests. *Guidelines for Proficiency Tests* represents the Department of Education's perspective on the *minimum* technical requirements for proficiency assessment instruments.

The guidelines are organized around the processes of test construction, validation, and documentation. Our recommended procedures are easy for school personnel to implement, yet rigorous enough that those who use tests developed in accordance with these procedures can be confident that the results of the tests will be accurate.

As important as the psychometric qualities of proficiency tests is the manner in which the test results are used. We see the concepts of test quality and test use as highly interrelated. By using the *Guidelines for Proficiency Tests,* testing specialists can ensure the technical quality of the testing instruments and the accuracy of the test results. But we ask that county and district staff look beyond the quality of proficiency tests and examine the use of such tests in light of the local curriculum and the interests of the local community.

If proficiency tests match the local curriculum, the tests gain both validity and utility. If proficiency tests accurately measure local standards for graduation, the tests have legitimacy, and the high school diploma gains respectability.

The development and uses of proficiency tests require much thought by district staff and community representatives. We hope that *Guidelines for Proficiency Tests* will enable testing and curriculum personnel to reconsider, and perhaps revise, earlier decisions on proficiency testing for students in California public schools.

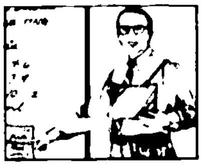
DONALD R. MCKINLEY
Chief Deputy Superintendent
of Public Instruction

ALEXANDER I. LAW
Chief, Office of Program
Evaluation and Research

A Note to the Reader

Although it is recognized that most districts have already developed their proficiency tests, the guidelines in this document are presented basically in procedural sequence. In this way they can be readily used to review and refine existing instruments or to develop new ones.

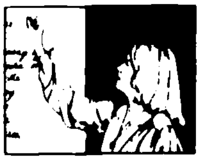
The illustrations and examples provided throughout this document are designed to highlight the steps that school districts should follow in revising or developing proficiency tests. The characters shown below, from the fictitious San Tomas Unified School District, are those that appear throughout the illustrations. They are identified here to help the reader better understand the roles played by various individuals in the revision or development process and to emphasize the need for participation by such individuals in these processes.



Associate Superintendent for Curriculum, in charge of proficiency assessment

Counselor

Teacher



Teacher

Teacher

Parent

Introduction

This document contains guidelines for use by school personnel in reviewing and improving locally developed proficiency tests used in meeting the requirements of the California Pupil Proficiency Law (Education Code sections 51215—51218).

The original legislation for proficiency assessment, Assembly Bill 3408 (Chapter 856, Statutes of 1976), was quite flexible with regard to testing options. In fact, the broader term *assessment*—rather than *test*—was used throughout the law. One of the few technical specifications in the law mandated criterion-referenced (as opposed to norm-referenced) test score interpretations. In *Proficiency Assessment in California: 1980 Status Report on Implementation of California's Pupil Proficiency Law*, the Department of Education reported that the predonderance of existing proficiency tests were objective, paper-and-pencil instruments. In some cases commercially published tests were being used, and occasionally performance tests or subjective assessments were being given. Fully 80 percent of the proficiency tests being used in California were locally developed, criterion-referenced tests.

This document is purposefully brief so that school personnel can easily identify the strengths and weaknesses of existing, locally developed tests and then go about the task of improving them. The Department's Office of Program Evaluation and Research (OPER) has also developed other mechanisms for test review and refinement, including the *Handbook for Proficiency Assessment* and the Proficiency Assessment Training Network. The handbook is an instructional manual that includes in-depth coverage of a variety of test development topics ranging from setting passing scores to scoring writing samples. It is much more "how-to" oriented than this document. Parties interested in obtaining a copy of the handbook should contact the Proficiency Assessment Team at the Office of Program Evaluation and Research, 721 Capitol Mall, Sacramento, CA 95814 (916-445-0297).

The Proficiency Assessment Training Network consists of school district personnel and office of county superintendent of schools personnel trained in numerous areas of test development and refinement. Consultative assistance is available on an on-call basis from network members. This assistance covers both psychometric and curricular issues related to proficiency assessment and basic skills instruction. Access to the Proficiency Assessment Training Network is also available through OPER's Proficiency Assessment Team.

Guidelines for Proficiency Tests is organized in three main chapters on test construction, test validation, and test documentation. The test construction chapter focuses on issues that should be addressed in the construction of a high-quality proficiency test. Although most districts have already developed proficiency tests, the techniques presented in this chapter may be useful in any examination or revision of established proficiency tests. For example, in 1980 it was learned that almost two-thirds of locally developed proficiency tests were not constructed from rigorous item specifications. Districts that skipped this step in the development process should review the appropriate information herein and rework their testing instruments.

The test validation chapter covers psychometric indexes of test quality. In order to have faith in the decisions based on proficiency tests (whether to provide remedial study for students or to graduate them), district staff should assess the reliability and validity of their assessment measures. Other procedures and indexes related to technical quality are also covered in this chapter.

The chapter on test documentation deals with the administration of proficiency tests and the reporting of proficiency test information. The focus of this chapter is on how to communicate accurately the intent, content, and results of proficiency testing. Also included is information on describing the tests to students who will take the tests and to their parents.

Throughout this document questions are posed in the page margins to stimulate reader inquiries regarding the completeness and quality of proficiency tests. These questions are repeated in checklist form at the end of the book for quick reference. (The pages are perforated for easy tear out.) In addition, materials for further reference are cited at the end of each section.

This publication can be used in many ways. At a minimum personnel responsible for developing proficiency tests should review the checklist at the back of the book to ensure that each guideline has been considered. It is not essential that all questions be answered in the affirmative. But for those questions answered negatively, rationale should be established for omission or substitution. For example, if a district did not conduct a statistical test for bias (because, perhaps, there were too few minority students to be statistically significant), this fact should be documented, and a subjective bias review should be substituted.

It is important to realize that these guidelines are not ironclad. They can be implemented in numerous ways; where possible, various options are recognized, and priorities are indicated or recommendations are made for their use. Since the Pupil Proficiency Law is flexible with regard to measurement options, it is difficult to cite any of the guidelines as being relevant for all proficiency tests. Nevertheless, the flexibility in the legislation was intended to give districts control of test content, not to permit serious variations in the technical adequacy of tests. The measurement techniques set forth herein are fairly well agreed upon and are generally applicable.

Test Construction

The test construction process involves four major steps: (1) developing proficiency standards; (2) developing item specifications; (3) writing test items; and (4) pretesting and revising items.

Although the test construction guidelines in this publication are primarily intended for use by districts that constructed their own proficiency tests, districts using commercially published tests or existing item pools from other districts may also benefit from the information on the development of proficiency standards, item review, and pretesting of items. (See *Handbook for Proficiency Assessment*, Section 111, pp. 95-118.)

The test construction process involves much more than simply writing items and assembling them into a test. This process is only part of the larger test construction plan that each district should have developed. The features of the test construction plan include:

1. Identifying the purposes and uses of the test These may include certification of secondary students for graduation, identification of students for remediation, and identification of gaps in the curriculum or instruction.

   Have the uses of the proficiency tests been identified and agreed upon?

2. Determining whom to test The law states that students must be tested at least once in grades four through six, at least once in grades seven through nine, and at least twice in grades ten through eleven. Districts must still decide which students to test and in which grades.

   Have procedures been established for testing at the grade levels specified by law?

3. Determining when and how often to test Districts must decide the time of year, the day of the week, and even the time of day to assess students' proficiency. They should also establish a policy on retesting. (See *Handbook for Proficiency Assessment*, Section I, pp. 37-46.)

   Have policies been established with regard to schedules for testing and retesting?

4. Assessing available resources and practical constraints Resources and constraints include the time available for test construction and validation; available personnel with specific areas of expertise within and outside the district; and money, facilities, and equipment.

   Have district staff identified local resources for use in the various proficiency assessment activities?

Developing Proficiency Standards

The test construction process should begin with the development of proficiency standards by the district and the community. Proficiency standards should describe the skills students are expected to demonstrate, the methods to be used to assess skill acquisition, and the level of performance at which the students are expected to perform at the time of testing. These skills include those in the required testing areas--reading comprehension, writing, and computation—and they may also include more general "life skills." Since proficiency standards are developed locally, they should reflect the community's commitment to local control and the local curriculum.

Community Involvement

Has the local community been involved in developing proficiency standards?

The law mandates that parents, administrators, teachers, and counselors be involved in the development of proficiency standards. Students must also be involved in developing standards for secondary schools. The type and extent of community participation is left up to each district. Community involvement may range from simply responding to questionnaires prepared by the district to participating on ongoing advisory committees. Districts should try to involve community members as much as possible in the development and periodic reexamination of proficiency standards.

Has community involvement in proficiency assessment reflected the demographic makeup of the district?

Community members participating in the development of proficiency standards should be representative of the community as a whole with regard to socioeconomic level, sex, age, and ethnicity. Students involved in developing standards for secondary schools should be representative of the students who will take the tests. A representative sample of community members should be surveyed at least once for their opinions regarding proficiency standards.

Articulation Between Schools

Have the proficiency standards of elementary and secondary schools been articulated?

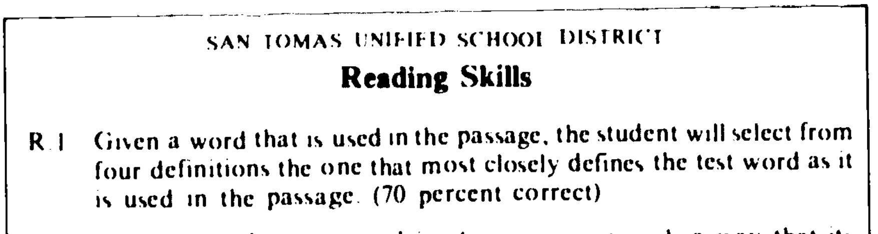
To help ensure continuity between the skills taught in the elementary and secondary schools, the law requires that educators from both elementary and secondary schools work together to ensure that the proficiency standards adopted for elementary schools in the district are consistent with those adopted for secondary schools. Representatives of elementary schools should work with those developing proficiency standards for secondary schools, and secondary school personnel should be familiar with the proficiency standards established for the elementary school students who will eventually enter their schools. Continual dialogue among representatives from both levels will maintain articulation and foster exchange of information about instructional methods and curricular materials.

Format of Proficiency Standards

Each proficiency standard should consist of three parts: (1) a statement of the skill the student should be able to demonstrate; (2) the conditions under which the student should be able to perform the skill; and (3) the acceptable level of student performance in demonstrating acquisition of the skill (see Fig. 1). To the extent possible, the same format and style should be used for proficiency standards in the three content areas (reading comprehension, writing, and computation). This makes communication with lay audiences easier and shows articulation among subject-matter specialists. In the case of writing skills, however, it may be difficult to use the same format, because writing standards are often more global than discrete math or reading skills.

Do proficiency standards include a statement of the skill being assessed, how the skill will be assessed, and the level of performance required?

Have a similar style and format been utilized for the proficiency standards in the three required content areas (reading comprehension, writing, and computation)?



SAN TOMAS UNIFIED SCHOOL DISTRICT

Reading Skills

R.1 Given a word that is used in the passage, the student will select from four definitions the one that most closely defines the test word as it is used in the passage. (70 percent correct)

R.2 Given a word that is used in the passage in such a way that its meaning can be inferred from context and that is at least three grade levels above the readability level of the passage, the student will select from four options the one that most closely defines the test word as it is used in the passage. (75 percent correct)

R.3 Given a statement or question derived from two or three sentences within the passage, the student will select from four options the one that completes the statement or answers the question correctly. (60 percent correct)

R.4 Given a question regarding the sequence of various elements within the passage, the student will select from four options the one that answers the question correctly. (60 percent correct)

R.5 Given a question or statement regarding a cause-and-effect relationship within the passage, the student will select from four options the one that correctly relates the cause and effect. (60 percent correct)

R.6 Given a statement regarding what the passage is mostly about, the student will select from four options the one that identifies the main idea of the passage. (75 percent correct)

Fig. 1. Sample proficiency standards

The skills identified in proficiency standards should be broad enough to cover desired aspects of the proficiency, but they should not be so broad as to encompass all possible skills. One reliable gauge of skill breadth is the amount of instruction devoted to the skill. For example, in the area of writing, "proper use of the semicolon" may require only two or three days of instruction, while "writing compositions" may demand two or three semesters. A more reasonable skill might be "developing the topic sentence," which might require three or four weeks of instruction.

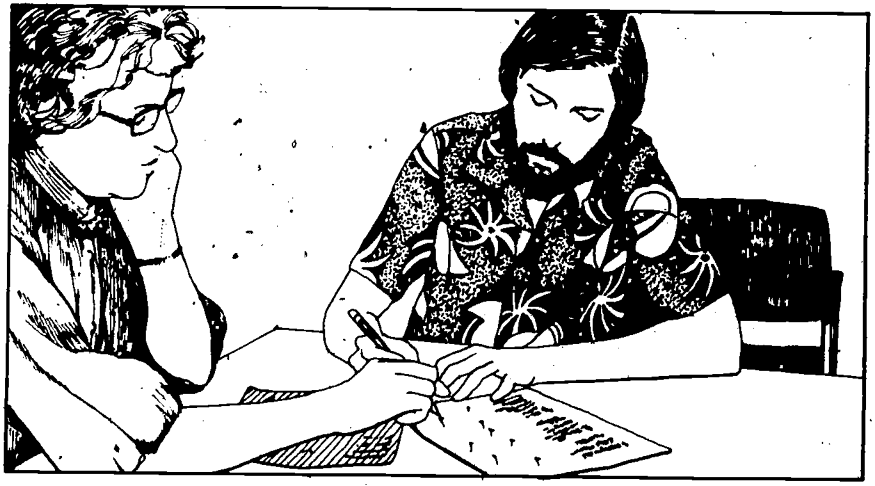
The conditions of performance stated in each proficiency standard should reflect both the conditions under which the skill was taught and the conditions under which skill development will be assessed. The stated acceptable level of performance should identify the *minimal* level of performance necessary for mastery of the skill in the community. (See p. 26 for information on the process to be used in establishing performance levels or passing scores.)

