DOCUMENT RESUME

ED 312 284 TM 014 037

TITLE	Monitoring and Improving Testing and Evaluation Innovations Project. State Level Activity. Annual Report.
INSTITUTION	Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Nov 88
GRANT	OERI-G-86-0003
NOTE	85p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Agency Role; Check Lists; Construct Validity; Educational Assessment; Educational Testing; Equated Scores; *Evaluation Methods; *Program Proposals; Proposal Writing; Research Proposals; *State Programs; *Student Evaluation; Trend Analysis
IDENTIFIERS	Large Scale Programs; *Monitoring Improving Testing Eval Innov Project; *Requests for Proposals

ABSTRACT

Three papers and a sample outline of Requests for Proposals (RFPs) represent the work of the Monitoring and Improving Testing and Evaluation Innovations (MITEI) Project during 1988. The skeleton of a sample RFP outline for large-scale assessment was expanded to provide a checklist of the type of information that should be included in a RFP. The checklist presented is a guide to RFP writing. Revisions were made to two papers written by members of the Joint Task Force of the Center for Research on Evaluation, Standards, and Student Testing/National Council on Measurement in Education concerning critical technical issues in large-scale assessment. The first paper, "Issues To Be Considered in the Equating Portions of Requests for Proposals for Large-Scale Assessment Programs" (Richard M. Jaeger), discusses placing multiple test forms on the same scale to make them useful for comparing performances and examining trends (test equating). The second paper, "Issues To Be Considered in the Content Validity Portions of Request for Proposals for Large-Scale Assessment Programs" (Ronald K. Hambleton), discusses the construct validity information that should be included in a RFP. The third paper, "Report on the MITEI Project Panel Given at the NCSL/ECS Meeting," reviews the November 12, 1988 meeting in Annapolis (Maryland) of the Education Staff Network of the National Conference of State Legislatures (NCSL) and the Education Commission of the States (ECS). An overview of large-scale assessment programs throughout the United States is provided. Sample rating/review forms and summary sheets; the Annapolis meeting agenda; and the "Code of Fair Testing Practices in Education" are presented. (SLD)

Center for Research on Evaluation, Standards, and Student Testing

Deliverable - November 1988

Monitoring and Improving Testing and Evaluation Innovations Project State Level Activity

Annual Report

Study Director: Pam Aschbacher

Grant Number: OERI-G-86-0003

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

Table of Contents

	Page
Introduction	1
Sample Outline of Requests for Proposals for Large-Scale Assessment	4
Issues to be Considered in the Equating Portions of Requests for Proposals for Large-Scale Assessment	16
Issues to be Considered in the Content Validity Portions of Requests for Proposals for Large-Scale Assessment Programs	30
Report on the MITEI Project Panel Given at the NCSL/ECS Meeting	49

Project on Monitoring and Improving Testing and Evaluation Innovations

The papers included here represent the work of the MITEI Project during 1988. Over this past year, the joint CRESST/NCME Task Force on Large-Scale Assessment has worked on three tasks. First, we expanded the skeleton of a sample Request for Proposal (RFP) outline inherited from the original NCME Task Force on Model RFPs. Second, we worked on revisions of the two papers written by Task Force members last year on critical technical issues in large-scale assessment. And third, we presented a panel at a meeting of the Education Staff Network of the National Conference of State Legislatures (NCSL) and the Education Commission of the States (ECS). Each of these activities is described briefly below.

The "Sample Outline of Requests for Proposals for Large-Scale Assessment" was developed as a supplement to our handbook for state and large district testing directors, Improving Large-Scale Assessment. The outline is based on a compilation of many different, successful RFPs from across the country and combines features of RFPs for both test development and administration. It will provide a

checklist for testing directors during the RFP process.

"Issues to be Considered in the Equating Portions of Requests for Proposals for Large-Scale Assessment Programs," by Richard Jaeger, is the first of three revised papers on critical technical issues. In the revision, he has expanded the paper, drawn distinctions, and clarified some points to increase the paper's usefulness for testing directors. "Issues to be Considered in the Content Validity Portions of Requests for Proposals for Large-Scale Assessment Programs," by Ron Hambleton, has also been revised, extended, and appended with sample forms to use in establishing content validity evidence. We plan to issue shortly, both technical papers as additional supplements to our testing directors' handbook.

We have included here a summary of our meeting with the NCSL/ECS Education Staff Network as well as the agenda and several handouts. We presented background information to legislative staffers on testing and measurement, an overview of current policies and processes in the states, and a brief view of the future of testing. In addition, we discussed with legislative staffers the relationship of policy

to test quality and its consequences, with a focus on how to improve collaboration among policymakers, state departments of education, and others to improve assessment programs.

Sample Outline of Requests for Proposals for Large-Scale Assessment

The purpose of this Sample Outline of Requests for Proposals (RFP) for Large-Scale Assessment is to provide a checklist of the types of information you might include in your RFPs. It is based on a compilation of many different, successful RFPs from across the country and combines features of RFPs for both test development and administration. It is not meant to be prescriptive since each RFP is unique. Your own may include only some of these topics and may also include others as well. You may also prefer a different organization of the sections to the one presented here. Regardless of the content or organization that you use, you may find it helpful, as have many states, to require that all proposals use the same paragraph or section numbering system as that used in your RFP.

I. INTRODUCTORY INFORMATION

A. PURPOSE AND INTENT

1. Clear statement of purpose of test; rationale (e.g., legislation)

2. Content areas to be covered
3. Grade levels
4. Approximate number of students to be tested per grade level
5. Time of year tests are to be administered
6. Special considerations (e.g., bilingual or handicapped students to be tested)
7. Any tasks or subtasks to be bid separately

B. KEY DATES

1. During bid process
   a. Bidders' conference
   b. Bidders' inquiries
   c. Bids due
   d. Contract awarded
2. During contract period
   a. Scheduled start date
   b. Completion date

C. BIDDING INFORMATION

1. Issuing office and address, contact person and phone
2. Number of copies due
3. Bidders' conference
   a. Mandatory?

b. Place and time

c. Recorded?

d. If, when, and how minutes will be available

4. Questions and inquiries

a. How to ask (e.g., in writing only?)

b. Whom to ask

c. Responses shared with all?

5. Revisions to RFP

a. When issued

b. Who will receive revision information

6. Level of effort

a. Expected cost

b. Fixed and variable costs

c. Funding amount and schedule set by legislature

d. Contract awarded in whole or in part

7. Bonding

a. Performance bond required?

b. Bid bond required?

8. Subcontracting

a. Allowed?

b. Subject to approval

c. Information about subcontractor to be

provided

1) Company name, address, officers, contact person

2) Organization support and experience

3) References

d. Who is responsible for which tasks

9. Particular requirements of state (e.g., Equal Employment Opportunity (EEO), percent minority staff, favoritism to in-state companies)

D. CONTRACT INFORMATION

1. Project monitoring

a. Planning documents after contract is let

b. Progress reports

c. Project officers and assistants (state department of education [DOE] and contractor)

d. Technical advisory committee (e.g., who, when meet, functions)

e. Other advisory or oversight committees

f. Schedule of reviews and approval of materials (e.g., who, when, length of review period)

g. Late work: penalties, whom to contact

h. Extension: possible length, how to notify contractor, how contractor must respond

2. Prime contractor responsibilities

a. Proposal, RFP contents, and minutes from bidders' conference become part of any contract awarded as result of RFP

b. Can contractor assign or transfer responsibilities without state's/district's approval?

c. Conditions under which contract may be terminated

d. Period for which accounting records are to be kept and made available

e. Effort required beyond scope of this RFP

1) Hearings, meetings, etc.

2) When, who, how decide when new contract needed

3) Costs (a part of contract or additional fee?)

3. Ownership of materials, data,

documentation: what belongs to state/district and what to contractor

4. Invoicing
   a. When rendered to a state/district
   b. When due and payable by state/district

E. PROPOSAL FORMAT AND CONTENT

1. Definition of "non-responsive" proposals
2. Contents
   a. Technical proposal
   b. Organization support and experience
      1) Personnel qualifications and loading
      2) Organizational capabilities: previous experience with projects of similar scope (give name of company of project officers)
      3) External consultants
      4) References
   c. Cost proposal
3. Format
   a. Proposal required to use same organizational structure as RFP?
   b. Specifications for cost proposal
      1) Under separate cover?

2) At the task level?

3) Standard format

F. EVALUATION OF PROPOSALS

1. Evaluation criteria

2. Point values or other indication of weight/importance

3. Open to creative approaches to particular problem?

4. Oral presentations

a. Mandatory/optional?

b. How request/assign date and time

II. BODY OF THE RFP

A. BACKGROUND INFORMATION

1. Relation of proposed assessment to related past, present, and future programs

2. Salient features of or quotes from relevant legislation

3. Important (e.g., legislated) dates

B. SCOPE OF WORK (Specify products and processes, let bidder recommend, or do both)

1. Specification of assessment type

a. Content area and grade levels to be assessed and when

b. Test objectives (e.g., provided or to be developed and how)

c. Assessment strategies (e.g., census testing, matrix sampling, duplex design)

d. Criterion-referenced, norm-referenced assessment, or both

e. Speed or power assessment

2. Composition

a. Item development (e.g., all original? number of items per objective)

b. Item review and editing: who, where, when, cost

c. Bias control: statistical and/or subjective review; who, when, what

d. Response mode(s) (e.g., essay, multiple choice, performance)

e. Relationship or role of state committees

f. Timelines

3. Trial testing

a. Pilot and field testing

1) Purpose

2) Contingent on review/approval

3) Supporting administrative procedures (e.g., training sessions)

4) Design (e.g., when, minimum number of responses per item, number of items per test form, minimum amount of test time per student, security)

5) Who decides on sampling plan and selects schools (DOE or contractor)

b. Contacts with schools

1) Liaisons

2) Who administers tests (DOE, contractor, Local Education Agency)

4. Developmental analyses: what, when, design (RFP may specify particular procedures or request that bidder describe proposed procedures, rationale, and types of statistics to be obtained)

a. Item analysis

b. Calibrations

c. Reliability of test forms

d. Validity (e.g., content, construct, concurrent, predictive)

e. Demographic data desired

f. Procedure for setting critical scores (i.e., cut scores, standards)

g. Forms (number of equivalent or parallel)

h. Norming

i. Equating to other tests or forms (e.g., anchor form?)

j. Sampling of items

k. Scaling

5. Distribution of pretests and final form

a. School-year timing

b. Delivery and return (who, when, where, number, overage, whom to contact for shortages and problems)

c. Packaging

d. Security

6. Data collection

a. Registration of examinees (if required)

b. Test administration

c. Training

d. Security

e. Quality control

7. Operational analyses

a. Scoring (formulas or plans)

b. Data processing

1) Data cleanup

2) Documentation

3) Hardware

4) Software

5) Required turnaround

8. Deliverables

a. Planning document (after contract let)

b. Reports (progress and final)

c. Tests

d. Manuals

1) Test administration

2) Interpretation

3) Technical

e. Training materials

f. Computer tapes

9. Reporting

a. Audiences

b. Formats

c. Publicity requirements

10. Cost proposal (note: RFPs may require that this be in the body of the proposal or in a separate document)

a. Organization

1) Budget at the task level?

2) Summary

b. Standard format

III. LICENSING, COMPLIANCE, CERTIFICATION, AND AFFIRMATION STATEMENT

Issues to be Considered in the Equating Portions of Requests for Proposals for Large-Scale Assessment Programs

Richard M. Jaeger
University of North Carolina at Greensboro

Because of test security problems and the evolution of school curricula, large-scale assessment programs require the creation of multiple forms of tests. For a variety of reasons--such as ensuring that each examinee has an equal opportunity to evidence his or her achievement, or a desire to examine growth or other temporal trends in the average achievement of students in schools or school systems--it is essential that multiple forms of tests used in large-scale assessments be placed on the same score scale. The process used to place multiple test forms on the same scale (and thus make the forms interchangeable, useful for comparing the performances of examinees who are tested with different test forms, and useful for examining trends in average student achievement) is termed test equating.

Developments in measurement theory and advances in computer technology and statistical software over the past 20 years have made routine equating of

multiple test forms far more feasible than was the case several decades ago. In addition, the development of mathematical models that provide specific descriptions of examinees' performances on test items has greatly increased the range of available test equating procedures. However, these models are based on strong assumptions and provide accurate and durable equating only if their assumptions are met.

Strictly speaking, tests that are to be equated must be psychometrically parallel. Frederic Lord (1980) has noted that two tests are parallel, and thus capable of being equated, only if it is a point of indifference to any examinee which test he or she completes. Although the score scales of any two measures can be made to appear the same (through a process called calibration), the process will not result in equating unless the measures are parallel. To illustrate this point, consider two contrived examples.

First, suppose you were to weigh two random samples of adult men. The first sample is weighed on a scale that measures in English units (pounds), and the second sample is weighed on a scale that measures

in metric units (kilograms). Suppose also that the first scale had been adjusted so that it added one pound to every person's weight, whereas the second scale had been adjusted so that, on average, it showed correct weights. Weights produced by the two scales could easily be equated (placed on the same score scale). If the samples of men were large enough, the formula needed to convert weight on the scale that weighs in kilograms to the scale that weighs in pounds would be estimated correctly as follows:

Weight in Pounds = 1 + 2.2046(Weight in Kilograms).

The 1 appears in the formula because the scale that measures in pounds adds a pound to everyone's weight, and the 2.2046 appears in the formula because it is the number of pounds in one kilogram. Now suppose that you wanted to apply this equating formula to the weights of two samples of women, half of whom had been weighed on the English-unit scale and half of whom had been weighed on the metric-unit scale. The equating formula derived from the data on men's weights would produce perfectly comparable scores for the women, just as it did for the men, because the two

measurement instruments (the scales) measure the same variable and are thus parallel instruments. Only if measurement instruments (e.g., tests) are parallel, will the equating formula developed using one sample of examinees apply correctly to other samples or populations of examinees. Our second example illustrates the converse situation:

Suppose you had weighed all of the sampled men, using the scale that measures in pounds, and that you had then measured their heights in inches, using a tape measure. You could use the height and weight data for the men to develop a calibration formula that would convert the men's weights in pounds to the scale of their heights in inches. Any of several calibration methods could be used. The simplest approach would be to calculate the mean (M_W) and the standard deviation (S_W) of the men's weights and the mean (M_H) and the standard deviation (S_H) of their heights. These statistics would be used in the following conversion formula:

$$\text{Height} = (S_H/S_W)\ (\text{Weight} - M_W) + M_H.$$

This formula would put the weights of the men on the

same scale as their heights, in the sense that, on the new scale, the men's weights and heights would have the same mean (average value) and the same standard deviation. Since the distribution of weights and heights of men follow a bell-shaped curve (are approximately normally distributed) in the adult population, creating score scales that had the same mean and standard deviation would make the score scales comparable at every score value.

If you followed this process, you would have calibrated the scale (measuring weight) and the tape measure (measuring height) for the sample of adult men--the numbers these measurement instruments produced when applied to the sample of men would be on the same score scale. However, you would not have equated the scale and the tape measure because they measure different variables; that is, they are not parallel. To verify this conclusion, you would merely have to apply your calibration formula to the heights and weights of a sample of women. Since the relationship between height and weight is different for women than for men, the calibration formula for men would not produce converted heights for women that were anywhere near their actual heights. More to the

point, the mean of the height values produced by using the men's conversion formula would not be the same as the women's actual mean height, and the standard deviation of height values produced by using the men's conversion formula would not be the same as the actual standard deviation of women's heights. Not only would the conversion formula for men result in converted scores (heights) that were wrong for most individual women, but the average converted score would be wrong as well. Although this example is contrived, and admittedly extreme, it applies directly to two tests that measure different psychological functions, and are therefore not parallel.* The scales of such tests can be made comparable for a single sample of examinees by creating a conversion formula, but the tests cannot be equated. The conversion formula will not produce trustworthy score conversions for other samples or populations of examinees when the tests are not parallel, regardless of the test equating method used.

Test Equating Specifications for RFPs

This section contains recommendations on the test equating specifications that should be provided in requests for proposals (RFPs). The recommendations

are necessarily general because specifics depend on the nature of the test forms or tests to be equated, and the constraints that govern collection of data for equating.

Since the psychometric literature is replete with methods for equating tests (cf. Angoff, 1984; Petersen, Kolen, & Hoover, in press) and none has been demonstrated to be universally superior, RFPs should specify a particular equating procedure only if the issuing state strongly prefers that equating procedure. In the latter case, proposers should be permitted to specify use of an alternative equating procedure, provided the specification is supported by a thoroughly-developed rationale.

RFPs should include the following three sections pertaining to test equating: "Rationale," "Procedures," and "Evaluation," as described below.

Rationale for Test Equating

If prospective bidders are to respond appropriately and completely, they must be fully informed about the purposes of test equating in the context of the assessment program operated by the issuing agency. The RFP must contain a detailed narrative description of the the tests to be equated

and the state's objectives in requesting that tests be equated. Among several potential objectives, listed in order of increasing problems and difficulties, are the following:

a. equating psychometrically parallel, multiple forms of a test,

b. equating a slightly customized norm-referenced achievement test (a test that incorporates some new development of item content specifications or some new item formats, but with at least three-fourths of the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form,

c. equating a moderately customized norm-referenced achievement test (a test that incorporates new development of item content specifications or new item formats, but with at least half the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form,

d. equating an extensively customized norm-referenced achievement test (a test that incorporates

substantial new development of item content specifications or substantial use of new item formats, with less than half the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form,

e. equating a curriculum-tailored, criterion-referenced test to a nationally standardized norm-referenced test, and

f. placing multiple levels of a test intended for different grade levels or age levels of students on a continuous, longitudinally-interpretable scale.

Authors of RFPs should realize that the current state of measurement science does not support the use of test equating for purposes (b) through (f) listed above. As noted earlier, it is widely known that test equating is not robust when applied to (1) tests that differ substantially in content, (2) tests that differ substantially in difficulty or reliability, (3) tests that are targeted to groups that differ substantially in ability, and (4) tests that assess a multiplicity of constructs that are differentially sensitive to instruction. The greater the differences among tests

on any of these factors, the weaker will be the generalization of equating results to populations that differ in composition from the equating sample. If tests differ substantially in what they measure, the result of using equating procedures will be calibration, rather than equating, as described in the hypothetical example considered earlier.

Although previous research has shown that pre-equating of test items (purposefully selecting test items for a new form that are similar in content, format, and difficulty to items in the old form that is to be replaced) is generally not sufficient to ensure equivalent test forms in operational use; every attempt should be made to construct test forms that are as nearly parallel in content distribution and psychometric properties as is possible. Careful attention to content parallelism and psychometric parallelism should be required in RFPs that call for the development of multiple forms of assessment instruments.