Review of Proficiency Standards

Have proficiency standards been reviewed periodically for curricular validity and instructional validity?

Proficiency standards should be reviewed periodically for curricular and instructional validity. Curricular validity is the extent to which the skills identified in the proficiency standards are consistent with the stated curricular objectives. Instructional validity pertains to the extent to which students have been provided instruction in, or have had an opportunity to learn, the identified skills. Obviously, students should not be tested on skills or material that they have not been taught.

A committee composed of school administrators, curriculum specialists, teachers, and community members should review and approve the completed set of proficiency standards before any other major steps in the test construction process are undertaken. Each standard



should be checked for consistency with the content of curriculum materials. Likewise, steps should be taken to ensure that the proficiency skills are being taught. (Reviewing lesson plans and making classroom observations are but two ways of making instructional validity checks.) Proficiency standards of doubtful curricular or instructional validity should be revised or discarded.

References

Gagne, R. M., and L. J. Briggs. *Principles of Instructional Design* (Second edition). New York: Holt, Rinehart and Winston, Inc., 1974, pp. 45—135.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section I, pp. 1—6, 11—16, and 49—53.

Developing Item Specifications

Have item specifications been used in the development of test items?

Once proficiency standards are established, a set of item specifications, or "blueprints," for writing the test items should be developed (see Fig. 2). The careful development of item specifications is important for at least three reasons. First, item specifications provide a set of rules to guide item writers. This may help a group of writers to produce a consistent set of items for each skill being assessed. It must be remembered that item specifications allow thorough domain description, which is the defining feature of criterion-referenced tests (and, by extension, proficiency tests). Second, item specifications provide the basis for interpretation of test results; that is, for mastery/

SAN TOMAS UNIFIED SCHOOL DISTRICT

Item Specification—Reading

Skill R.2 The student will demonstrate the ability to determine word meanings from context.

Performance mode: Given a word that is used in the passage in such a way that its meaning can be inferred from context and that is designated as being at least three grade levels above the readability level of the passage, the student will select from four options the one that most closely defines the test word as it is used in the passage.

Item stem characteristics: The test word will be underlined in the passage. The item stem will direct the student to use the passage to identify the meaning of the word and will consist of words designated at a grade level equal to or lower than the readability level of the passage.

Distracter characteristics: Distracters will be definitions consisting of words used in the passage; or, if necessary, they may consist of other words designated at a grade level equal to or lower than the readability level of the passage. Each distracter will be as grammatically parallel to the correct response as possible.

Sample item:

Read the following story, and answer the question.

It was a perfect night for a barbecue. The day had been hot, but it had cooled off to the point where all the kids were in sweaters and sweatshirts. The long days of competition created a real hunger for the team. Just as the sun was setting with a rosy glow, the charcoal briquets gave off the same, soft color. Just around the hottest part of the fire, there were grey talcumy ashes, and we all knew it was time to cook.

R.2.6 You can tell from the story that "talcumy" means:

A. rocky B. powdery C. frosty D. sticky

Fig. 2. Sample reading skill item specification

nonmastery judgments. Third, item specifications communicate what will be assessed for students (so that they can prepare) and for teachers (so that they can target initial and remedial instruction).

Developing item specifications is a process that need involve only school personnel (specifically, subject-matter and testing specialists). The content specialists translate the broad skills identified in the proficiency standards into smaller, more measurable skill components. The testing specialists ensure that the items for each skill are psychometrically sound. Consultative assistance from county personnel (for example, Proficiency Assessment Network trainers), state personnel, (for example, OPER personnel providing training in the use of the *Sample Assessment Exercises Manual*), or university personnel may be helpful in developing item specifications.

Format of Item Specifications

Each item specification should contain four parts: (1) a general description of the skill being assessed; (2) the item stem characteristics; (3) the distracter characteristics; and (4) a sample item.

The general description should identify the skill to be performed and the performance mode, or the manner in which skill acquisition will be tested. Often, the skills identified in proficiency standards are too broad to be measured precisely and need to be broken down into subskills. But there is a trade-off between skills that are too broad to be measurable and skills that are so specific that they are trivial, resulting in a test of unwieldy length. Content specialists developing item specifications should identify skills at a level of specificity that allows precise and meaningful measurement. The skills should not be so specific that (1) the test results would be uninterpretable or trivial; or (2) the test would be overly long if a sample of representative skills for each content area were included.

Do item specifications include descriptions of the manner in which each skill is to be assessed (i.e., performance mode)?

The performance mode identified for each skill should match the manner in which instruction in that skill was provided and should simulate the way the skill will actually be used in school or life situations.

Item stem characteristics set limits on the stimulus portion of each item. The item stem presents the problem to be solved or question to be answered by the student. Item stem characteristics should include (1) a description or list of acceptable content; (2) the readability level at which the item should be written; and (3) the expected difficulty level in terms of p-values (see p. 23) or grade levels.

Are item stem characteristics included in the item specifications?

The distracter characteristics section describes the features of both the correct response and the incorrect response alternatives. Careful construction of the incorrect alternatives is just as important as careful construction of the correct response, because the distracter characteristics affect the difficulty level of the item. The features to be described include the number of distracters, the types of errors to be included in the distracters, and the content limits of the distracters.

Are distracter characteristics described in the item specifications?

Are sample items included in the item specifications?

The sample item part of each item specification, including directions to the student, should exemplify what the content specialists and testing specialists think a good item should include. Sample items may be as helpful to the item writers as clear statements about each of the other three parts of the item specification. Sample items should reflect (1) the difficulty level desired for that set of items; and (2) appropriate language, format, style, length, and so forth.

The above description of item specifications applies specifically to multiple-choice items. Slight modifications can make the development and use of item specifications appropriate for writing samples, performance tests, oral spelling, and so on.

References

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section II, pp. 1—39.

Popham, W. J. "Domain Specification Strategies," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 15—31.

Sample Assessment Exercises Manual for Proficiency Assessment, Grades 4—6, Vol. I, *Sample Exercises.* Sacramento: California State Department of Education, 1978, pp. 1—349.

Writing Items

Even though the item specifications contain clear descriptions of what each item should be like, writing items requires attention to detail on the part of the writer and can be a time-consuming process. Except in the case of generating computation items, good writers can rarely produce a large number of items in a day. This should be kept in mind when allocating time and personnel to this step in the test construction process.

Have enough staff been assigned to item writing?

Teachers and other content specialists who have been trained in the use of item specifications can serve as writers. More than one writer per content area is desirable because this practice (1) reduces the work load for each writer; (2) may increase the range of coverage in the items produced; and (3) allows for critique of items from others involved in the process.

Number of Items

Have enough test items been written to allow the creation of multiple test forms?

The purpose of the initial stage of item writing is to produce a pool of draft items from which to choose those items that will be included in the actual test. Item review and field testing are used to reduce the number of items from the initial item pool. Within the available time constraints and without compromising quality, item writers should try to produce as many items as possible for each specification. The minimum number of items to be written for each item specification depends on the number of test forms being constructed at the same time and on the proposed amount of overlap between forms.

Although experience has shown that few items are thrown out when tight specifications are used, there are still benefits for writing a substantial number of items at once. These advantages include (1) economy—it is cost efficient to train item writers only once; (2) ease of field-testing alternate forms—it is easier to construct and field-test comparable forms of the test by starting with a large pool of items; and (3) domain specificity—it is easier to stay within the bounds of an item specification by writing items at one time rather than having on-going item writing.

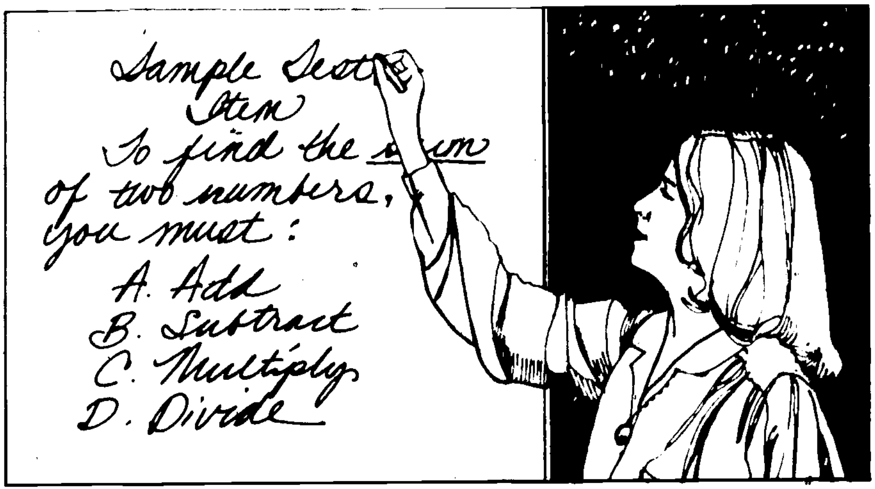
Characteristics of Good Test Items

Questions written for proficiency testing can take many forms, but most districts have elected to use the multiple-choice format. Its advantages include economy and objectivity for scoring large numbers of tests, as well as diagnostic utility. Still, other types of exercises will work equally well. True-false, completion (fill-in-the-blank), and essay items can be, and are, used for proficiency testing. The item format should match the manner in which the skill is presented in the curriculum.

Four parts must be written for each multiple-choice item (these parallel the parts of the item specification): (1) the directions to the student; (2) the item stem; (3) the correct answer; and (4) the distracters. Several good sources are available on how to write good multiple-choice items (see the partial listing on p. 12). Some of the most important guidelines are listed below:

Do all test items conform to item-writing rules?

- Each item should have one, and only one, correct answer.
- The position of the correct response alternative should be varied across items.



- The language used in the item stem should be simple, direct, and free of ambiguity.
- Double negatives should be avoided.
- The item stem should pose a complete, clear question for the examinee. Such a question is one that the student should be able to answer without reading the distracters.
- Information that can be placed in the item stem should not be repeated in each response alternative.
- If the same passage, problem, graph, chart, or other stimulus material is to be used for two or more items, the directions to the student should clearly state this fact.
- When several items are based on the same passage, graph, or chart, each item should be independent; that is, the student's determining the correct answer for an item should not depend on the student's having correctly answered a previous question.
- All distracters should be stated clearly and concisely.
- All distracters should be approximately the same length.
- Distracters that overlap or include each other should not be used (synonymous distracters should be avoided).
- All distracters should be grammatically consistent with the item stem and should be parallel in form.

References

Gronlund, N. E. *Constructing Achievement Tests* (Third edition). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1982, pp. 36—80.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section II, pp. 7—70.

Sample Assessment Exercises Manual for Proficiency Assessment, Grades 4—6, Vol. I, *Sample Exercises.* Sacramento: California State Department of Education, 1978, pp. 1—349.

Swezey, R. W., and R. B. Pearlstein. *Developing Criterion-Referenced Tests.* Alexandria, Va.: Applied Science Associates, Inc., 1974, pp. 4.1—4.7.

Pretesting and Revising Items

Do all test items conform to item specifications?

Before the items are assembled into a provisional form of the test, each one should be reviewed to ascertain whether it (1) matches the item specifications; (2) is free from bias; and (3) is written in accordance with good item-writing principles. At this stage the review process may be informal, involving only written comments by reviewers on suspect items. More formal review procedures, involving ratings of every item by reviewers, may also be undertaken.

It is important that the draft form of the item include the artwork that will be used on the final version of the item. If stimulus material is added at the last moment, the prior development and review procedures can be compromised. For example, if an item is based on an ad from *TV Guide*, substituting an ad from *Scientific American* could change the item significantly. For item review and pretesting, items should be in the most final form possible.

During the initial review, items should be checked for several properties, which requires the expertise of several types of specialists. The team of reviewers should include teachers and other content specialists, individuals familiar with the curriculum and its stated objectives, persons familiar with the actual instructional practices used in the district, test construction specialists, and possibly community members.

Have all items on the proficiency test been reviewed by teachers and other educational specialists not involved in developing the items?

Sources of Bias and Irrelevant Difficulty

An initial check for bias should be made for each item before the items are pretested on students. (For more information on identifying biased items, see p. 30.) Item bias exists when some characteristic causes the item to be offensive or excessively difficult for a particular ethnic, cultural, or sex subgroup. Sources of bias include idiomatic expressions, words that have different meanings for different groups, or concepts that are not taught in school or are unfamiliar to a subculture. Items containing potential sources of item bias should be revised or designated for close examination when the results of field testing become available. (For information on statistical methods for identifying biased items, see p. 31.)

Have all items been reviewed for bias and irrelevant difficulty?

Irrelevant difficulty exists when an item characteristic causes the item to be more difficult than intended for all students. Sources of irrelevant difficulty include (1) complex sentences or difficult words in items that are designed to measure skills other than reading comprehension; and (2) in the item stem, information that is not needed to answer the question and that may cause confusion.

Pretesting of Items

Pretesting is the informal tryout of items on students who are similar to those who will be taking the final form of the test. Pretesting differs from field testing in at least three ways. First, pretesting is concerned only with determining how good individual items are, while field testing involves both individual items and the test as a whole. Second, pretesting is more informal than field testing. The choice of student samples and the testing conditions are largely a matter of convenience in pretesting; but in field testing, the samples of students and testing conditions must be more rigorously selected. Third, the information sought in pretesting and field testing differs. Pretesting focuses on students' opinions about which items are ambiguous, difficult to understand, and so on. The focus of field testing is on students' performance on the items and on the test as a whole; statistical item analyses are used to help make inferences about the items.

Have all items been pretested on a small but representative group of students?

Students should be asked to complete the items as if they were actually taking the test and to indicate which items (1) were unclear, poorly worded, or confusing; (2) seemed to have more than one correct alternative; (3) seemed to have no correct alternative; or (4) contained content or involved skills that had never been addressed in their instruction. After the students take the test, they should go over their own tests and provide indications of problematic items.

If a substantial number of the pretested students identify the same flaw in an item, the item should be reviewed again by the team of reviewers and revised or discarded.

References

Gronlund, N. E. *Constructing Achievement Tests* (Third edition). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1982, pp. 93—96.

Hambleton, R. K. "Test Score Validity and Standard Setting Methods," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 84 -97.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section II, p. 48; and Section III, pp. 1 4.

Addressing Other Factors in the Test Construction Process

Before beginning the test construction process, districts need to determine how long each step in the process will take and how those steps will fit into the proposed test administration schedule. Districts should also determine the personnel available for test development, including the types and extent of community and other professional support available. Other factors to be considered include reproduction of the test, test length, test security, and the use of multiple test forms.

Design and Reproduction of the Test

The design and reproduction of the test should not be overlooked in the test construction process. In fact, these are critical factors that can influence how students respond to proficiency testing. The layout of the test requires decisions about how many items to place on a page and the grouping of items by stimulus material (all items based on a passage or graph should be placed on the same page). If items are bunched too closely together, students will have difficulty concentrating on the item at hand and may confuse the sequence and mismark the answer sheet.

A related topic is the readability of the print to be used. Research on the readability of various typefaces shows that print with serifs (short lines stemming from, and at an angle to, the upper and lower ends of the strokes of a letter) are easier to read than those without serifs. The type size to be used depends on the grade level and age of the students for whom the test is intended, but a rule-of-thumb is to match the type size to that used in the classroom reading materials used by the students taking the test.

Copyright infringement is a serious legal matter. If magazine ads and similar stimuli are to be used in the test items, several factors merit consideration. Copyrighted materials must not be used without the expressed written permission of those who hold the copyright. Permissions are not required for material in the public domain. The reader should be aware that most large circulation magazines and some newspapers are copyrighted in their entirety. If it is not known whether material is copyrighted, one of the following actions should be taken:

- Do not use the material.
- Contact the publisher (or the holder of the copyright if other than the publisher) to secure permission.
- Consult an attorney.

The quality of illustrations to be reproduced is important. Generally, color photos do not reproduce well in black and white. If possible, the size of the type in the stimulus items should be as large as that used in the test questions. Magazine and newspaper items quickly become dated. Prior to test production and administration, the appropriateness and relevance of the stimulus materials should be checked.

Other test production considerations include quality reproductions, administration directions, and test security. Ditto masters usually do not produce sharp copy. A slight reduction in size may result when items are reproduced on ordinary copying machines. This is especially important if exact size is required in the reproduction (for example, on a measurement task, where reduced stimulus material could cause students to become confused). The purpose of the test and the directions for administering it should be printed on a single cover page. This helps students to focus their concentration on the directions and ensures that all students start at the same time. For security considerations (see p. 18), it may be worthwhile to stamp each test with a serial number and use a tape clasp on the edge of each test booklet.

Test Length

Is the testing time commensurate with available resources, staff/student time, and the purpose(s) of proficiency assessment?

In general, long tests provide more information about a student's performance than short tests. But the relationship between test length and the reliability of decisions made about students on the basis of test results is not so direct and simple. The results of reliability analyses from field testing (see the discussion on reliability, p. 36) can provide estimates of the number of items each test (reading comprehension, computation, writing, and so forth) should contain.

Other factors that affect test length include overall costs (of development, administration, scoring, and interpretation); testing time available; and, of course, the importance of the decisions to be based on the test scores. Proficiency tests used at the lower grade levels (four through nine) are used to identify students in need of remediation and may require fewer items than the tests used for the diploma sanction.

No hard-and-fast rules exist about how many items should be included in each separate proficiency test. The appropriate number of items may vary for each test and item type. Tests used to certify students for graduation or to assign students to an extensive remedial program should have at least 50 items. A mastery/nonmastery determination for a narrowly defined skill (for example, the student can make change for a consumer purchase) can be made with reasonable assurance with relatively few items (four to ten). If proficiency test results are to be used for diagnostic purposes, reliability and validity indexes should be developed for each subtest.

Multiple Test Forms

Multiple test forms are different forms of a proficiency test that are comparable in terms of difficulty, content, and the item specifications on which the items are based. They differ, however, in the specific items they contain.

Have multiple forms of the proficiency test been developed?

Constructing multiple forms of a proficiency test is important for several reasons, the most important of which is test security. If the same test is used repeatedly, students in subsequent testings may learn which items are on the test from those who took it earlier.

Second, valid interpretation of scores from repeated testing may require multiple forms. If a student is retested with the same test after a period of remediation, it may be difficult to determine whether improvements in scores should be attributed primarily to newly acquired mastery or to the student's memorizing the answers to specific items.

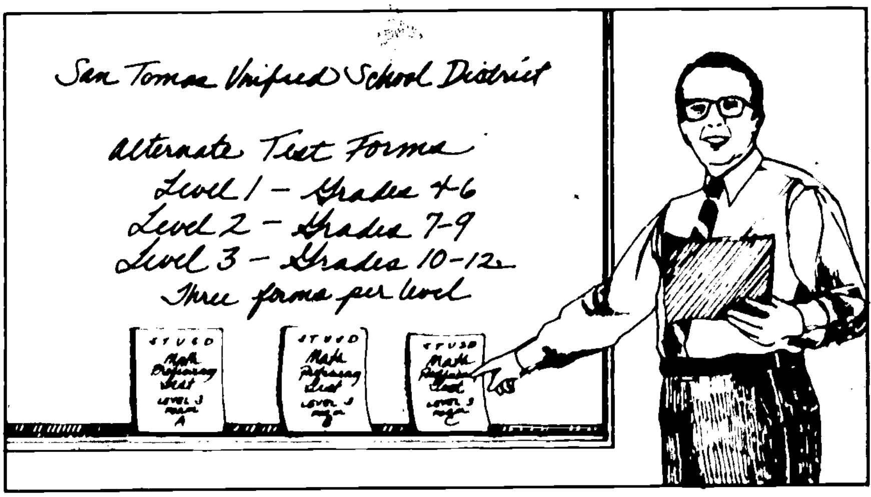
Third, an ongoing test construction process in which multiple forms of tests are produced facilitates incorporation of curricular and instructional changes. The requirement that proficiency tests measure only what is taught in the curriculum may be met more easily if a district develops multiple test forms. As new test items are developed, they should match the item specifications and reflect any changes in curriculum and instruction.

Do all test forms demonstrate curricular and instructional validity?

In constructing multiple test forms, districts should be concerned with (1) the comparability or equivalence between multiple forms; and (2) the number of multiple forms to develop. Often, test developers create multiple test forms that contain some common items. This practice reduces the number of items that have to be written and maximizes the equivalence between the multiple forms. Each multiple form must undergo field testing and adjustment for differences in difficulty levels among forms. One way to ensure equal difficulty levels between multiple forms is to administer all items measuring a single proficiency standard to a sample of students at one testing. (Counterbalancing the presentation of items will mask any order effects.) The items should then be ranked by p-values (see p. 23) and then randomly assigned to the multiple forms.

Are all forms of proficiency tests comparable?

The number of multiple forms that should be developed depends on a district's policy about retesting and the number of times that students will be tested each year. A separate test form should be developed for each regularly scheduled testing during the year. The same forms should not be used on consecutive occasions. If resource constraints preclude the development of multiple test forms, scrambling the order of items will create the illusion of different tests and help to prevent cheating.


San Tomas Unified School District
Alternate Test Forms
Level 1 - Grades 4-6
Level 2 - Grades 7-9
Level 3 - Grades 10-12
Three forms per level
STUVD Math Proficiency Test Level 3 Form A
STUVD Math Proficiency Test Level 3
STUVD Math Proficiency Test Level 3

Test Security

Have the proficiency tests been kept secure?

When decisions about students are based on proficiency test data, it is assumed that (1) all students had an equal opportunity to learn the material or acquire the skills on which they were tested; and (2) no students were given an unfair advantage in the testing situation, such as longer time limits for timed tests or prior practice on the exact items on which they were tested. It is particularly important that the content of proficiency tests be kept secure; that is, to ensure that test items not be made available to students, teachers, or the public before the actual time of testing.

One person at each school or testing site should be responsible for the safekeeping of proficiency tests, usually in a locked storage area. This responsibility also includes supervising the distribution and collection of test booklets at the time of testing.

When testing is scheduled for several sessions over a short period of time, care must be taken to ensure that the test content cannot be discussed with students who have not yet taken the test. This problem can be avoided by administering the test to all students at once, either in a large-group session or during the homeroom period. Another technique is to use multiple forms of the test so that items vary within and across these administrations.

Some teachers like to provide students with practice on the *types* of items on which they will be tested. Similarly, parents and students deserve to know what kinds of questions will be asked. Districts should provide sample items and descriptions of the test to teachers, students, and parents (see "Test Documentation," p. 40), but no proficiency test or subtest should be available for review in its entirety prior to the time of testing. Even reviewers should be given only parts of a test so that the greatest degree of security can be maintained.

Have sample test items and test descriptions been shared with teachers without compromising test security and so that instruction is linked to assessment?

References

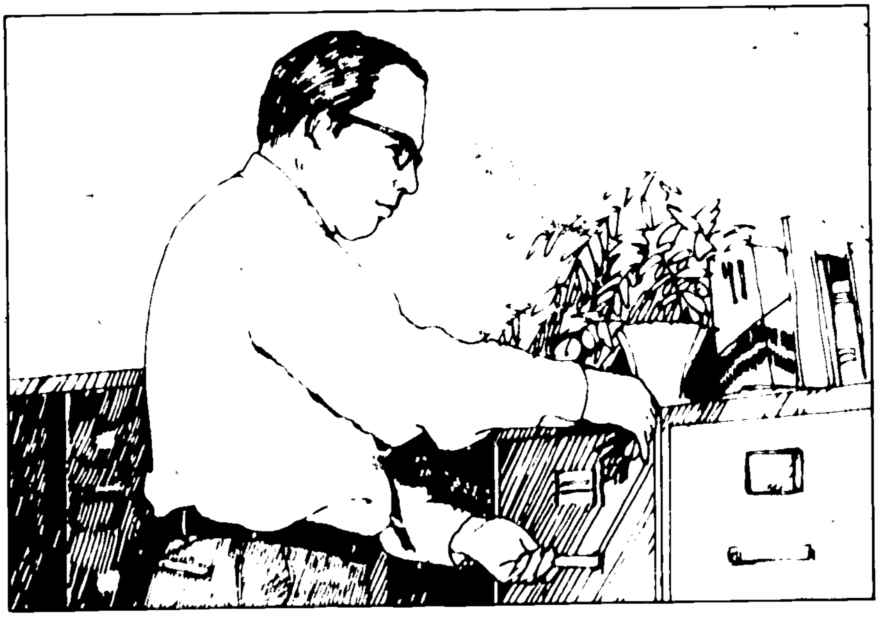
Gronlund, N. E. *Constructing Achievement Tests* (Third edition). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1982, pp. 96 100.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section I, pp. 37 53; Section II, pp. 42 45, 89 120, and 114 18.

Hambleton, R. K, and others., "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments," *Review of Educational Research,* Vol. 48 (1978), 23 25.

Swezey, R. W., and R. B. Pearlstein. *Developing Criterion-Referenced Tests.* Alexandria, Va.: Applied Science Associates, Inc., 1974, pp. 3.1 3.5.

Thorndike, R. L. "Reproducing the Test," in *Educational Measurements.* Edited by R. L. Thorndike. Washington: American Council on Education, 1971, pp. 160 87.



Test Validation

Important decisions about student remediation and graduation warrant the use of high-quality tests. This chapter deals with methods for gathering technical evidence of test quality. The word "technical" is used because the methods described are generally systematic, and the evidence is summarized numerically.

Several factors must be considered when determining the technical quality of a test: item analysis, passing scores, bias, validity, and reliability. If the evidence suggests that the quality of the test is high in all, or most, respects, district personnel, students, parents, and the community can feel confident about the decisions based on the test scores. If the evidence suggests marginal quality, revision or other adjustments are required to ensure that decisions based on test score interpretations are valid.