Equating Procedures

RFPs should require that proposals include detailed discussion of the procedures to be used in equating tests or test forms to achieve each purpose

specified in the RFP. Among the procedures that should be discussed in bidders' proposals are the following:

a. the data-collection design to be used, including plans for sampling examinees and plans for the administration of tests or test forms to be equated;

b. the sizes and composition of samples of examinees to be used in the equating study, including specification of the sampling frames to be used, the sampling units to be used, and backup sampling to compensate for nonresponse; and

c. the analytic equating methods to be employed, including discussion of the use of anchor tests or items (if any), and the specific statistical procedures to be used in constructing a comparable score scale for all tests and forms to be equated.

The RFP should require that the proposal contain a detailed justification of the data-collection design, sampling procedures, and data-analytic methods proposed for each equating purpose, including reasons for selecting the proposed design and methods instead of viable alternatives.

Evaluation of the Test Equating

The RFP should require that the proposal contain a detailed discussion of the methods to be used to evaluate the quality of the equatings that result from the data collected and the analytic procedures employed. In particular, the proposal should describe methods that will be used to estimate the degree of random equating error overall, at the mean, and at various points on the score scale including values at or near any cut-off scores that the contracting state intends to use in classifying or selecting individuals on the basis of test scores. In situations where equating is to be applied to a sequence of tests over a period of years, methods to be used to estimate the resulting degree of scale drift should be described and justified.

The RFP should also require that the proposal include a description of procedures the prospective contractor will use to obtain an independent validation of the equating, so as to verify its accuracy and the appropriateness of all procedures used to collect and analyze equating data.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (in press). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan.

Footnote

*Parallel is used here to mean test forms that measure, within acceptable limits, the same psychological function. The operational definition of parallelism, according to Angoff (1984) is: "Two tests may be considered parallel if, after conversion to the same scale, their means, standard deviations, and correlations with any and all outside criteria are equal." It is the last requirement that would be violated in the second contrived example (conversion of weights to heights) cited earlier.

Issues to be Considered in the Content Validity Portions of Request for Proposals for Large-Scale Assessment Programs

Ronald K. Hambleton
University of Massachusetts at Amherst

According to the AERA, APA, and NCME Standards for Educational and Psychological Testing (1985), content validity evidence requires reviewers to "assess the degree to which the sample of items, tasks, or questions on a test are representative of some defined domain of content" (p. 10). Expert judgment is the main mode of investigation of a test's content validity (Messick, 1988). In assessing content validity, test content is matched to the content specifications for the test.

In preparing content validity specifications for a Request for Proposal (RFP), the RFP writer has the choice of (1) asking bidders for a content validation plan, or (2) providing details of the types and nature of content validity evidence which are of interest. Four categories of content validity evidence are typically needed to support the uses of tests in large-scale assessments:

a. Objective Representativeness - Are the

objectives that are selected for inclusion in the test, representative of the objectives included in the domain of content of interest? For competency tests, normally the domain of content of interest is based upon a state curriculum or an agreed upon set of state objectives. The objectives themselves are often reviewed for appropriateness by a committee. Appropriateness can be assessed by judging how well the set of selected objectives covers the most important parts of the state's objectives or provides an adequate sampling of the full set of objectives. In the case of professional exams, the domain of content of interest may be based upon the results from job analyses or role delineation studies. Another possibility is that the content is based on a review of college curricula in required courses.

b. Item Representativeness - Are the items measuring each objective in the test, representative of the domain of content defined by the objective? To address this category, well-developed objectives such as those that highlight a model test item, content specifications, and distractor specifications (with multiple-choice items), are commonly used (e.g., see Popham, 1978). The set of test items can be judged

for their representativeness by asking reviewers to comment on how well the set covers the full domain of items spanned by the item specifications for the objective.

c. Item-Objective Congruence - Is the item a valid indicator of proficiency of the objective to which it is matched? Does successful performance on the test item require the same cognitive processes as those specified in the objective the item was prepared to measure? Measurement specialists can be especially helpful here. Unlike (b), which focuses on the assessment of sets of test items, (c) refers to the evaluation of individual test items.

d. Technical Adequacy of Items - Do the items satisfy standard item writing principles? Are the chosen item formats appropriate to permit valid assessments of the objectives of interest? Measurement specialists are well-qualified to comment on the suitability of the item formats. In some cases, empirical evidence would be desirable.

It is common to address the four categories of evidence using rating forms. Four examples from Hambleton (1984) are provided in Appendices A, B, C, and D. Interested readers are referred to Hambleton

(1984) for more information about these categories of content validity evidence and approaches for addressing the categories.

In preparing the content validity section of an RFP, the point must be made with prospective bidders that when building a test, amassing content validity evidence should not be viewed as a one-shot activity carried out at the completion of the test development process. Rather, content validity evidence should be compiled throughout the test development process and used in a timely way to make adjustments to the items in the test and items that are selected. Content validity evidence should be collected and used to guide the test development process at several important places. Some important places and appropriate questions to ask at each place follow:

1. At the item development stage, are the items representative of the domains of content they were intended to measure? Is each item technically sound? Is there evidence of item-objective congruence? When the answer to one or more of the questions is no, revisions can be made to the test items, or, in some cases, they can be discarded.

2. At the item tryout stages, is there evidence of

technical adequacy of items as reflected by the results from an item analysis? Comments from the field may also be useful.

3. At the final test development stage, are the topics, sub-topics, or objectives that have been selected for inclusion in the test, representative of the domain of content of interest? If not, new content selections can be made. Similarly, item representativeness with respect to each objective should be assessed at this stage.

4. At the final test development stage, were content validity considerations used in test development? How? And what evidence is there concerning the content validity of the test? Documentation of content validity is handled at this stage.

At each stage in the test development process, content validity evidence can guide the item writing process (where are items needed to meet needs?), item-writing training, and item selection.

A few additional points concerning content validity studies follow:

a. Representativeness means assessing the more important or critical objectives, and reflecting the

proportional size of the domains of content for objectives. In other words, for the representativeness criterion to be met in content validity studies, objectives which are more important or broader in scope than others need to be emphasized in test construction.

b. Judging item or objective representativeness may involve stratifying the domain of content prior to obtaining the reviewers' ratings. For example, in organizing a set of mathematics objectives, categories such as "computations," "measurement," "geometry," and "problem solving" could be useful for stratifying the objectives, prior to evaluating the representativeness of the set selected for inclusion in the test.

c. Content validity studies are techincal in nature, but the evidence can also meet political agendas as well. Designers must therefore seek out not only groups who can comment on content validity concerns, but also groups who are apt to raise concerns about the test if they have not had the opportunity to review and influence the choice of test content early in the test development process.

d. Minority representation on item review committees is particularly important in conducting

meaningful content validity studies. Therefore the RFP should make this point.

e. On some occasions, the number of test items may be too large for judges to review in the time available to complete the work. (There is also a practical limit on the number of test items that judges are willing to review.) On such occasions, a sampling plan must be developed to insure that each test item is reviewed by an acceptable number of judges. Obviously, more judges will be needed when the number of items to review is large.

f. In the early stages of the test development process, judges should be encouraged to offer editorial changes to test items when they see shortcomings. At the final stages, editorial changes may be less useful because the proposed changes would need to be reviewed, and time may not be available to carry out these reviews. Less than ideal items can be withheld from the test and reviewed again later for inclusion in a future form of the test.

g. The composition of review committees should be given considerable attention. Technical as well as political considerations must be addressed in the selection of reviewers for committees.

Possible details to request from prospective contractors in an RFP include proposed methods for selection and training of judges or reviewers, the number of judges to be used, the intended review process and sample rating forms, methods for resolving conflicts, intended data analyses, and approaches for reporting and using content validity data. These details will be addressed again in the next section.

Information Needed in an RFP

A well-written RFP should address six parts of a content validity study:

1. Ask for the types of content validity information that bidders feel are needed and why. Alternately, the state may wish to tell prospective bidders the nature and/or scope of the content validity studies they want.

2. Ask for details on the group or groups of persons who will be involved in the item and objective review tasks, along with desired numbers, and how persons will be selected and by whom.

3. Ask for details on the nature and amount of training for reviewers.

4. Ask for examples of item rating forms and approaches for data analysis and reporting.

5. Ask for details on the timing of content validity studies (in relation to the stages of test development) and how the available data will be reported and used.

6. Ask for details on analysis of content validity data.

Of course, a prior question before writing the content validity phase of the RFP is for the state to review its own resources (available time and expertise) to determine its role in the content validity process. The state may vary its involvement from essentially none (except observing the content validity meetings) to total involvement. State departments of education normally have the technical knowledge on staff to carry out content validity studies without assistance from contractors. Seldom, however, do the departments have sufficient numbers of staff and the time to direct the work themselves. Assuming sufficient resources, the main argument against total state involvement is the question of conflict of interest. Some might argue that a state department of education has too much at stake to identify a test as lacking in content validity--the state's judgment in selecting a competent contractor would be questioned, and

relations with the contractor would become very difficult. On the other hand the contractor may not be the best agency either. Contractors know the test best, but they have the most to gain from a positive review. It is hard to imagine a contractor who would design a study to show its test lacked content validity. An intermediate position might involve the formation of a neutral committee under the direction of (say) an independent consultant. Ben Shimberg, George Madaus, and others have called for the formation of an independent auditing agency that could conduct validation studies which would include content validity evidence in the scope of their work.

Also, it is important for state departments of education to insure that a contractor schedules the collection of content validity evidence at a time in the test development process when changes to the test can still be made. Normally, this time would be (1) following the item writing phase, (2) following the pilot-testing, and (3) following the subsequent construction of the test but prior to printing the test.

To this point in the report, we have described the content validity evidence that is needed during

the test development process. On some occasions, an "off-the-shelf" test may be proposed for use in a large-scale state assessment (e.g., selecting one of the major standardized achievement tests may be of interest). Here the review task shifts to judging how well the test content matches the state's objectives for assessment and the intended curriculum and instruction. Again, bidders need to be instructed to provide complete details on their plan for reviewing test items and for making a final test selection.

Additional Research and Development Issues

At least four aspects of content validity studies require additional research:

1. Guidelines for helping to decide when a sufficient amount of content validity evidence to support the intended use of the test scores has been collected would be helpful (e.g., see Smith, 1985). The particular test use and the feasibility of collecting the criterion data are important considerations.

2. Guidelines for documenting (reporting) content validity evidence would be helpful.

3. More research on the actual procedures for carrying out the four types of analyses described

above are needed. Content validity evidence is greatly valued, but the process of collecting the relevant data, unlike the standard-setting problem for competency tests, for example, appears to be understudied.

4. Extensions to the methods proposed in this report for collecting content validity evidence are needed to handle subjective item formats such as performance items (e.g., writing assessments).

References

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

Hambleton, R. K. (1984). Validating the test scores. In R. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore, MD: Johns Hopkins University Press.

Messick, S. (1988). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Smith, I. L. (1985). Content validity study of the AASPB item bank. Professional Practice of Psychology, 6, 233-250.

Appendix A

An Example of a Judge's Item Rating Form

Item Content Review Form

Reviewer: _____ Date: _____ Content Area: _____

First, read carefully through the lists of domain specifications and test items. Next, please indicate how well you feel each item reflects the domain specification it was written to measure. Judge a test item solely on the basis of the match between its content and the content defined by the domain specification that the test item was prepared to measure. Please use the five-point rating scale shown below:

Poor	Fair	Good	Very Good	Excellent
1	2	3	4	5

Circle the number corresponding to your rating beside the test item number.

Objective	Test Item	Item Rating					Comments
1	2	1	2	3	4	5	
	7	1	2	3	4	5	
	14	1	2	3	4	5	
2	1	1	2	3	4	5	
	3	1	2	3	4	5	
	8	1	2	3	4	5	
	13	1	2	3	4	5	
3	4	1	2	3	4	5	
	6	1	2	3	4	5	
	12	1	2	3	4	5	
4	5	1	2	3	4	5	
	9	1	2	3	4	5	
	10	1	2	3	4	5	
	11	1	2	3	4	5	

Appendix B

An Example of a Judge's Summary Sheet for the Items/Objectives Matching Task

Items/Objectives Matching Task

Reviewer: _____ Date: _____ Content Area: _____

First, read carefully through the lists of domain specifications and test items. Your task is to indicate whether or not you feel each test item is a measure of *one* of the domain specifications. It is, if you feel examinee performance on the test item would provide an indication of an examinee's level of performance in a pool of test items measuring the domain specification. Beside each objective, write in the test item numbers corresponding to the test items that you feel measure the objective. In some instances, you may feel that items do not measure any of the available domain specifications. Write these test item numbers in the space provided at the bottom of the rating form.

Objective	Matching Test Items
1	
2	
3	
4	
No Matches	

Appendix C

Instructions for Using the Multiple-Choice Item Review Form

1. Obtain a copy of the objective and the test items written to measure it.

2. Place the objective number, your name, and today's date in the space provided at the top of the Item Review Form.

3. Place the numbers corresponding to the test items you will evaluate in the spaces provided near the top of the Item Review Form. The numbers should be in ascending order as you read from left to right. (This must be done if the processing of your data along with the data from many other reviewers is to be done quickly and with a minimum number of errors.)

4. Read the objective statement carefully.

5. Read the first test item carefully and answer the first 15 questions. Mark "/" for "yes"; mark "X" for "no"; and mark "?" if you are "unsure."

   The last question requires you to provide an overall evaluation of the test item as an indicator of the objective it was written to measure.

   There are five possible ratings:

   5 - Excellent
   4 - Very Good
   3 - Good
   2 - Fair
   1 - Poor

6. Write any comments or suggested wording changes on or beside the test item.

7. Repeat the rating task for each of the test items.

8. Staple your Item Review Form, objective, and copy of the test items together, and return to the coordinator.

Appendix D

An Example of a Technical Review Form for Items

- Item Review Form -
(Multiple Choice)

Objective No.:_____ Reviewer:_____ Date:_____

Test Item Characteristics (Mark "√" for Yes, "X" for No, and "?" for Unsure)	Test Item Numbers					
1. Is the readability level of the test item stem and answer choices suitable for the examinees being tested?						
2. Does the item stem describe a single problem for an examinee?						
3. Is the item stem free of ambiguities and/or irrelevant material?						
4. Is the content of the test item matched closely to the goal statement, objective, or task?						
5. Are all negatives underlined?						
6. Do the item stem and answer choices follow standard rules of punctuation, capitalization, and grammar?						
7. Are the answer choices arranged logically (if such an arrangement exists)?						
8. Is there one correct or clearly best answer?						
9. Is the placement of the correct answer made on a random basis?						
10. Are the answer choices free of irrelevant material?						
11. Are numbers or letters used to label the answer choices?						
12. Is any material provided in another test item that will provide a clue to the correct answer?						
13. When pictorials, tables, or figures are used, are they printed clearly and labelled correctly?						
14. Can the test item be answered by simple logic or common sense?						
15. a. Have words that give verbal clues to the correct answer such as: "always," "may," "none," "never," "all," "sometimes," "usually," "generally," "typically," etc. been avoided?						
b. Have repetitious words or expressions been removed from the answer choices?						
c. Will the distractors be plausible and appealing to examinees who do not know the correct answer?						
d. Are the answer choices of approximately the same length?						
e. Has the use of "all of the above" or "none of the above" as answer choices been avoided?						
f. Are four or five answer choices used?						
g. Have double negatives been avoided?						

	Test Item Numbers					
h. Have "clang" associations with the stem been avoided for the correct answer?						
i. Have distractors that mean the same thing or are opposites been avoided?						
j. Are the answer choices for an item similar in type, concept, and focus so that they are as homogeneous as possible?						
k. Is the correct answer stated at the same level of detail as the other answer choices?						
16. Disregarding any technical flaws which may exist in the test item (addressed by the first 25 questions), how well do you think the content of the test item matches with some part of the content defined by the objective? (Remember the possible ratings: 1=poor, 2=fair, 3=good, 4=very good, 5=excellent)						

Author Notes

The author is grateful to Richard Jaeger, Robert Linn, and Jim Popham for providing evaluative comments on an earlier draft of this report.

Report on the MITEI Project Panel Given
at the NCSL/ECS Meeting

Several members of the joint CRESST/NCME Task Force on Large-Scale Assessment from the MITEI Project served as panelists at a meeting on November 12, 1988, in Annapolis, MD, of the Education Staff Network of the National Conference of State Legislatures (NCSL) and the Education Commission of the States (ECS).

The Education Staff Network is responsible for studying state education issues, providing information to state legislatures about these issues and the methods being taken to deal with them, and facilitating dialogue between legislators and legislative staff from different states regarding the improvement of the educational enterprise.

The idea of our Task Force meeting with legislative staffers was conceived at our June, 1988 meeting. The purposes of our panel were to present to the legislative staffers some basic information on the technical requirements of educational testing and measurement, and to discuss with them how to improve the formation of testing policies.

Approximately twenty members of the Education Staff Network attended the meeting from such states as

Maryland, Pennsylvania, California, New Jersey, Colorado, Utah, Idaho, and Iowa. Members introduced themselves and reviewed the most pressing educational testing issues in their states. These issues included the following: accountability and school appraisal, incentives for school improvement being viewed as entitlements, the need for teacher proficiency testing, excessive testing driving the curriculum, the need for early identification of "weak" students for remediation, the need to assess the adequacy of home schooling, and the lack of training for teachers and administrators in giving increasingly sophisticated tests.

Ed Roeber, Supervisor of the Michigan Educational Assessment Program, presented an overview of the types of testing programs currently utilized across the states and distributed the most recent survey data on large-scale assessment programs collected for the Asso tion of State Assessment Programs. He underscored the importance of collaboration in defining reasonable goals and publicly specifying program objectives to help everyone put the tests in perspective.

Bob Linn, Professor of Education at the

University of Colorado and Co-Director of CRESST, presented some considerations in designing a testing program. Among the points he discussed were the following:

1. A test designed for one purpose may be dysfunctional or inadequate when used for another.

2. The meaning of test results can change with different uses of those results.

3. The degree of match between the test and curriculum is critical to both the results and their interpretation.

4. The use of multiple forms of a test can increase the breadth and depth of information available.

5. The choice of shelf, customized, or locally developed tests affects the nature of the results as well as the cost.

6. The level and nature of the thinking skills required by the test can affect instruction as much as the choice of content catgories on the test.

7. Multiple-choice items are efficient and effective in many areas but have their limits. Alternative measures may provide better instructional targets.

8. Norm-referenced and content-referenced interpretations each have strengths and weaknesses. Both need careful explanation and often become most useful in monitoring trends.

9. Global scores may satisfy accountability demands, but multiple scores related to specific content and process domains are needed in evaluating components of the curriculum.