To understand or implement the guidelines in this chapter requires minimal statistical expertise. However, readers who are unfamiliar with the statistical indexes or methods suggested herein can refer to the cited references to obtain simple but complete definitions, instructions, and computational algorithms. Calculations should require only a small calculator; computer calculations are not necessary, but they may simplify the data entry and manipulation procedures.

All the topics addressed in this chapter involve both empirical data and the expert judgment of content specialists. Statistics and other numerical summaries are strictly estimates and should serve only as tools for the deliberations of content and testing specialists. The activities recommended in the previous chapter (including the development and refinement of proficiency statements, item specifications, and test items) may well be the most convincing and soundest evidence for test validation.

Conducting Item Analyses

Items that survive pretesting may still be faulty; item analysis is a further check on item quality. A field test should be administered to a sample of students to try out the items and the test as a whole. The results from the field test can be analyzed with simple formulas to help identify potentially poor items. However, decisions about whether or not to include items in the final test form should not be based solely on these results; input from teachers and other content specialists should ultimately be used as a basis for these decisions.

Samples for Field Testing

Have students in the field test sample been carefully selected?

The group of students involved in field testing should be selected carefully. The results from the field testing will be used to modify the test so that the test scores can be used intelligently in making important decisions about students' futures. The students selected for field testing should be assigned to one or more of the following groups:

- Group A—Students of both sexes from all ethnic groups represented in the population of students for whom the test is intended.
- Group B—Students like those in group A and who would be expected by teachers to perform well (masters)
- Group C—Students like those in group A and who would be expected by teachers to perform poorly (nonmasters)
- Group D—Students like those in group A and who are expected by teachers to have scores right around the passing score (borderline students)

Groups A through D are designed to enable district staff to conduct the validation studies described below (for each validation technique the necessary groups are listed). Students in groups B, C, and D may also be counted in group A (see Fig. 3).

Students in groups B and C should be as similar as possible in all respects except proficiency in the content domain. Variables on which they should be as similar as possible include ethnicity, sex, socioeco-

nomic status (SES), and age (or grade level). The assignment of students to these groups is a very complex task, and it may be difficult to ensure that SES levels are comparable.

Another difficulty may lie in identifying students who are truly masters (group B), nonmasters (group C), or borderline students (group D). Teachers, counselors, and other personnel who are familiar with students' achievement levels should select the students who are to be included in the masters, nonmasters, and borderline groups. Subsequent validation techniques are based on the accuracy of these classifications, and so it is important to have discrete groups identified. Teacher judgments regarding student mastery should be made at the level at which the test is designed. For example, if the proficiency

Have teachers categorized students in the field test as masters, nonmasters, or borderline students?

SAN TOMAS UNIFIED SCHOOL DISTRICT

Reading Field Test Sample

	Sample groupings				
	Population parameters*	Group A (stratified random sample)	Group B (masters)	Group C (nonmasters)	Group D (borderline students)
Total	324	Ta=200	Tb=80	Tc=80	Td=60
Sex					
Male	152	90	35	40	22
Female	172	110	45	40	38
Ethnicity					
Indian	0	0	0	0	0
Asian	17	11	5	3	1
Filipino	24	15	6	7	6
Black	4	2	1	1	1
Hispanic	112	68	28	32	21
White	167	104	40	37	31
Language fluency					
LEP					
Spanish	78	55	20	21	12
Pilipino	17	8	2	5	3
FEP	25	13	4	4	6
All others	204	124	54	50	39

*Demographic breakdown of entire eleventh grade class.

NOTE: This district does not have enough students who are clearly masters or nonmasters, and so groups B and C do not have the recommended number of subjects. Also, note the overlap in membership between groups; group A overlaps with groups B, C, and D; but groups B, C, and D are mutually exclusive.

Fig. 3. Sample field test sample

test is designed as a survey test, then the judgment should be made at the broader content area level. But if the test contains several subtests, for example whole numbers and fractions within computation, then judgments should be made at the subtest level. Several ways exist to assign students to the master, nonmaster, and borderline groups:

- Have teachers discriminate masters from nonmasters, using sample skills assessed on the proficiency test.
- Have teachers or other staff administer short oral quizzes or performance tests to estimate mastery levels.
- Use surrogate measures of student ability, such as reading groups, textbook levels, or individualized education programs.
- Use other test data to separate masters from nonmasters; exercise caution here, as norm-referenced tests are often "contaminated" criteria.

These classifications need to be made for each content area, too. Students who are masters in computation may be borderline students or nonmasters in reading or writing. If possible, the sample should include at least 100 students in each group; a total sample of more than 300 students is rarely necessary. An equal number of students in groups B and C is strongly recommended (this requires that nonmasters [group C] be oversampled). Oversampling for this purpose means selecting a disproportionately high number of nonmasters (relative to the total student population) so that group C has the same number of students as group B.

Administration of the Field Test

Was the field test administered like an actual assessment?

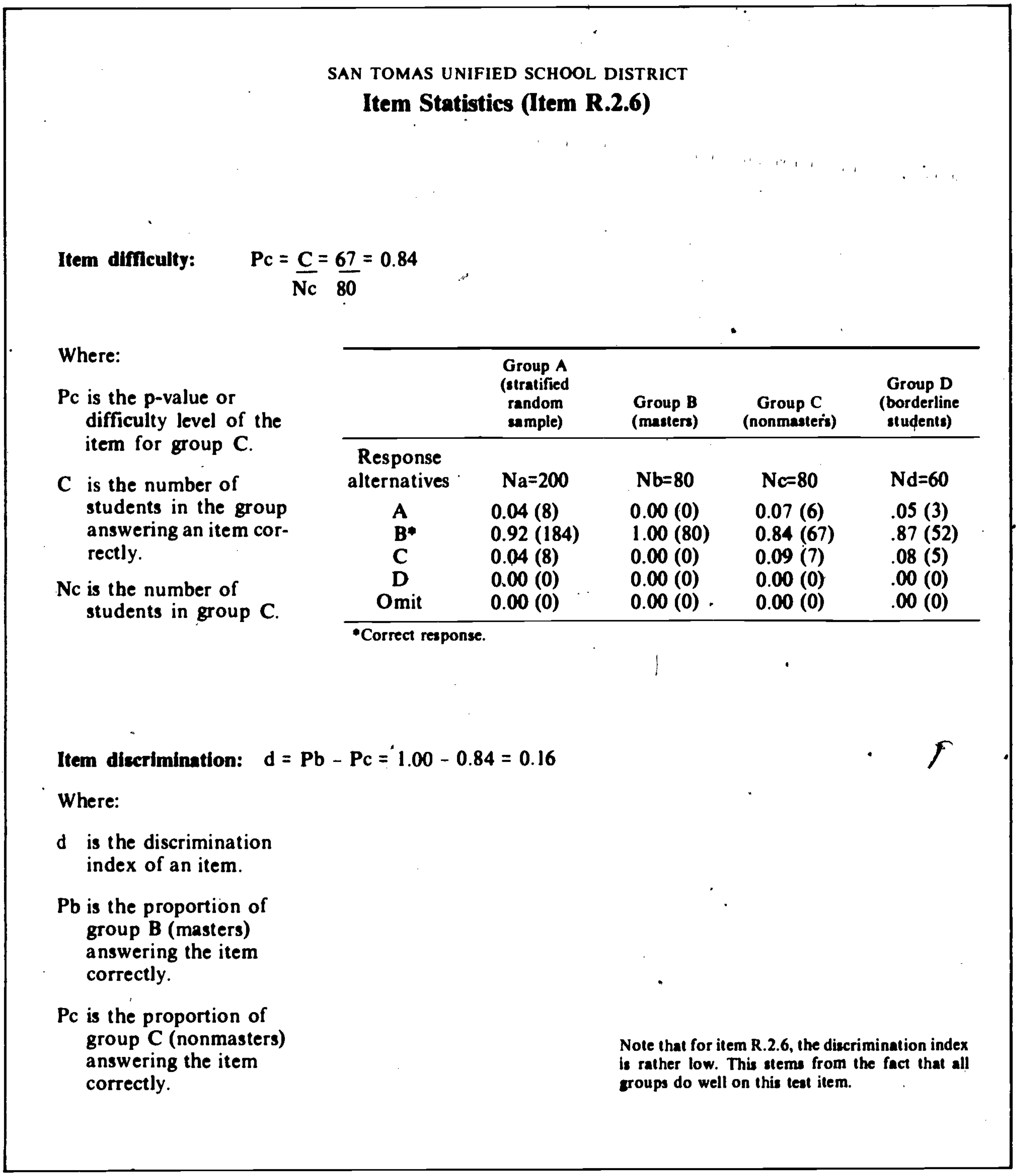
The field test is also used for trying out the instructions, time allotment (if any), format, and the like. Therefore, the field test should be administered to the field test sample as if the results were to be counted; that is, in the same manner in which the final test form will be administered. Since most proficiency tests are already in use, revised tests may be field-tested in the context of the regular proficiency testing schedule.

The students and the test proctors should comment on the test administration procedures immediately after the test is given. In this way the procedures can be modified as necessary. Several students and proctors should be asked to critique test items on a separate sheet of paper.

Item Difficulty Index

Have p-values been computed and analyzed for various subgroups (for example, by ethnicity, sex, and ability levels)?

The item difficulty index is simply the proportion of students who answered the item correctly; it is commonly called the "p-value" (see Fig. 4). P-values range from 0 to 1.00, with high p-values indicating that a high proportion of students get the item right; conversely, low p-values indicate that a low proportion answer the item correctly.



SAN TOMAS UNIFIED SCHOOL DISTRICT

Item Statistics (Item R.2.6)

Item difficulty: $Pc = \frac{C}{Nc} = \frac{67}{80} = 0.84$

Where:

Pc is the p-value or difficulty level of the item for group C.

C is the number of students in the group answering an item correctly.

Nc is the number of students in group C.

	Group A (stratified random sample)	Group B (masters)	Group C (nonmasters)	Group D (borderline students)
Response alternatives	Na=200	Nb=80	Nc=80	Nd=60
A	0.04 (8)	0.00 (0)	0.07 (6)	.05 (3)
B*	0.92 (184)	1.00 (80)	0.84 (67)	.87 (52)
C	0.04 (8)	0.00 (0)	0.09 (7)	.08 (5)
D	0.00 (0)	0.00 (0)	0.00 (0)	.00 (0)
Omit	0.00 (0)	0.00 (0)	0.00 (0)	.00 (0)

*Correct response.

Item discrimination: $d = Pb - Pc = 1.00 - 0.84 = 0.16$

Where:

d is the discrimination index of an item.

Pb is the proportion of group B (masters) answering the item correctly.

Pc is the proportion of group C (nonmasters) answering the item correctly.

Note that for item R.2.6, the discrimination index is rather low. This stems from the fact that all groups do well on this test item.

Fig. 4. Sample item analysis statistics

These p-values should be computed for each item and for groups A, B, and C separately. Items that have relatively high p-values (more than 0.85) for nonmasters should be marked for further scrutiny. A high p-value is not necessarily an indicator of poor item quality; important skills are worth testing even if most students do well on the related items. Potential causes of spuriously high p-values for nonmasters include the following: (1) group C students knew the information beforehand (they really were group B masters); (2) the distracters were poor, and, thus, the correct answer was obvious; (3) the item was actually measuring some other domain (skill area); and (4) the item was just too easy.

Items with relatively low p-values (less than 0.50) for masters should also be examined. Potential causes of low p-values for masters include the following: (1) the wording was ambiguous; (2) group B students were really group C nonmasters; (3) the item measured some other domain; (4) the answer key was wrong; (5) the distracters were confusing; and (6) more than one answer could be defended.

If a proficiency test is designed so that there are several subtests measuring a broader content domain, then the p-values of the items within a subtest should be relatively homogenous. P-values across subtests, however, may vary.

Item Discrimination Statistics

Item discrimination statistics (see Fig. 4) are determined by comparing the item responses from the two extreme groups of students, masters (group B) and nonmasters (group C). Students only in group A are excluded from these analyses. Of course, the masters are expected to outperform the nonmasters on each item. The magnitude of the discrepancy indicates the discriminating value of the item.

Have item discrimination indexes been used in determining item validity?

The difference in p-values of masters and nonmasters for each item should be computed. The resulting item discrimination index can be used to determine item validity, because test items are intended to discriminate masters from nonmasters. Items with low indexes (less than about 0.25) should be scrutinized further (although good items can exhibit low discrimination indexes). Items with negative discrimination indexes should probably be discarded.

Item Statistics and Judgment

Decision makers should use the results of the field testing, along with expert judgment, in making practical decisions regarding the final form of the test. Item statistics can be used to identify items in need of additional scrutiny. The directions or format of the test may also be modified after completion of the field test.

The content specialists involved in item reviews (based on field testing) should be different from those who wrote the items. Their job will be to consider simultaneously item difficulty, item discrimination, effects of distracters, and each item's congruence with proficiency statements.

Have content specialists reviewed each item, using item analyses to guide decisions about inclusion, exclusion, or revision?

The content specialists should be certain that the format is appropriate, no irrelevant difficulty exists, the domain is adequately sampled, and the required skills and knowledge were actually taught. The field test data and the reports of content specialists and students should be used to revise test items, format, instructions, and time limits. If major changes are made in the test, the field testing should be repeated.

References

Berk, R. A. "Item Analysis," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 49—79.