10. Both judgmental reviews and statistical analyses are needed to avoid unintentional item bias and potentially offensive content.

11. Guidelines of acceptable practice are needed regarding appropriate test preparation. It is important to distinguish familiarization with test format, practice on similar tests, and practice on specific test items, as they differentially affect results and interpretations.

12. Security policies on test access can affect outcomes.

Eva Baker, Professor of Education at UCLA and Co-Director of CRESST, elaborated on current measurement research at CRESST and discussed the value of well-conceived and designed assessment tools for diagnosis and prescription. For example, she

discussed how multiple choice item distractors can be constructed to diagnose the types of errors made by students and prescribe appropriate instruction, and how multiple choice items may be created to assess higher order thinking skills. In addition, she discussed assessment of writing and critical thinking skills via essays written in response to visual or written stimuli (e.g., computer graphics displays or a speech from the Lincoln-Douglas debates).

Tom Kerins, Manager of Student Assessment and Program Evaluation in Illinois, distributed copies of the Code of Fair Testing Practices in Education, prepared by the Joint Committee on Testing Practices, a cooperative effort of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). He recommended that legislators and legislative staffers consider themselves test consumers, and take a more active stance toward the testing programs they legislate. He emphasized the importance of pressing the state departments of education for the information the legislators need to make informed assessment policies. He discussed the following points and illustrated them with specific

examples from Illinois' experience with its new reading and math tests. Legislators and staff should:

1. define the purpose for testing and the population to be tested, and then should select a test for that purpose and population, based on a thorough review of the information;

2. become familiar with how and when the test was developed and piloted;

3. examine the tests, directions, manuals, and score reports before selecting a test; and they should actually take the tests themselves to understand what will comprise the assessment program;

4. obtain information about the scales used in reporting results, the norms or comparisons groups used, and the limitations of the scores; and

5. evaluate the test development procedures for avoiding bias and insensitive language or content.

Pam Aschbacher briefly described the MITEI Project and its interest in facilitating the collaboration of state departments of education and state policymakers to create testing policies that work. She shared with the group the insights and viewpoints she had gained from talking with a number of state testing directors, department of education

staffers, measurement experts, and a few policymakers over the past several months; and she asked the audience to correct her perceptions and provide comments and suggestions. She reported that the input she had thus far received revealed a downward cycle of lack of trust and communication on "both sides of the river." The school reform movement has led policymakers to inquire whether new policies are working and whether American education is significantly better than it was. However, distrust of the educational system to provide unbiased information has led to a great deal of testing, and now even the test results are suspect for a variety of reasons. In order to create better policy and implement better assessment programs, four key factors were suggested as necessary: (1) proactive stance by both the department of education and the policymakers, (2) ongoing communication and trust on both sides to avoid premature commitment to inappropriate assessment plans, (3) clear delineation of purposes, and (4) recognition of the critical importance of developmental and technical requirements for good assessment.

Members of the Education Staff Network generally

agreed, and shared experiences to illustrate some of the difficulties of collaborating in a situation constrained by such problems as the relatively low priority of testing within educational legislation issues, restrictive requirements for test security purposes, and the lack of continuity of key legislators and staff.

Education Staff Network

Annapolis, Maryland *Agenda*

Saturday, November 12, 1988

1:00 p.m. **Welcome and Introduction**
Presiding: Ray Stark (ID), Education Task Force Vice Chairman

1:15 p.m. **Recent Events in the States: Staff Sharing**
Presiding: Chris Pipho, Education Commission of the States

1:25 p.m. **Student Testing: Recent State Actions**

Presenter: Ed Roeber, Supervisor of the Michigan Educational Assessment Program

1:35 p.m. **Tests and Measurement 101: A Primer for Legislative Staff**

- o What Makes a Good Test?
- o What Can Tests Tell Us?
- o What Are the Myths and Realities of Testing?

Moderator: Ray Stark
Presenters: Bob Linn, Co-Director--CRESST
Eva Baker, Co-Director--CRESST
Tom Kerins, Manager, Student Assessment and Program Evaluation, Illinois State Board of Education

2:15 p.m. **Dialogue Between Presenters and Participants**

3:00 p.m. **BREAK**

3:15 p.m. **A Panel Discussion Regarding *The Future of Testing***

- o Will We Ever Be Able to Test Higher Level Thinking Skills? Creativity?
- o New Technology--What Will It Mean for State Testing Programs?
- o What Are the Emerging Issues?

Moderator: Ray Stark
Presenters: Eva Baker, Co-Director--CRESST
Bob Linn, Co-Director--CRESST

EDUCATION STAFF NETWORK
Annapolis, Maryland

3:30 p.m. **Dialogue Between Presenters and Participants**

3:45 p.m. **Collaborating to Make Testing Policies That Work: Presentation and Discussion**

- o How Should State Departments of Education and Policymakers Interact?
- o How Can Effective Testing Policies Be Developed Via the Legislative Process?
- o How Does Policy Affect the Quality of Testing?

Moderator: Ray Stark
Presenters: Eva Baker, Co-Director--CRESST
Pam Aschbacher, Project Director--CRESST
Ed Roeber, Supervisor of the Michigan Educational Assessment Program

4:15 p.m. **Dialogue Between Presenters and Participants**

4:30 p.m. **Evaluation**

4:35 p.m. **Adjourn**

ABOUT THE PRESENTERS:

For over 15 years the *Center for the Study of Evaluation* (CSE) at UCLA has been at the forefront of efforts to improve the quality of education and learning in America through systematic evaluation practices. CSE has helped to pioneer valid and sensitive evaluation and testing techniques and has vigorously promoted the use of evaluation for reasoned decision making, seeking to ensure the best use of time and organizational resources.

In 1985 OERI funded the *Center for Research on Evaluation, Standards, and Student Testing* (CRESST) to address a broad array of research and development issues, serving the diverse interests of practitioners, researchers, and policymakers.

Bob Linn is Co-Director of CRESST and Professor of Education at the University of Colorado, Boulder.

Eva Baker is Director of CSE, Co-Director of CRESST, and Assistant Dean for Research at the Graduate School for Education at UCLA.

Pam Aschbacher has worked as Project Director at CRESST for over five years.

Ed Roeber is Supervisor of the Michigan Educational Assessment Program of the Michigan Department of Education.

Tom Kerins is Manager of the Illinois State Board of Education's Student Assessment and Program Evaluation.

Association of

State

Assessment

Programs

Survey of Large-Scale Assessment Programs

Fall 1988

Compiled by
Edward D. Roeber
Michigan Department of Education
Michigan Educational Assessment Program
P. O. Box 30008
Lansing, Michigan 48909

1. In the space below, please briefly describe your assessment/testing/ competency testing program(s). Include grades and subjects testing, how tests are developed, and what uses are made of the results.

Alabama: Basic Competency Tests (BCT) - grades 3, 6, 9, Alabama High School Graduation Exam (AHSGE) - grades 11 and 12. These tests are for acquisition of minimum skills. Skill deficiencies are identified and are to be remediated. The AHSGE is required for a high school diploma. Students must pass all three sections, reading, language and mathematics.

Stanford Achievement Test/Otis-Lennon School Ability Test - grades 1,2,4,5,7,8,10. Reports at student, school, system and state level are intended to be used for instructional planning and curriculum evaluation. The philosophy of all of our program is to use tests as tools for instructional improvement.

Alaska: We are proposing a 4-6-8 grade basic skills testing program (reading/math/lang. arts) using an "off the shelf" test. We will be doing a writing assessment pilot tryout in grade 8.

Alberta: Diploma Examinations Program consists of examinations administered annually in selected Grade 12 courses: English 30, English 33, Social Studies 30, Mathematics 30, Biology 30, Chemistry 30, Physics 30 and Langueet Literature 30. Results are used to certify individual student achievement which serve in part as a basis for university and (see attachment).

Arizona: Norm-referenced standardized achievement tests for all pupils in grades 2 through 22 and 1000 pupil sample in grades 1 and 12 in reading, grammar and mathematics. Writing assessment every 3 years of 1000 pupils each in grades 4, 8 and 11.

Arkansas: State developed criterion referenced tests in reading and math are administered to all students in grades 3, 6 and 8. Grades 6 and 8 are also tested in the subject areas of language arts, science and social studies. The MAT-6 is administered to all students in grades 4, 7 and 10.

Colorado: The Colorado Student Assessment Program tests statewide samples of students in a variety of learning areas to develop a state profile. Besides state results, results are provided to students, teachers, and schools. A schedule is attached.

Commonwealth of Northern Marianas: For Title VII Federal Grant, we use Language Assessment Scale Test (LAS I) to measure oval English, proficiency of all LEP students in grades K-3. For Statewide Assessment we use the California Achievement Test (CAT, Form C & D) in grades 1-12 using a random sample. To measure vernacular language proficiency we use a locally developed instrument.

Connecticut: See Attachment A

Delaware: By legislative mandate, Delaware is required to assess students in grades 1-8 and 11 using a nationally normed standardized achievement test. For the past five years, we have employed the Comprehensive Tests of Basic Skills (CTBS) testing students in the content areas of reading, language arts, and mathematics. In addition, science and social studies have been assessed in grade 11. Test results are available at the student, classroom, school, district, and state levels and are primarily used to identify individual and group weaknesses so that instruction can be improved and better targeted.

Florida: See Attachment

Georgia: The Georgia statewide testing program includes criterion-referenced tests (CRT's) in grades 1, 3, 5, 8, and 10 (reading and mathematics). Norm-referenced tests (NRT's) in grades 2, 4, 7, and 9 (reading, language skills, science, mathematics, and social studies). Also, a first-grade readiness test in kindergarten. CRT's for grades K, 2, and 4 are used at local system options.

Hawaii: Mandated Statewide Testing Program at Grades 3,6,8,10 using combination standardized achievement and criterion. referenced competency tests. Graduation testing grades 9-12 using criterion ref. test. We test basic skills as well as affective areas and oral, writing skills, student, remediation, class, school district, statewide evaluation.

Idaho: Comprehensive standardized achievement tests are administered at grades 6,8 and 11 with state developed direct writing assessments conducted at grades 8 and 11. The Department of Education actively promotes utilization of test results for comparative (local, state, national) purposes;tracking achievement trends across grade levels and over time; as supplemental information in assessing LEA curriculum and instructional practices, screening students, placement and advisement, supporting public relations efforts and identifying SDE consultation priorities.

Illinois: The Illinois Goal Assessment Program, when it is fully implemented in 1993, will assess all 3rd, 6th, 8th, and 11th grade students on state goals in language arts, math, science, social sciences, fine arts, and physical development and health. Assessment instruments are being developed collaboratively by the state education agency and educators. Results will be distributed to local education agencies, who will make . . . available to the press and public.

Indiana: Grades 1,2,3,6.8,9,11; subjects: English/language arts, mathematics, social studies, science. Tests are developed to measure student achievement of state educational proficiencies in cooperation with test contractor. English/language arts and mathematics results are used in grades 1,2,3,6,8 to make individual student decisions on remediation (summer) and promotion/retention; other uses.

Kansas: State developed criterion-referenced reading and mathematics tests-grades 2,4,6,8,10. Test results are to be used by districts to identify students reading remediation. Statewide results published annually. Individual students building, and district results reported to districts.

Kentucky: Beginning with the 1988-89 school year, Kentucky will move to the CTBS/4 to be administered in grades K,1,2,3,5,7,10. The test is published by CTB/McGraw-Hill covering reading, spelling, language, math, library skills, science and social studies. The department will score tests administered at other grades, but LEAs must purchase materials.

Louisiana: The Louisiana Educational Assessment Program includes state developed CRT's based on state language arts and mathematics curriculum standards (grades 3, 5, and 7), a graduation test, NRT component (grades 4, 6, and 9) and kindergarten developmental readiness screening program. CRT and NRT results will be used in school district progress profile programs. The high school CRT measures student competence in English language arts, mathematics, science, and social studies and is a graduation requirement. Results of the Kindergarten Screening Program are used for placement with the regular kindergarten instructional program.

Maryland: See attachment.

Massachusetts: Two testing programs: (1) Biennial assessment testing (modeled after NAEP) in reading, math,science and social studies at grades 4, 8 & 12. Matrix sampling permits building, district and state reports. (2) Basic skills testing every year of all students in grades 3,6, & 9 in reading, math and writing for purposes of identifying students in need of remedial assistance. Individual, building, district reports provided. Results used to target funds to low-performing schools.

Michigan: All fourth, seventh and tenth grades are tested in mathematics, reading and science. Tests in health education, social studies career development and writing are given on a voluntary, state-paid basis.

Minnesota: Statewide sampling of Essential Learner Outcomes with regional local districts utilization on this cycle. Tests are developed in-house and results are used by state and local policy makers.

Mississippi: Stanford achievement testing in grades K,1,2,4,6. Basic skills testing in grades 3,5,8. Functional literacy testing beginning in grade 11. Subject area testing in biology, algebra I and algebra II.

Missouri: The Missouri Assessment Program continues as it has for the last two years, that is, a representative sample of youngsters in grades 3,6,8 and 10 complete the Missouri Mastery and Achievement Test (MMAT). From those data a state report is developed and made public. Currently the MMATs are available for Reading/Language Arts/English, Mathematics, Science and Social Studies/Civics. Students receive information on their respective key skills, a standard score which reflects the comparable national percentile rank which is based on the most recent education of the Iowa Test of Basic Skills and the Test of achievement and Proficiency. The primary use of the results of the MMAT is to improve instruction and to assist districts in determining the efficiency of their curriculum

Montana: For the 1988 year, Montana does not have a statewide student assessment program. Student assessment is a local district option. A statewide survey indicates that all school districts have in place a local student assessment program.

Nebraska: We are still using the Neb. Assessment Battery of Essential Training Skills N-ABELS. It is a mastery based instruction instrument for local use. Participating schools pay for materials (cost).

New Jersey: New Jersey High School Proficiency Test of Reading, Mathematics, and Writing - administered to grade 9 students as a graduation requirement and used as one factor in monitoring school program quality - developed with the aid of state committees and a contractor.

New York: See attachment.

North Carolina: The testing program in North Carolina is as follows:

1. Grade 3 - normative tests in reading, language and mathematics; curriculum referenced testing for science and social studies; minimum skills diagnostic testing in reading, math and language for those scoring below the 25th national percentile.

2. Grade 6 & 8 tests the same area as in grade 3 but has an additional essay assessment.

3. a minimum competency test required for high school graduation is given at grade 10 and must be re-taken if failed initially.

4. Tests are given at the end of the course for the following courses: Algebra I & II, Biology, US History. Tests being developed this year include 9th grade English, Geometry, Physics and Chemistry.

North Dakota: It is recommended that schools administer standardized achievement tests to two grade levels for grades 1-6 and 1 at grade 7 and 8, and two at grade levels 9-12 for accreditation purposes which is voluntary. Tests are generally administered to grades 3, 5, 7, 9, and 11.

Ohio: Locally selected or developed tests in the areas of reading, English composition, and mathematics are administered at three grades in every district as part of the State Board adopted competency-based education (CBE) programs. Beginning in 1989-90, districts will administer standardized achievement and ability tests selected from a state-approved list in grades 4,6 & 8. Results will be collected, aggregated and reported by the state. Beginning in 1990-91, students will take tests in reading, writing, mathematics and citizenship. To earn a regular diploma in 1993-94, students must establish ninth grade proficiency in those four areas, as well as meet all other curriculum requirements. To earn one of the high-level diplomas, students must establish twelfth grade proficiency and meet other specified criteria.

Ontario: Multiple matrix sampling of provincial sample of schools in main subject areas of math, first language, and science. Five-year cycle rotating through various grades and subjects. Tests developed in Ministry of Education, drawing upon previously developed pools. Provides assessment of provincial levels of achievement. School boards (districts) can join in review process.

Oregon: Every other year, we test a sample of seventh grade students in reading, mathematics, and writing using state developed tests. The results are used to set state level targets for improvement. We also collect, annually, grades 3 and 5 norm-referenced test data from local districts. This information is used to provide general achievement status information and trend data.

Pennsylvania : Tests of Essential Learning Skills (Reading and Mathematics) presently at grades 3, 5, and 8 developed by teachers and educators from state committees. Used both to identify students in need of additional help and as a measure of districts' and schools' ability to provide students with these basic skills.

Puerto Pico: Spanish (2nd to 9th grade); Mathematics (2nd to 9th grade); English (4th to 9th grade); Social studies (3rd, 6th and 9th); Science (3rd, 6th & 9th). The tests are developed by subject specialists from the Evaluation Division and assisted by the Programs. The results are used to improve the quality of the academic programs and to assist policymakers and program managers in making more concise decisions.

Rhode Island: Basic skills achievement (MAT) at grades 3,6,8 & 10. Direct writing assessment (developed by RI teachers at grades 3,6. Health knowledge (developed by contract at grades 3,6,8,10. Physical fitness (MAHPRED Physical Best Program) at grades 3,6,8,10. Merit recognition, a voluntary grade 12 program (written and performance tests in over 20 subject areas). Merit tests are developed by RI committees.

Tennessee: See attachment 1.

Texas: The Texas Educational assessment of Minimum Skills (TEAMS) testing program assesses annually 1.5 million Texas public school students in Grades 1,3,5,7,9 and 11/12 with criterion-referenced tests in mathematics, reading, and writing. Students must demonstrate mastery of the 11/12 (exit level) test in order to receive a high school diploma. Test items are written by a commercial contractor according to objectives and measurement specifications developed by Texas educators. The TEAMS tests are designed to identify students needing remediation in the basic skills. Schools are required to offer remediation to these students.

Utah: Utah Statewide Educational Assessment Program collects data on a stratified random sample of 65 elementary schools and 30 high schools for each assessment period. Approx. 5,000 fifth graders and 3,000 eleventh graders are tested for the program. Areas measured include mathematics, reading, English, art and music achievement, academic self-concept, career exploration, peer relations and numerous other scales. Instruments used in the program include both standardized tests as well as state measures developed by the state specifically for the assessment program. Mathematics, reading and English are assessed with the Comprehensive Tests of Basic Skills, Form U.

Virginia: There are seven components, two optional and five mandatory or partially mandatory. The option components are readiness testing in grades K-1 and career assessment in grades 7-12. Kindergarten screening is required of underage children whose parents request that they be permitted to enter kindergarten early. Criterion-referenced testing, using either the state's Standards of Learning Program or an alternative, is requested in all grades in all subjects. Literacy testing is conducted in

Virginia (con't): reading, writing, mathematics in grade 6 (beginning in 1989-90; passing will become a requirement for promotion to ninth grade). The state's assessment includes ability testing in the fall of first grade and achievement testing in the spring of grades 4,8 & 11.