Hambleton, R. K., and others. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments," *Review of Educational Research,* Vol. 48 (1978), 34—42.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section III, pp. 4—19; and Section III, pp. 75—93.

Sample Assessment Exercises Manual, Vol. II, *Item Statistics.* Sacramento: California State Department of Education, 1978.

Setting Passing Scores

Passing scores on proficiency tests are set to identify those students who require remediation or do not qualify to receive a high school diploma. The passing score is designed to help minimize errors in classification of students, and many factors influence what that score should be. These factors include both expert judgment and empirical data. All methods of setting passing scores involve judgment to some extent. Students with scores just below the passing score deserve special attention; additional information should be considered in determining their classifications.

Definitions of Mastery and Nonmastery

Have district staff and community representatives been involved in defining levels of mastery (setting passing scores)?

It is imperative that clear-cut definitions of the minimal level required for mastery at each testing level be established. These definitions will eventually be expressed numerically as passing scores on the proficiency test. A committee composed of community members and school personnel should define as precisely as possible the minimum level of mastery to be attained in each domain. The school curriculum and goals of the community are important factors in arriving at these definitions.

The definitions of nonmastery may distinguish between graduation and remediation. In general, more stringent requirements (and ultimately higher passing scores) should be set for purposes of remediation than for graduation, because errors in determining the need for remediation are far less critical than those that deny 'students diplomas.

Judgmental Methods

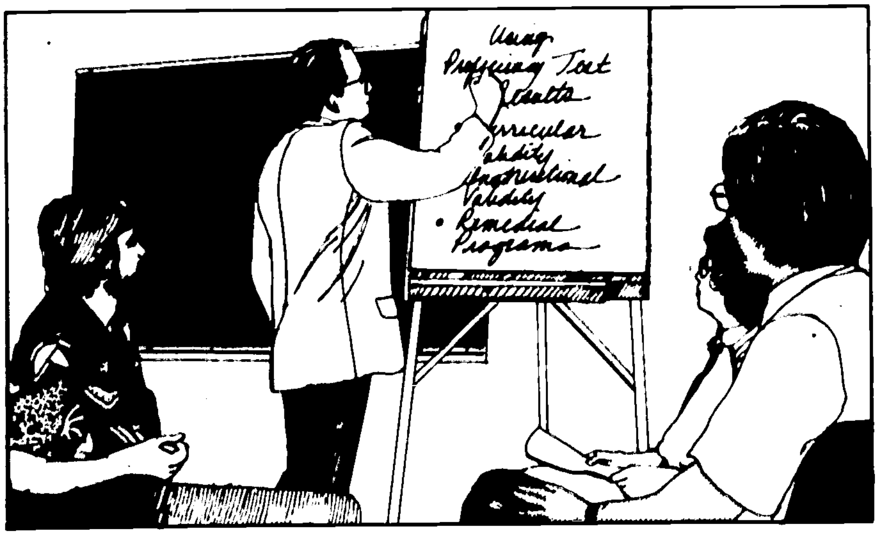
Were judgmental methods used in setting passing scores?

With judgmental methods for determining passing scores, greater consideration is given to the test items than to the ability levels of students taking the proficiency test. Basically, judgments are rendered about the difficulty levels of items on the test. The empirical methods discussed in the next section focus on student performance data in the setting of passing scores.

Content specialists can determine preliminary passing scores by taking the definitions of minimal mastery into account and by examining the individual items. The items should be judged on the basis of their importance in the curriculum and their difficulty for borderline students.

Three formalized item review methods for setting passing scores on proficiency tests can be used. These methods, named after their developers, are the Nedelsky, Angoff, and Ebel methods. Each method results in a preliminary passing score, which should be carefully scrutinized by participants in the passing-score-setting process.

The references at the end of this section provide complete details about each of these procedures, but the Angoff method is briefly described here for illustrative purposes. Judges using the Angoff method are directed to estimate the probability that "minimally proficient" students will answer each item correctly. The probabilities of



success (correct responses) are summed across all items on the test, and the total represents the passing score for that test.

Empirical Methods

Were empirical methods used in setting passing scores?

Empirical methods require judgments about students' performance rather than analysis of the items. A sample of student responses to items can be used in setting the passing score. The borderline method or the contrasting groups method, or both (described below), should be used for setting preliminary passing scores.

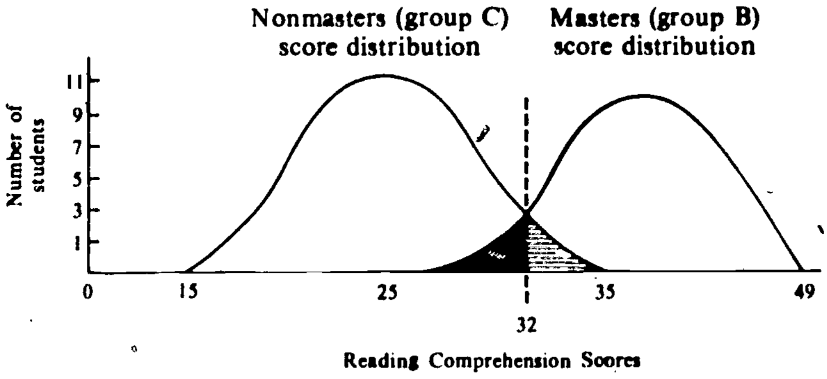
In the contrasting groups method (see Fig. 5), the field test results from groups B and C are plotted. The passing score is placed where the distributions of scores of masters and nonmasters intersect. For the contrasting groups method, each group should include the same number of students.

For the borderline method, test data are collected from students who are judged to be so close to the borderline between mastery and nonmastery of the skill that the judges are uncertain which way to classify them. The passing score is placed at the median of the test scores for the borderline cases. For the borderline method roughly 100 borderline students need to be identified for group D.

If possible, the contrasting groups method, rather than the borderline method, should be employed because with the contrasting groups method, classification errors are minimized and can be identified on an individual basis. The contrasting groups method also provides an indication of the magnitude of classification errors.

SAN TOMAS UNIFIED SCHOOL DISTRICT

The Contrasting Groups Method



NOTE: Nonmasters to the right of the passing score (32) are incorrectly classified, as are masters to the left of the passing score.

Fig. 5. Using the contrasting group method to set passing scores

Other Considerations About Passing Scores

Have passing scores been reviewed since they were first determined?

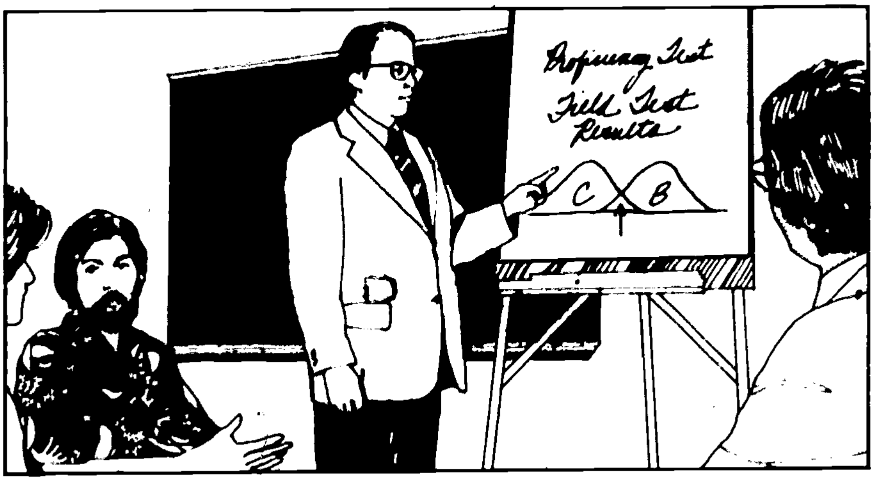
Different passing scores can be expected to result from each judgmental and empirical method employed. The methods in which judges have the greatest confidence (with respect to fairness) should receive the most weight in the final determination of the passing score. Methods that result in passing scores that are very different from the passing scores arrived at through the use of other methods should be checked for faulty utilization. No matter what standard-setting process is used, the resultant passing score should be reviewed by the community and other lay audiences.

Statistical confidence in the reliability of passing score decisions increases as the number of items increases. However, if a district has developed a proficiency test composed of several subtests, then passing scores should be set at the subtest level. While there will be less statistical confidence in these decisions because most subtests have fewer items than do complete tests, passing scores on subtests do provide increased precision in pinpointing student deficiencies for remedial instruction.

It is also necessary to keep in mind that the passing score ultimately determines how many students will require remediation. Therefore, the passing score should not be set so high that not enough resources exist for remediation.

Alternatives to a Single Passing Score

If possible, decisions on some students should not be made solely on the basis of whether their scores are above or below the passing score. Since no test is perfectly reliable, students scoring just below (or above) the passing score may indeed be masters (or nonmasters). A band of scores around the passing score should be used to identify borderline students. The band could be defined as one or two standard errors of measurement (from exhibit 1, p. 3, Section IV, of the



Handbook for Proficiency Assessment), depending on the available resources for providing remediation to students whose sçores are located within the band. The standard error of measurement can be estimated, or it can be calculated from a formula given in any elementary measurement textbook.

Another method of setting the band width is to use the passing scores determined from two or more different methods of determining passing scores. This should be considered when two methods are judged equally accurate but the resulting passing scores are different. Students whose scores are above the higher boundary of the band should be allowed to graduate. Additional information should be used to determine the mastery status of students whose scores lie within the band. This additional information should include some, if not all, of the following: retest score, relevant course grades, teacher remarks, results of parental and student discussions, and other test scores. Those students whose scores fall below the lower boundary of the band should not be eligible for graduation.

Is additional academic information used to determine mastery status of students whose scores are near the passing score?

References

Hambleton, R. K. "Test Score Validity and Standard Setting Methods," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 80—123.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section IV, pp. 1—57.

Reviewing the Test for Bias

With respect to bias in proficiency testing, a revision of the Pupil Proficiency Law (AB 3369, Chapter 1333, Statutes of 1980) states:

> It is the intent of the Legislature that the governing board of each school district make every effort possible to periodically screen assessment instruments for racial, cultural, and sexual bias.

Bias occurs when some facet of a test or of the test administration procedures distorts a subgroup's true achievement level. In fact, the Legislature is so concerned with the potential for disproportionate impact of proficiency testing that it has directed the Department of Education to study the effects of proficiency assessment on linguistic and ethnic minorities. It is important to realize, however, that differential performance by minority subgroups does not necessarily indicate biased tests. Many factors may account for differential performance on proficiency tests.

Detecting bias is even more difficult than defining it. Bias detection techniques should be employed during the development, field testing, and administration of a test. Although no procedure is infallible, a variety of bias inquiries helps ensure a sound, fair assessment. Content reviews are subjective analyses of bias in test items and administration procedures. These reviews usually involve a group process in which representatives of the community and school district scrutinize each test item. Statistical methods for identifying bias make use of field test data (for each subgroup) to identify items that may be biased.

Have both subjective and statistical bias reviews been conducted?

Subjective Content Reviews

A preliminary step in conducting a subjective content review is to identify within the district the subgroups that the test may be biased against. In the conduct of the content review, it is important to involve individuals who adequately represent the identified subgroups. The representatives of a linguistic minority, for example, should know the language and the culture of the subgroup. It is a good practice to have more than one person representing each subgroup. The school district representatives should include curriculum and testing specialists.

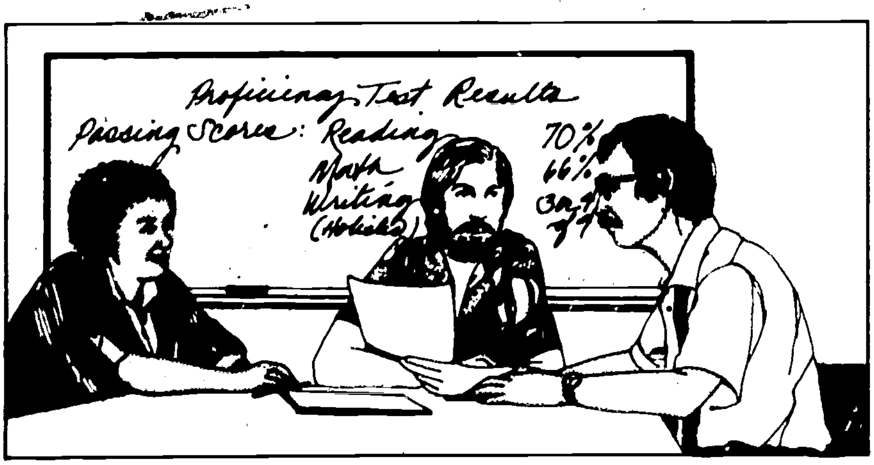
Have subjective reviews for bias included representation from significant minority groups?

The process for conducting the content bias review should be developed before the actual review begins. Sample forms, ratings, and items from other districts may be made available to reviewers for study before a meeting takes place. Training in the bias review procedure may be necessary if the reviewers are novices or if the procedure is somewhat complex. As part of the content review procedure, it is essential to have a rule for reaching consensus on whether an item is biased and whether it can be revised or should be deleted.

Statistical Bias Reviews

Field test data can be used to detect item bias if the data have been categorized on the basis of the various subgroups of interest. Rigorous, well-documented field test procedures are important, because the integrity of the item bias review may depend on the quality of the data

Have statistical bias reviews been based on quality field test data?



collected. For field testing it is necessary to have a sufficient number of examinees (50 to 100) from each subgroup.