Virgin Islands: Student achievement testing--grades 3, 5, 7, 9, & 11--off the shelve - student progress and attainment. Preliminary scholastic achievement testing--grades 10 & 11--off the shelve - student scholastic progress, scholarships. Scholastic achievement testing--grade 12--off the shelve - college eligible.

Washington: Every student testing each October in grades 4-8-10. Currently using MAT6 basic battery (reading, math, language). Science and social studies are optional. Results used for federal programs needs assessment and allocation of state learning assistance funds.

West Virginia: CTBS/U grade 3,6,9,11
LOGAT grade 3,9
Writing Assessment 8,10
Instructional Improvement

Wisconsin: a) Voluntary CBT program; reading, math, language arts, usually at grades 3, 7, 10; results used for remediation, other uses optional; b) statewide 3rd grade reading test, beginning in 4/89, district/school comparisons; c) districts must test in reading, math, and language arts using locally-selected tests aligned with the curriculum; no district comparisons are made.

Wyoming: 1988 concurrent national assessment in reading, writing and civics - grades 4,8,12 - 20% of students.

2. How has your program changed during the last year? What changes do you foresee for the coming year?

Alabama:
- funding allowed grade 7 to be added back into the norm-referenced testing
- Basic Competencies for grades 3,6,9 are being revised and strengthened. New tests will e developed and administered in spring 1990.
- New graduation exams will begin to be developed based on the new competencies and the new assessments will begin to be given in fall, 1991.
- The test-selection process for a new norm-referenced instrument will take place in 1988-89 with plans to administer the new tests in spring, 1990.

Alaska: We will move from collecting and aggregating local test information to adopting a uniform test.

Alberta: Programs have been maintained during the last year. Development has been initiated (expected completion in 1992) of diagnostic materials in: Reading (gr 7-10), Writing (gr 7-10), Oral Proficiency (gr 7-10), Mathematics (Gr. 7-10).

Arizona: Testing of pupils in grades 1 and 12 is now optional. Districts may elect to test their pupils in these grades. Research & Development Unit and Essential Skills Unit will be developing and pilot testing criterion-referenced tests for essential skills in grades 3,8 and 12. We must also conduct various studies (see attached House Bill).

Arkansas: This is the first year that eighth grade students have had to pass our state competency test in order that they may be promoted to the ninth grade. Results of the 1988 Arkansas Minimum Performance Testing Program show that Arkansas students have improved in every subject area when compared with 198 results. No major changes are anticipated for the coming year.

Colorado: A shift from the initial "every-student" approach (1985-86) to acceptance of sampling approach; a shift from spring to fall testing and acceptance by districts of statewide testing.

Commonwealth of the Northern Marianas: The state used to assess students in grades 1-12. Last year this changed and only odd grades are tested annually (random sample).

Connecticut:
- No changes during 1987-88.
- New assessment program, titled the Common Core of Learning (CCL), being developed for 1990-91. This program will primarily be performance testing at grades 11 and 12 and will focus on intellectual challenges which require students to demonstrate "knowledge-in-use," inquiry and expression.

Delaware: The CTBS 1981 edition was last used in our Spring 1988 test administration. After the solicitation and evaluation of competitive bids, the Department has adopted the Stanford Achievement Tests (SAT 8) for statewide use in 1988 and 1990. In addition, we are planning to conduct one or more assessments on a sampling basis this year in response to the recently adopted Delaware Agenda for Education. The most likely candidates for assessment this year are Health and Physical Fitness.

Florida: See attachment.

Georgia: Implemented a readiness assessment for all students entering public school first grade.

Hawaii: Adoption of NAEP state/state comparisons as driving force for much of our future testing. No real changes in program last year.

Idaho: The sixth grade comprehensive achievement component was added this past year.

Illinois: The program is being implemented gradually. Census assessment began in April 1988 with reading (part of the language arts) and will be expanded to include math in April 1989.

Indiana: No, except that social studies and science will be added in the 1988-89 school year, and the writing sample will be administered in December (1988) instead of March (1989).

Kansas: Remediation element added in 1988 legislative session. State Board is currently in the process of developing a proposal for statewide testing after 1989. Current statutory requirement end in 1989. Proposal being considered calls for a testing program similar to the current program

Kentucky: Kentucky has dropped the combination of norm estimation and criterion referenced testing (Kentucky Essential Skills Test). The philosophy motivating the change in that the state can best provide solid normative tests, that LEAs can best provide objective skills based tests.

Louisiana: The state testing program has been revised and upgraded froma competency program to a program that is intended to measure grade-level academic skills. Full program implementation is scheduled for the 1988-89 school year.

Maryland: No major changes occurred last year, and no major changes are foreseen for the coming year.

Massachusetts: Improved reporting formats to serve multiple audiences. Performance assessment in science/math will be conducted in spring of 1989 with state-wide sample of 4th and 8th grade students. Analysis and reporting on open-ended questions used in 1988 assessment will be released late in year.

Michigan: Next year, science,plus voluntary testing will shift to grades 5,8 & 11.

Minnesota: The program has remained philosophically intact with various technical improvements. It is anticipated that the program will expand.

Mississippi: Basic skills testing was dropped from grade 11 Stanford Achievement testing will be in grades K,4,6 & 8.

Missouri: The testing program has remained the same across the last couple of years. During the last year and into the next two years, however, we are in the process of developing a pre-school, kindergarten and first grade testing instrument.

Montana: During 1988, the State Board of Public Education, at the direction of the 1987 Legislature, has put in place a policy on gathering student assessment information from local school districts. The Office of Public Instruction will begin collecting and aggregating information on standardized testing presently carried out in local school districts. The information will be collected for grades 3,8 & 11 in subject matter areas of language arts, mathematics, science, reading & social studies during the spring of 1989.

Nebraska: See attached recommendation to the State Board .

New Jersey: Much greater emphasis on test security in 1988. We expect the HSPT to be replaced by a more difficult grade 11 test by 1994.

New York: See attachment.

North Carolina:
1. We have deleted formal assessment at grades 1 & 2 and are piloting an informal assessment relying on teacher judgments.
2. We plan to add additional high school courses to our testing program.
3. We are determining the feasibility of using a screening test to reduce the number of students subject to the minimum competency test.

North Dakota: None. 1990--new standards may reduce testing at the secondary level.

Ohio: The State Board has adopted (1) an initial list of achievement and ability tests that meet all established criteria and (2) a set of rules governing the administration, grading and scoring of the tests. School districts have one year to select and use one of the approved tests at the designated grades. Work continues on the high school proficiency testing program.

Ontario: 1987-88 Pilot provincial services. Current science review more sophisticated; some use of non-objective questions. Future reviews to involve some observation of students. Also increased participation by school boards.

Oregon: Has not changed since last year. However, the governor is considering a plan that would test all students at grades 3, 5, 8, and 11 using state developed tests.

Pennsylvania: The program is in a major t. nsition to establish a new statewide comprehensive test ng program. Educational Quality Assessment has been ferred until it is combined with the testing of reading , d mathematics. The future program will probably include grades 2 to 10. The form of the new program is still not decided.

Puerto Rico: It has not changed. None.

Rhode Island: The writing assessment has expanded to include grade 6. The number of merit recognition subject areas continues to increase.

Tennessee: See attachment.

Texas: The scope of the program has not changed during the past year. In order to prepare students for a written composition requirement to be added to the 1990 exit level test, Grade 9 failing compositions will be rescored analytically beginn ng with the February 1989 test.

Utah: In addition to the Utah Statewide Educational assessment Program, the state of Utah has embarked on an ambitious program of developing measurement resources for use by Utah's school districts. The focus of this program is on the identification and development of item pools, as well as the development of end-of-level and end-of-course criterion-referenced tests. All of these instruments are referenced to Utah's K-12 core curriculum. Major projects are currently underway in building end-of-level tests for elementary reading, science, and math. End-of-course tests are being constructed for eleven secondary math courses and eight secondary science courses.

Virginia: Selected new NRTs for State Assessment Program (see #3). The requirement for Minimum Competency Testing has been eliminated. Literacy Testing Program for grade 6 (reading, writing, and math) added. Under the criteri -referenced testing program Standards of Learning material for science, physical education, art and music have been added.

Virgin Islands: Yes. Preliminary scholas ic achievement testing program will include sophomores.

Washington: No change from last year. We are planning to shift the emphasis at grades 8 & 10 to guidance and planning as well as program review.

West Virginia: Will change in 1990-91 to: a) Kindergarten Test for Readiness; b) CRT in Reading, Composition, Math gr. 1-4; c) NAEP Trial Assessment

Wisconsin: Termination of CTBS state sample testing and DPI-developed CRT testing. New third grade reading test, new district basic skills testing requirement, and participation in the NAEP state assessment program are new state directions.

Wyoming: No.

Some Considerations in Designing a Testing Program.

1. **Purposes:** A test that is good for one purpose may be inadequate or even dysfunctional when used for another.

2. **Uses:** The effects and the meaning of test results can change with changes in the uses that are made of test results.

3. **Alignment:** The degree of match or mismatch between the test and the curriculum is critical to both the results and their interpretation. The right degree of match depends on purpose.

4. **Breadth and Depth:** With a single test for all there is a tradeoff, but assessments with multiple forms can reduce problem.

5. **Shelf, Customized, or Local:** The choice between a publishers off-the-shelf test, a test customized by a publisher to local specifications, or a locally developed test affects the nature of results as well as cost.