Two of the many statistical techniques are described briefly here, but the reader should consult the references at the end of this section for information on other methods. The adjusted item difficulty approach is based on the use of the simple p-value computed for each comparison group and adjusted for differences in subgroup performance on the total test. This method has the advantages of (1) being easy to present visually; and (2) correlating well with other, more complex techniques. In the Chi-square approach (see Fig. 6), the expected performance on each item is compared to actual item performance for each subgroup. This approach is easy to use, and it too correlates well with other methods.

Other Considerations for Bias Reviews

An important consideration in planning bias reviews is deciding whether to conduct the statistical review before or after the subjective content review. Conducting the statistical review first gives the content review panel important data to help identify biased items but tends to limit the panel's discussion to the statistical approach rather than the judgmental approach.

It is also important to think about how the two approaches—content review and statistical treatment—should be combined. The question of which approach is "right" need not be asked. All items identified as biased by means of either method should be considered suspect and revised or deleted. *NOTE:* All items suspected of being biased should undergo additional pretesting and field testing before being used on a proficiency test.

Have the bias review results been integrated and acted upon?

References

Angoff, W. H. "A Technique for the Investigation of Cultural Differences." Paper presented at the American Psychological Association's annual meeting, Honolulu, 1972. (Available from the author through the Educational Testing Service, Princeton, NJ 08540.)

Scheuneman, J. "A New Method of Assessing Bias in Test Items," *Journal of Educational Measurement,* Vol. 16 (1976), 143—52.

Shepard, L. G.; G. Camilli; and M. Averill. "Comparison of Six Procedures for Detecting Test Item Bias Using Both Internal and External Ability Criteria." Paper presented at the National Council on Measurement in Education's annual meeting, Boston, 1980, pp. 1—79. (Available from L. G. Shepard through the Laboratory of Educational Research, University of Colorado, Boulder, CO 80302.)

Technical Assistance Guide for Proficiency Assessment. Sacramento: California State Department of Education, 1977, Appendix K, pp. K1—K8.

Identifying Biased Subtests

1. Construct score intervals for the subtest so that:
   a. Each cell has 10 to 20 observed (correct) responses per cell.
   b. The number of intervals is less than or equal to 5, but greater than or equal to 3.

Score intervals	Males	Females	Total
7-8	24	28	52
6	22	32	54
0-5	44	50	94
	90	110	200

No single way exists to set up score intervals. If the intervals meet the criteria in step 1, the Chi-square design will work. In the example the males' score distributions were divided into three intervals (because males were the minority group). The lowest interval could have been divided at 5 and 0-4.

2. Compute the expected values for each cell by:
   a. Listing the observed cell frequencies by row and column
   b. Multiplying the row marginal by the column marginal for each cell and dividing by the grand total; e.g., for cell 1, the expected value equals (90 x 52) ÷ 200.
3. Compute the Chi-square by:
   a. Subtracting the expected value from the observed value (d = o–e)
   b. Squaring the difference (d^2) and dividing by the expected value (d^2/e)
   c. Summing the resultant figures

Row (r)	Column (c)	Observed value (o)	Expected value (e)	d	d^2	d^2/e
1	1	24	23.4	0.6	0.36	0.02
1	2	28	28.6	-0.6	0.36	0.01
2	1	22	24.3	-2.3	5.29	0.22
2	2	32	29.7	2.3	5.29	0.17
3	1	44	42.3	1.7	2.89	0.07
3	2	50	51.7	-1.7	2.89	0.06
						$x^2 = 0.52$

4. Find the critical value of Chi-square by:
   a. Calculating the degrees of freedom (df = (r-1)(c 1))
   b. Looking up the value, using the desired confidence level, in an elementary statistics textbook
5. Compare the critical value of Chi-square to the computed value. For this example, the hypothesis is that this subtest is not biased, since the critical value of Chi-square for two degrees of freedom at the 0.05 level of significance is 5.991, which is greater than the computed value (0.52).

Identifying Biased Items

Bias among items can be checked in the same way, except that proportions of correct responses (p-value x 100) for the items are used instead of frequency counts. The set-up for this study would appear as shown for item R.2.6 in Figure 2:

Score intervals	Males	Females	Total
7-8	88	100	188
6	76	97	173
0-5	52	92	144
	216	289	505

The computations for the example follow:

Row	Column	Observed value (o)	Expected value (e)	d	d^2	d^2/e
1	1	88	80.4	7.6	57.8	0.72
1	2	100	107.6	-7.6	57.8	0.54
2	1	76	74.0	2.0	4.0	0.05
2	2	97	99.0	-2.0	4.0	0.04
3	1	52	61.6	-9.6	92.2	1.50
3	3	92	82.4	9.6	92.2	1.12
						$x^2=3.96$

$x^2(df=2,\alpha=0.05)=5.99$.

Here, too, the hypothesis is that this item is not biased. (At α = 0.20, the critical value of Chi-square would be 3.22, and the item would be identified as biased.)

Fig. 6. Using the Chi-square approach to identify test and item bias

Assessing the Validity of the Test

Have validity indexes been computed for each content area?

Validity information provides evidence that tests actually measure what they were designed to measure. This section deals primarily with the systematic assessment of validity that should be conducted after the pretesting of items. Validity indexes are required for each content area.

Decision Validity

Does the validity coefficient show that the proficiency test accurately distinguishes masters from nonmasters?

Decision validity refers to how well the test results can be used to distinguish masters (the higher scorers) from nonmasters (the lower scorers) (see Fig. 7). An important factor in the validity of an instrument used for making certain decisions is the location of the passing score.

On the basis of the agreed upon passing score, students in groups B and C should receive an additional classification as above (pass) or below (fail) the passing score. The proportion of right decisions is the validity index. It should be remembered that wrong decisions may be the fault of the initial grouping decision and not the test.

	Group C (nonmasters)	Group B (masters)
Above passing score (mastery)	Wrong decision (Cell A)	Right decision (Cell B)
Below passing score (nonmastery)	Right decision (Cell C)	Wrong decision (Cell D)

Fig. 7. One method of arranging data to determine decision validity

A useful statistic for determining decision validity is the phi-coefficient (see Fig. 8). The phi-coefficient is a special case of the Pearson product-moment correlation in which the variables of interest are dichotomous. In proficiency assessment one is interested in the amount of agreement between students' mastery or nonmastery and their passing or failing the proficiency test; both variables are truly dichotomous, since a student has either mastered the skills or not and will either pass or fail the proficiency test.

The phi-coefficient is computationally simple. It is a conservative estimate of validity that can range from –1 to +1 (except where differences in marginal proportions cause a maximum value to be reached). An important consideration in using the phi-coefficient in validity studies is that it can also be employed to determine the reliability of a proficiency test, which is discussed later.

Once a validity coefficient is determined, the problems of interpretation arise. The primary dilemma is: What value of validity is acceptable? As with most statistical indexes, little agreement exists as to what constitutes the right amount of validity. It depends on the test and its use.

For tests exhibiting low validity coefficients, methods can be employed to increase the validity coefficient. One method is to reconsider the initial classifications of groups B and C and determine whether or not some students should be reclassified. Another method is to change the passing score so that the classifications from the test (pass/fail) agree with the grouping judgments.

Have passing scores been reexamined or adjusted in light of validity considerations?

Curricular Validity and Instructional Validity

As described earlier (see p. 6), curricular validity and instructional validity refer to the linkage between the proficiency test and what is actually taught in the classrooms. These terms were used by Merle McClung, an expert on legal issues in proficiency testing, to focus attention on the match between what students are taught and what they are held responsible for. According to Mr. McClung, without such linkage, "it would be unfair to deny students their diplomas because they did not learn to be functionally competent."[1]

Have the curricular validity and instructional validity of the proficiency test been checked?

A variety of techniques can be used to ensure curricular validity and instructional validity. For curricular validity each test item should be matched to the curriculum materials on which it is based. Often, district staff identify multiple curricular areas in which the

[1] Merle Steven McClung, "Developing Proficiency Programs in California Public Schools: Some Legal Implications and a Suggested Implementation Procedure," in *Technical Assistance Guide for Proficiency Assessment.* Sacramento: California State Department of Education, 1977, p. K-4.

SAN TOMAS UNIFIED SCHOOL DISTRICT

Validity of the Reading Subtest

	Group C (nonmasters)	Group B (masters)	
Above passing score (mastery)	Cell A (67)	Cell B (78)	145 (a + b)
Below passing score (nonmastery)	Cell C (13)	Cell D (2)	15 (c + d)
	(a + c) 80	(b + d) 80	160

$$\phi = \frac{bc - ad}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

$$= \frac{1014 - 134}{\sqrt{(80)(80)(145)(15)}} = \frac{880}{3731} = 0.24$$

NOTE: This validity coefficient is relatively low, which is the result of nonmasters' demonstrating mastery (Cell A). In this case, the initial classifications into groups B and C may have been based on academic performance unrelated to the proficiency suhtest on reading comprehension.

Fig. 8. Using the phi-coefficient to determine decision validity

material is covered. For instructional validity it is important to document that instruction is offered in each curricular area. This can be done through examination of lesson plans, classroom observations, and the like. The particular methods used to assess curricular validity and instructional validity are less important than communicating the intent of the match to teaching staff.

References

Hambleton, R. K., and others. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments," *Review of Educational Research,* Vol. 48 (1978), 31—42.

Hambleton, R. K. "Test Score Validity and Standard Setting Methods," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 80—123.

Millman, J. "Reliability and Validity of Criterion-Referenced Test Scores," *New Directions for Testing and Measurements,* Vol. 4 (1979), 75—92.

Technical Assistance Guide for Proficiency Assessment. Sacramento: California State Department of Education, 1977, Appendix K, pp. K1—K8.

Assessing the Reliability of the Test

Have reliability studies been conducted for all proficiency tests?

Reliability is a measure of the consistency of the test scores or of decisions based on those scores. Consistency of results is necessary for the results to be meaningful. Student scores across multiple administrations of a single test form or administration of multiple test forms should be close enough that the same decisions would be made regardless of the precise time of testing or form of the test (assuming the skill level of the student has not changed). Conceptually, reliability is related to validity. In validity studies one measures the association between test performance and subjective teacher ratings of mastery. In reliability studies one measures the association between test results on one occasion and test results on another occasion (see Figs. 9 and 10). Both cases involve a check on the accuracy of test score interpretations. For this reason the phi-coefficient can be used for both purposes.

Decision Consistency

For school personnel and others to have confidence in the classifications of students as masters or nonmasters, it is imperative that their classifications be consistent over two or more testing sessions. In other words, a given student's test score and the test passing score should be such that the student will be consistently classified as either master or nonmaster over repeated testings (covering a short time interval). If possible, results from multiple forms of the test or from two administrations of the same test should be compared for consistency. A sample of representative students (like group A on p. 22) should receive two test administrations. If two test forms exist, each should be administered. If only one form exists, it should be administered twice. The two test environments should be as similar as possible.

		Administration 2	
		Nonmasters	Masters
Administration 1	Masters	Wrong decision (Cell A)	Right decision (Cell B)
	Nonmasters	Right decision (Cell C)	Wrong decision (Cell D)

Fig. 9. One method of arraying data to determine reliability

The time span between the two test administrations should be based on two factors: (1) memory—the tests should be administered far enough apart in time so that memory of items on the first test will have little or no effect on responses on the second test; and (2) maturation and knowledge acquisition—the tests should be administered close enough together in time so that maturation and additional knowledge gained will have minimal effects. For most purposes this time span between test administrations should be about one to three weeks. Care should be taken to ensure that virtually no new information and training relevant to the tests are given to the students after the first test administration and before the second test administration.

Was instruction relevant to the proficiency test avoided during the time between the two test administrations?

From their test performance students should then be classified as either masters or nonmasters. Unlike the case of validity indexes, changes in the passing score to raise reliability are prohibited. If the passing score were changed to enhance reliability, in the extreme case the passing score would require 100 percent correct answers, and reliability would approach 1.00 (but few students would pass).

As mentioned earlier, the phi-coefficient can be used to gauge the degree of association between decisions based on two test administrations. Figure 10 shows a computational example of the phi-coefficient used for reliability purposes. Other indexes, such as Cohen's Kappa, can also be used for determining test reliability (the references at the end of this section include studies on the relative merits of other reliability indexes). For writing samples, interrater reliability (that is, concordance among judges) can be computed for checking the consistency of ratings between judges.

Ideally, a decision reliability coefficient would be close to the perfect value of 1.00. This rarely happens in practice. Some reasons for a low reliability value are (1) the tests contain too few items; (2) different forms of the tests are unequal in difficulty, and the relative passing scores of the two tests do not reflect this difference in difficulty; and (3) the tests do not measure the same skills.

Adequate reliability is important. It is neither fair nor in the best interest of the student to have a decision about a student depend on which form of a test is administered. Reliability can be increased by using more items, employing proper equating procedures, and generating the items in each test form from the same item specification.

Number of Items to Include on the Tests

Has the number of test items been reconsidered or adjusted to increase the reliability of the proficiency test?

Although the number of test items can be an important issue in validity, it is traditionally related more to reliability. In general, the greater the number of items, the higher the reliability of the test. The test should contain as many items as item quality, cost, student fatigue, and other factors will allow. Sophisticated techniques exist for determining the optimal number of items to be included on a test (see the references below), and these may be used to estimate the number of items necessary to reach a given reliability index.

SAN TOMAS UNIFIED SCHOOL DISTRICT

Reliability of the Reading Subtest

		Administration 2		
		Nonmasters	Masters	
Administration 1	Masters	Cell A 2	Cell B 174	176 (a + b)
	Nonmasters	Cell C 10	Cell D 14	24 (c + d)
		(a + c) 12	(b + d) 188	200

Group A

$$\phi = \frac{bc - ad}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

$$\frac{1740 - 28}{\sqrt{(12)(188)(176)(24)}} = \frac{1712}{3087} = 0.55$$

NOTE: This reliability index is reasonably high, but it demonstrates the effects of classification errors (cells A and D) on decision consistency.

Fig. 10. Using the phi-coefficient to determine reliability

References

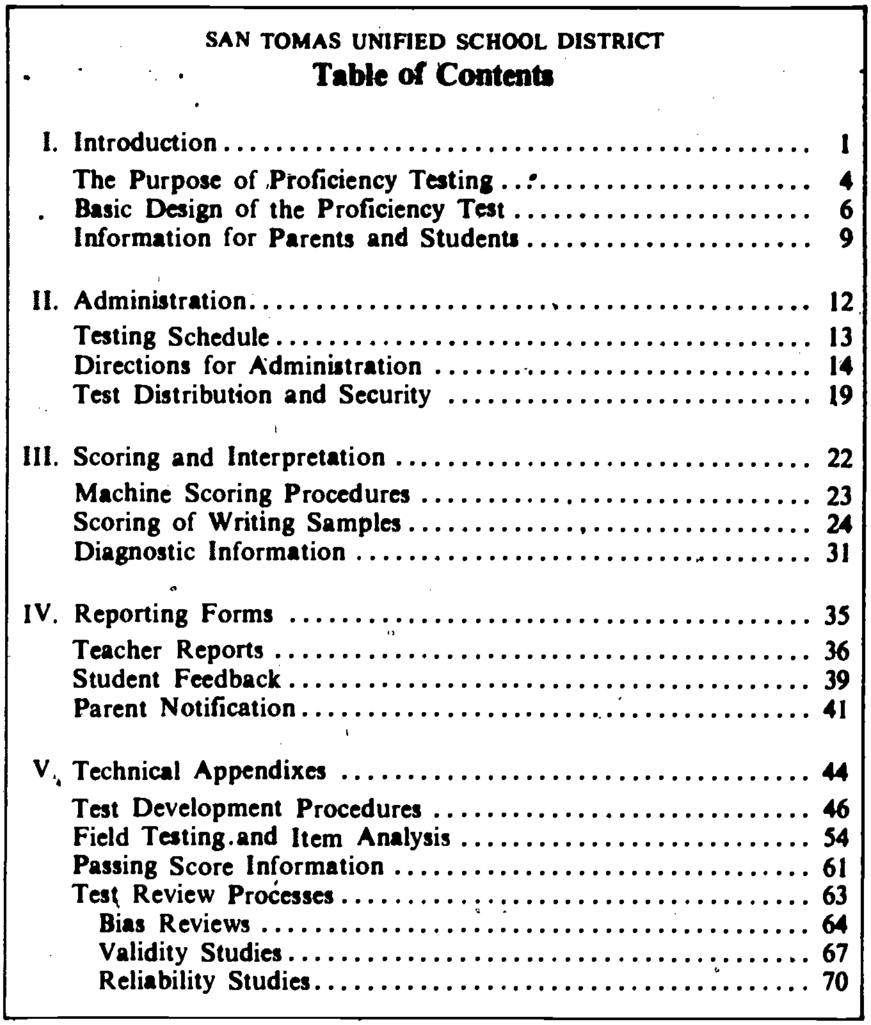
Hambleton, R. K., and others. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments," *Review of Educational Research,* Vol. 48 (1978), 20—25 and 38—42.

Millman, J. "Reliability and Validity of Criterion-Referenced Test Scores," *New Directions for Testing and Measurements,* Vol. 4 (1979), 75—92.

Subkoviak, M. J. "Decision Consistency Approaches," in *Criterion-Referenced Measurement: The State of the Art.* Edited by R. A. Berk. Baltimore: The Johns Hopkins University Press, 1980, pp. 129—85.

Test Documentation

Test documentation is the process of describing a test and the procedures used in its development and use; especially important is a complete set of directions for administering the test. (A sample test manual contents page is shown in Fig. 11.) Although school district test developers typically devote a great deal of effort to the test construction and validation processes, they often fail to document these processes. As a result many of the developmental aspects, which are crucial for verifying the validity of the tests, are soon forgotten. Commercial test publishers have long realized the importance of test documentation. They go to great lengths to develop manuals and test specimens for potential users. Because test results are used to make decisions about students' futures, thorough documentation is essen-



SAN TOMAS UNIFIED SCHOOL DISTRICT

Table of Contents

I. Introduction 1
The Purpose of Proficiency Testing 4
Basic Design of the Proficiency Test 6
Information for Parents and Students 9

II. Administration 12
Testing Schedule 13
Directions for Administration 14
Test Distribution and Security 19

III. Scoring and Interpretation 22
Machine Scoring Procedures 23
Scoring of Writing Samples 24
Diagnostic Information 31

IV. Reporting Forms 35
Teacher Reports 36
Student Feedback 39
Parent Notification 41

V. Technical Appendixes 44
Test Development Procedures 46
Field Testing and Item Analysis 54
Passing Score Information 61
Test Review Processes 63
Bias Reviews 64
Validity Studies 67
Reliability Studies 70

Fig. 11. Sample test manual contents page

tial so that the tests can survive scrutiny from both within and outside the educational community.

In *Proficiency Assessment in California: 1980 Status Report on Implementation of California's Pupil Proficiency Law,* it was reported that the documentation of locally developed proficiency tests varied greatly from district to district. For some district tests no field test data were collected to show whether or not the tests could be used to make accurate and consistent decisions about individual students. Other tests included complete manuals for administration and validation information on both test items and the tests used in field testing. In *Standards for Educational and Psychological Tests,*[1] testing experts agree that better testing results when thorough documentation of the test construction and validation processes is evident.

Three areas are emphasized in this chapter on test documentation:

1. District tests should be described in writing for students, parents, and local groups that want to know more about the local testing process.
2. The documentation for a test should include strict directions for test administration, scoring, reporting, and use of test results.
3. The processes of test construction and validation should be recorded for review (and possible revision) at a later time.

Has test documentation been addressed over the course of proficiency test development and field testing?

Test documentation does not occur in a vacuum. Tests should be documented as they are being developed and field-tested. In this way the writing effort is spread out over time, and the documentation is more accurate because the description is being written at the same time as test development and validation are occurring. Comprehensive documentation allows more informed, open, and critical review of the proficiency testing process and ultimately strengthens the assessment procedure. Furthermore, it ensures confidence in the uses of the tests and gives the examinees the best possible chance for demonstrating their proficiency. Finally, experience suggests that documentation directly influences the test development process, and that keeping a log prompts assessment staff to "rethink" the necessary steps in the development process.

If a district's research or validation procedures have not been completed when a test manual is put together, the developers should acknowledge this omission and set a target date for completion. If additional research or information about the test is too extensive to include in the test manual or documentation package, it should be summarized, referenced in the manual or package, and made available upon request. Language should be used that teachers and parents can understand. Test development, administration, and scoring are procedures that lay persons should be able to understand if the descriptions are written expressly for a lay audience. In many cases

[1] *Standards for Educational and Psychological Tests.* Washington: American Psychological Association, Inc., 1974.

separate documents would be written for various types of readers (examples include a technical report for testing specialists and a summary description for parents and students).

Providing a Test Description for Lay Audiences

Has some means of communicating proficiency testing information to lay audiences been developed and disseminated?

A test description may be developed in the form of a brief brochure or pamphlet designed to communicate information about the proficiency test to various lay audiences. A test description should promote awareness of, and trust in, the proficiency test. Therefore, it should be written in nontechnical, pithy language. In districts with linguistic minorities, the test description should be translated into the primary languages of the minorities. For districts with students with many different native languages, a reasonable effort should be made to translate the description into the prominent languages.

Purposes and Uses of the Test

Have the purposes of, and uses for, proficiency testing been made clear to students and community groups?

A district may design its proficiency test for a number of mutually consistent uses, including promotion, remediation, and diploma sanction. These uses should be stated in the test description. The skills that the test measures (in the form of proficiency statements) should be listed along with information about the community's input in the selection of those proficiencies.

Suggested Test Content

Have appropriate audiences received information about the test content, sample items, and the test administration schedule?

Parents, students, and other interested parties will benefit from knowing the test content and the types of items comprised in the test. In addition to the list of proficiencies mentioned above, a sample item for each proficiency should be included in the test description. Some districts find it useful to give brief suggestions on how to take a test and how to study for a test.

Each district will have a unique testing schedule. A schedule for administration of proficiency tests should be part of the test description.

References

Handbook for Proficiency Assessment: Sacramento: California State Department of Education, 1979, Section VI, pp. 25-26.

Thorndike, R. L., and E. Hagen. *Measurement and Evaluation in Psychology and Education.* New York: John Wiley and Sons, Inc., 1977, p. 202.

Documenting the Test Administration and Scoring Procedures

Routine procedures for administering and scoring a proficiency test serve to standardize the instrument. The test must be administered the same way to all students for the results to be comparable. Without standardized procedures for test administration and scoring, the reliability and validity estimates obtained in the field testing may be misleading.

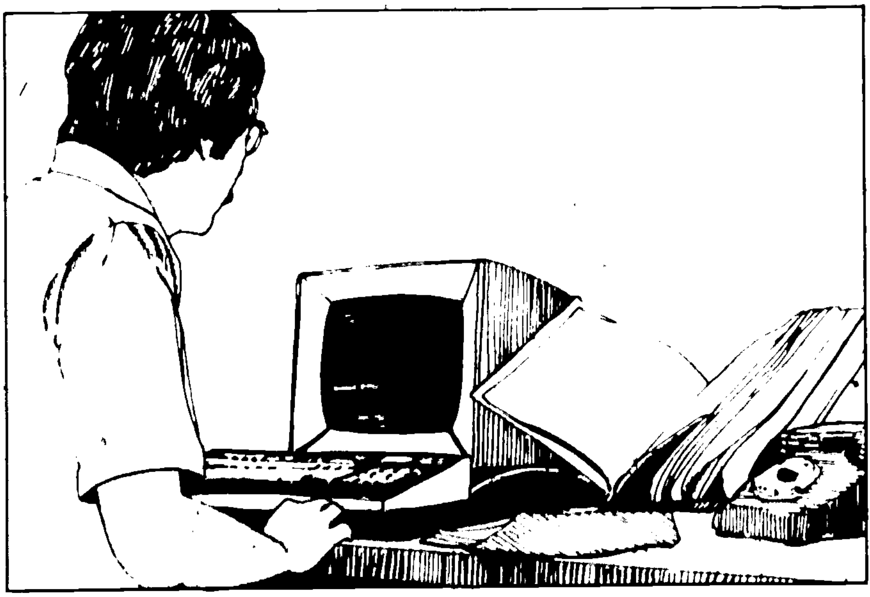
Administration Procedures

The directions to those administering the proficiency test should be distributed in written form to school administrators as well as to those who administer the test. In addition, those who administer the test should receive training in following the test administration procedures. Special training should be given to those giving individualized, make-up, or alternative modes of proficiency assessment.

Have complete, standardized test administration directions been prepared for examiners and examinees?

Complete Script of Directions

For a test to be standardized, it must include complete directions for both the examiners and examinees. The directions should include procedures for test administration and directions to be read verbatim to the examinees. The directions to the examinees should include the purpose for, and uses of, proficiency testing. (The diploma sanction should be made explicit without generating undue anxiety.) Also, sample items should be included, and practice in marking answer sheets should be provided.



Test Setting and Conditions

Do the test administration directions include information on the testing time and setting?

Information about the proper test setting and conditions should be included as part of the test administration package. For example, some districts stipulate the day of the week, time, and place for administering proficiency tests. The circumstances under which assistance can be provided to students and the limits of such assistance should be part of the directions for administering the test. Any time limits should also be specified, in advance, for the examiners and the examinees.

Scoring Procedures

Have all scoring procedures been standardized and verified?

For many proficiency tests, machine scorable answer sheets are used. In these cases students should be directed to erase stray marks completely. A clerk (or proctor) should double-check the answer sheets for accurate student identification and correct marks. When responses are keypunched for computing, they should be verified. If answers are to be hand-scored, scorers should receive appropriate training and practice. If subjective ratings are to be made, as in the direct assessment of writing, studies of interrater reliability (that is, concordance among judges) should be conducted.

References

Baker, F. B. "Automation of Test Scoring, Reporting and Analyses," in *Educational Measurement*. Edited by R. L. Thorndike. Washington: American Council on Education, 1971, pp. 202—34.

Clemans, W. V. "Test Administration," in *Educational Measurement*. Edited by R. L. Thorndike. Washington: American Council on Education, 1971, pp. 188 201.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section V, pp. 14—21.

Documenting the Test Construction Process

Documenting test construction includes recording information about how proficiency statements and test items were developed and how each is linked to the local curriculum and instructional practices. Documenting the test construction process also includes collecting and reporting information on field testing and test revision. The time line and staffing for these activities are normally described in this section of the documentation information. Such information facilitates subsequent test analysis and revision. For example, to incorporate curriculum changes in the proficiency test, those revising the test simply need to refer to the proficiency statements and item specifications and to revise the affected test items accordingly.

Development Procedures

The procedures followed in developing the initial draft of the proficiency test should be recorded. (For a complete description of necessary test development practices, see the chapter on test construction.) The methods used to obtain community input on proficiency statements and item specifications are especially important. Any time lines or planning documents related to test development should be cited.

Have test development procedures been recorded?

Test Revision

Documenting the test revision process involves keeping track of the item data and describing the procedures used in making revisions. As mentioned earlier, many persons involved in the initial test development effort fail to realize the need for continual test refinement. Updating proficiency tests from year to year allows minor psychometric flaws to be eliminated and increases congruence between the required skills and the curriculum. The test revision documentation should include specification of the data and other materials that were used in the revision process as well as the names of those persons who were involved, including their professional qualifications. Revisions in the proficiency test should be consistent with curriculum changes and should be reflected in the test manual. The dates of any revisions should be noted. The effects of changes in test content or the comparability of results across years should also be assessed and reported.

Have test revisions and the rationale for changes been documented?

Reference

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section II, pp. 105 - 9.

Documenting the Test Validation Process

One aspect of test validation involves assembling persuasive data and arguments that a test accurately measures what it is intended to measure. Another is that the data provide reasonable and useful information for making decisions about students' instructional programs. Both expert judgment and empirical data must be used in making decisions about the appropriateness of a test for its intended purpose. The process must be documented so that district staff have a record of how various decisions were made. The data collected via field testing are crucial for decisions about item analysis, bias review, and test validity and reliability. Each of these topics is discussed in greater detail in the chapter on test validation.

Field Testing

Complete documentation of the field testing procedures is important for interpreting the results of subsequent validation studies (for

Have field test methods and results been listed?

example, bias and reliability studies). Documentation of the field test administration procedures should include identification of the form of the test used and how it was administered. The directions, time limits, physical setting, and test administrator should be identified. The sample of students to whom the test was administered should be described in terms of the number and types of subgroups represented. Moreover, demographic and ability data should be recorded for further analyses.

Item Analysis

Have item analysis procedures and results been fully described?

Documentation of item analyses should be similar to that for bias review, especially if the same field test data are used for both. Documentation of the item analyses should include a list of statistical treatments, a description of how they were used in the context of proficiency assessment, and supporting rationale for each procedure employed. The results of the item analyses should be recorded along with procedures for subsequent revision or deletion. A list of items that require revision or deletion and other item analysis material should also be kept.

Bias Review

Documentation of the bias review procedures and results should include the names of the reviewers, their affiliations, and their demographic characteristics. The directions to the reviewers and any forms they used to record their comments about items should be kept. Suggestions for item revisions should be recorded and forwarded to the test development staff. Copies of the revised items should be filed with the other review materials described in this chapter.

Have the procedures and personnel involved in the subjective bias review been documented?

Statistical approaches used for identifying biased items should also be documented. The field testing procedures used to collect data for quantitative bias reviews need not be repeated if they are described in other documentation. The organization of field test data and their statistical treatments for detecting bias should be listed. The treatment of biased items should be noted, and the procedures for treating items identified as biased in the content review only or in the statistical analysis only should be documented.

Have statistical bias review techniques and results been recorded?

Validity and Reliability Estimates

No test documentation is complete without evidence of validity and reliability. Preliminary estimates should be made after the field testing has been completed and should be revised each year the proficiency test is used. This is especially important as the curriculum, instructional methods, test items, and student population change.

Are validity and reliability estimates available for each test administration?

The major steps in documenting validity and reliability include describing the data collection effort on which the estimates were based, listing the statistical and judgmental methods used to provide evidence of validity and reliability, and supporting the claims that a proficiency test is valuable for making decisions about students.

References

Henrysson, S. "Gathering, Analyzing and Using Data on Test Items," in *Educational Measurement* (Second edition). Edited by R. L. Thorndike. Washington: American Council on Education, 1971, pp. 130 59.

Handbook for Proficiency Assessment. Sacramento: California State Department of Education, 1979, Section VI, pp. 43 47.

Glossary

Articulation is the process of coordinating segments of an instructional program to provide for continuity between schools or levels.

Bias results when some facet of the test or of the test administration procedures distorts a subgroup's true performance level.

Curricular validity is a measure of how well test items represent the objectives of the curriculum.

Instructional validity is a measure of whether school districts are providing students with instruction in the knowledge and skills measured by the test.

Item specifications are detailed descriptions of the skills to be tested. They provide a "blueprint" for constructing test items to assess students' skill acquisition.

Proficiency standards describe skills and the minimum levels of performance at which the student is expected to perform.

Reliability refers to (1) the degree to which the results of testing with one sample of items matches the results of testing with another sample of items at a later time; or (2) the degree to which the results from multiple administrations of a sample of items at different times are similar.

Validity refers to the effectiveness of a test instrument in representing the content domain that the user is interested in.

Guidelines Checklist

TEST CONSTRUCTION

	Yes	No	Don't know
1. Have the uses of the proficiency tests been identified and agreed upon? See page 3.	□	□	□
2. Have procedures been established for testing at the grade levels specified by law? See page 3.	□	□	□
3. Have policies been established with regard to schedules for testing and retesting? See page 3.	□	□	□
4. Have district staff identified local resources for use in the various proficiency assessment activities? See page 3.	□	□	□

Developing Proficiency Standards

	Yes	No	Don't know
5. Has the local community been involved in developing proficiency standards? See page 4.	□	□	□
6. Has community involvement in proficiency assessment reflected the demographic makeup of the district? See page 4.	□	□	□

	Yes	No	Don't know
7. Have the proficiency standards of elementary and secondary schools been articulated? See page 4.	□	□	□
8. Do proficiency standards include a statement of the skill being assessed, how the skill will be assessed, and the level of performance required? See page 5.	□	□	□
9. Have a similar style and format been utilized for the proficiency standards in the three required content areas (reading comprehension, writing, and computation)? See page 5.	□	□	□
10. Have proficiency standards been reviewed periodically for curricular validity and instructional validity? See page 6.	□	□	□

Developing Item Specifications

	Yes	No	Don't know
11. Have item specifications been used in the development of test items? See page 7.	□	□	□
12. Do item specifications include descriptions of the manner in which each skill is to be assessed (i.e., performance mode)? See page 9.	□	□	□

	Yes	No	Don't know
13. Are item stem characteristics included in the item specifications? See page 9.	□	□	□
14. Are distracter characteristics described in the item specifications? See page 9.	□	□	□
15. Are sample items included in the item specifications? See page 10.	□	□	□

Writing Items

	Yes	No	Don't know
16. Have enough staff been assigned to item writing? See page 10.	□	□	□
17. Have enough test items been written to allow the creation of multiple test forms? See page 10.	□	□	□
18. Do all test items conform to item-writing rules? See page 11.	□	□	□

Pretesting and Revising Items

	Yes	No	Don't know
19. Do all test items conform to item specifications? See page 12.	□	□	□

	Yes	No	Don't know
20. Have all items on the proficiency test been reviewed by teachers and other educational specialists not involved in developing the items? See page 13.	□	□	□
21. Have all items been reviewed for bias and irrelevant difficulty? See page 13.	□	□	□
22. Have all items been pretested on a small but representative group of students? See page 13.	□	□	□
23. Is the testing time commensurate with available resources, staff/student time, and the purpose(s) of proficiency assessment? See page 16.	□	□	□
24. Have multiple forms of the proficiency test been developed? See page 17.	□	□	□
25. Do all test forms demonstrate curricular and instructional validity? See page 17.	□	□	□
26. Are all forms of proficiency tests comparable? See page 17.	□	□	□

	Yes	*No*	*Don't know*
27. Have the proficiency tests been kept secure? See page 18.	□	□	□
28. Have sample test items and test descriptions been shared with teachers without compromising test security and so that instruction is linked to assessment? See page 19.	□	□	□

TEST VALIDATION

Conducting Item Analyses

29. Have students in the field test sample been carefully selected? See page 20.	□	□	□
30. Have teachers categorized students in the field test as masters, nonmasters, or borderline students? See page 22.	□	□	□
31. Was the field test administered like an actual assessment? See page 23.	□	☑	□
32. Have p-values been computed and analyzed for various subgroups (e.g., by ethnicity, sex, and ability levels)? See page 23.	□	□	□

	Yes	*No*	*Don't know*
33. Have item discrimination indexes been used in determining item validity? See page 25.	□	□	□
34. Have content specialists reviewed each item, using item analyses to guide decisions about inclusion, exclusion, or revision. See page 26.	□	□	□
Setting Passing Scores			
35. Have district staff and community representatives been involved in defining levels of mastery (setting passing scores)? See page 26.	□	□	□
36. Were judgmental methods used in setting passing scores? See page 27.	□	□	□
37. Were empirical methods used in setting passing scores? See page 28.	□	□	□
38. Have passing scores been reviewed since they were first determined? See page 29.	□	□	□
39. Is additional academic information used to determine mastery status of students whose scores are near the passing score? See page 30.	□	□	□

	Yes	*No*	*Don't know*
Reviewing the Test for Bias			
40. Have both subjective and statistical bias reviews been conducted? See page 31.	□	□	□
41. Have subjective reviews for bias included representation from significant minority groups? See page 31.	□	□	□
42. Have statistical bias reviews been based on quality field test data? See page 31.	□	□	□
43. Have the bias review results been integrated and acted upon? See page 32.	□	□	□
Assessing the Validity of the Test			
44. Have validity indexes been computed for each content area? See page 34.	□	□	□
45. Does the validity coefficient show that the proficiency test accurately distinguishes masters from nonmasters? See page 34.	□	□	□

	Yes	No	Don't know
46. Have passing scores been reexamined or adjusted in light of validity considerations? See page 35.	□	□	□
47. Have the curricular validity and instructional validity of the proficiency test been checked? See page 35.	□	□	□

Assessing the Reliability of the Test

	Yes	No	Don't know
48. Have reliability studies been conducted for all proficiency tests? See page 36.	□	□	□
49. Was instruction relevant to the proficiency test avoided during the time between the two test administrations? See page 37.	□	□	□
50. Has the number of test items been reconsidered or adjusted to increase the reliability of the proficiency test? See page 38.	□	□	□

TEST DOCUMENTATION

	Yes	No	Don't know
51. Has test documentation been addressed over the course of proficiency test development and field testing? See page 41.	□	□	□

Providing a Test Description for Lay Audiences

Yes *No* *Don't know*

52. Has some means of communicating proficiency testing information to lay audiences been developed and disseminated? See page 42. □ □ □

53. Have the purposes of, and uses for, proficiency testing been made clear to students and community groups? See page 42. □ □ □

54. Have appropriate audiences received information about the test content, sample items, and the test administration schedule? See page 42. □ □ □

Documenting the Test Administration and Scoring Procedures

55. Have complete, standardized test administration directions been prepared for examiners and examinees? See page 43. □ □ □

56. Do the test administration directions include information on the testing time and setting? See page 44. □ □ □

57. Have all scoring procedures been standardized and verified? See page 44. □ □ □

Documenting the Test Construction Process

	Yes	*No*	*Don't know*
58. Have test development procedures been recorded? See page 45.	□	□	□
59. Have test revisions and the rationale for changes been documented? See page 45.	□	□	□

Documenting the Test Validation Process

	Yes	*No*	*Don't know*
60. Have field test methods and results been listed? See page 46.	□	□	□
61. Have item analysis procedures and results been fully described? See page 46.	□	□	□
62. Have the procedures and personnel involved in the subjective bias review been documented? See page 47.	□	□	□
63. Have statistical bias review techniques and results been recorded? See page 47.	□	□	□
64. Are validity and reliability estimates available for each test administration? See page 47.	□	□	□

User Questionnaire

The Office of Program Evaluation and Research is interested in your feedback on the clarity, quality, and utility of *Guidelines for Proficiency Tests.* Please answer the following questions, fold and staple the questionnaire, and mail it to the address indicated on the reverse side of the questionnaire. Your comments will be greatly appreciated. Thank you very much.

1. **Comments on the Test Construction Guidelines:** Please give your reactions to the chapter on test construction.

2. **Comments on the Test Validation Guidelines:** Please list your reactions to the technical section on test validation.

3. **Comments on the Test Documentation Guidelines:** Please comment on the appropriateness and completeness of the information on test documentation.

4. **Comments on the Examples and Illustrations:** Please indicate your reactions to the illustrations, figures, and examples included in this document. For example, did you understand the relationship of the illustrations to one another?

(continued)

5. **Comments on the Guidelines in the Margins:** The guidelines are presented in question form throughout the book. Did this assist you as a reader (or was it a distraction)?

6. **Comments on the Overall Ease of Utilization:** What suggestions would you have for making the guidelines easier to use?

7. **Workshops or Training:** Would you be interested in attending training sessions or workshops on the issues and procedures set forth in the guidelines? If so, which ones would you like to see highlighted?

Affix
Postage
Here

USER QUESTIONNAIRE
California State Department of Education
Office of Program Evaluation and Research
721 Capitol Mall—Fourth Floor
Sacramento, CA 95814