6. **Process Demands:** The level and nature of the thinking skills required by the test can affect instruction as much as the choice of test content categories.

7. **Item Format:** Multiple-choice items are efficient and can be effective for many areas, but they have their limits and other alternatives may provide better instructional targets.

8. **Type of Score:** Norm-referenced and content-referenced interpretations each have strengths and weaknesses. Both need careful explanation and often become most useful in monitoring trends.

9. **Number of Scores:** Global scores may satisfy accountability demands, but multiple scores related to specific content and process domains are needed in evaluating components of the curriculum.

10. **Bias:** Judgmental reviews as well as statistical analyses are needed to avoid unintentional bias and potentially offensive content.

11. **Test Preparation:** Distinctions among familiarization with test format, practice on similar tests, and practice on the specific items of the test are critical to results and interpretations. Guidelines of acceptable practice are needed.

12. **Security:** Policies on test access can affect outcomes.

Education Staff Network
National Conference of State Legislatures

Annapolis, Maryland
November 12, 1988

Examples of Test Items

---

1. Which is worth the most?

   ____ 11 nickels
   ____ 6 dimes
   ____ 1 half dollar
   ____ I don't know

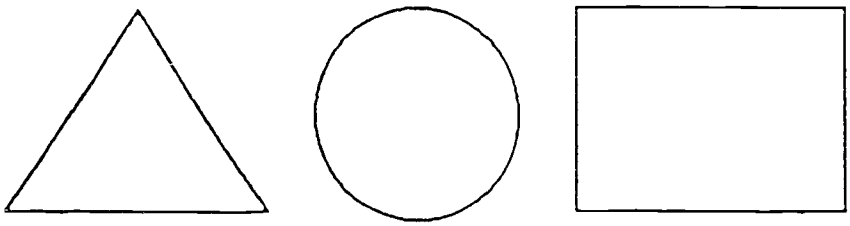
---

2. Suppose you have 10 coins and have at least one each of a quarter, a dime, a nickel, and a penny. What is the least amount of money you could have?

   ____ 41¢
   ____ 47¢
   ____ 50¢
   ____ 82¢

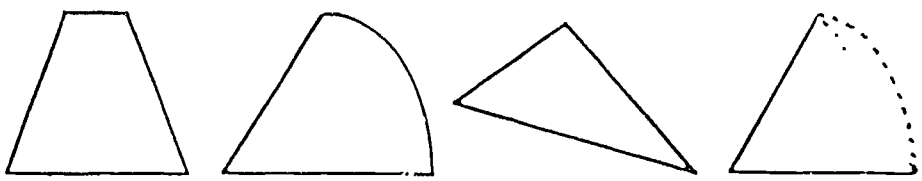
3.

Choose the Triangle



4.

Choose the Triangle



5. Which of the statements below best support the statement "Watching TV is damaging to children"?

A Eight-year-olds watch about 30 hours of TV a week.
B Children spend more time watching TV than working in school.
C By high school graduation, a student will have seen 800,000 hours of commercials.
D Violence among children increases in proportion to the amount of time they spend watching TV.

---

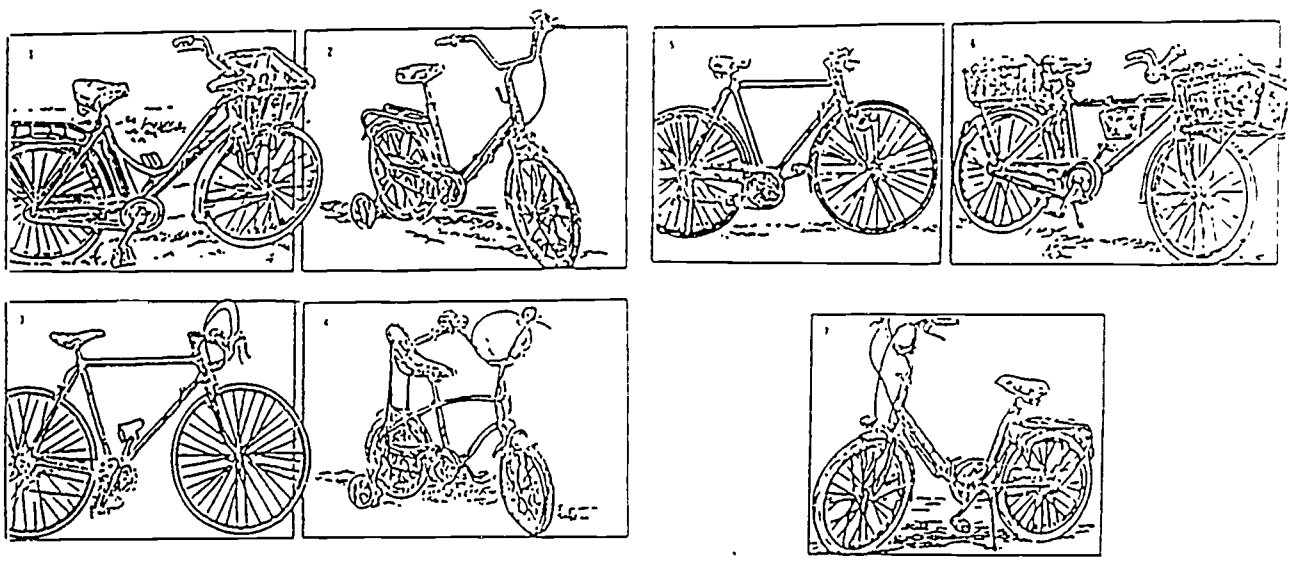
6. You arrive home after school to discover that no one is there. You expect that a member of your family will be home soon. You want to go and visit a friend for an hour or two, but you do not want your family to worry about where you are. Write a short message to a member of your family, explaining where you will be, who you will be with and when you will be home again.

---

7. Imagine that you visited your uncle who lives in another town. There you saw a fine collection of bicycles in a shop; in the pictures below you see some examples of these bicycles. The shop has bicycles for boys and girls, many models and colors with many kinds of extra parts (baskets, lights, horns). Your uncle wrote a letter promising that he will buy one for you as a birthday present. Choose a bicycle from the illustrations. Complete the following letter with a well-organized paragraph describing the bicycle you have chosen in such a way that he will be able to buy the model you really want.

Dear Uncle,

I think it's wonderful to get a bicycle for my birthday! I have thought about the one I would like and now I think I know. The bicycle I want ...



8. As they campaigned for the office of Senator from the State of Illinois in 1858, Abraham Lincoln and Stephen A. Douglas held seven joint debates throughout the state. Below is a speech taken from those debates. The speech was delivered on July 10, 1858. Although the senatorial race was a local one, the issues that were debated during the campaign were of national importance.

The "Nebraska bill" refers to the Kansas-Nebraska Act of 1854. This bill repealed the Missouri Compromise of 1820 and reopened the possibility of extending slavery into the newly organized Kansas and Nebraska territories.

Read the speech carefully. Lincoln states, "A house divided against itself cannot stand...It will become all one thing or the other." Using information from the speech and from what you know about events that led to the Civil War, write an essay in which you explain what Lincoln means by this statement. In your explanation, be sure to:

- identify the issue(s) or problem(s) which threaten to "divide the house";
- summarize how Lincoln believes that the issue(s) would best be resolved;
- explain what you consider to be the most convincing evidence he uses to support his proposed solution.

Education Staff Network
National Conference of State Legislatures

Annapolis, Maryland
November 12, 1988

Future of Testing

1. What will be tested?

   Higher order thinking

   Subject matter areas

2. How will testing occur?

   The role of technology

   Individual vs. group tests

   Portfolio development

3. For whom will tests be useful?

   Rapid development of display options will expand audience

   Policy relevance

CODE OF FAIR TESTING PRACTICES IN EDUCATION

Prepared by the Joint Committee on Testing Practices

The Code of Fair Testing Practices in Education states the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of testing. Although the Code has relevance to many types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs. The Code is not intended to cover tests made by individual teachers for use in their own classrooms.

The Code addresses the roles of test developers and test users separately. Test users are people who select tests, commission test development services, or make decisions on the basis of test scores. Test developers are people who actually construct tests as well as those who set policies for particular testing programs. The roles may, of course, overlap as when a state education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

The Code has been developed by the Joint Committee on Testing Practices, a cooperative effort of several professional organizations, that has as its aim the advancement, in the public interest, of the quality of testing practices. The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. In addition to these three groups, the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association are now also sponsors of the Joint Committee.

Code of Fair Testing Practices in Education. (1988) Washington, D.C.. Joint Committee on Testing Practices. (Mailing Address. Joint Committee on Testing Practices, American Psychological Association, 1200 17th Street, NW, Washington, D.C. 20036.)

The Code presents standards for educational test developers and users in four areas:

A. Developing/Selecting Tests
B. Interpreting Scores
C. Striving for Fairness
D. Informing Test Takers

Organizations, institutions, and individual professionals who endorse the Code commit themselves to safeguarding the rights of test takers by following the principles listed. The Code is intended to be consistent with the relevant parts of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985). However, the Code differs from the Standards in both audience and purpose. The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests. The Code is not meant to add new principles over and above those in the Standards or to change the meaning of the Standards. The goal is rather to represent the spirit of a selected portion of the Standards in a way that is meaningful to test takers and/or their parents or guardians. It is the hope of the Joint Committee that the Code will also be judged to be consistent with existing codes of conduct and standards of other professional groups who use educational tests.

A Developing/Selecting Appropriate Tests*

Test developers should provide the information that test users need to select appropriate tests.

Test Developers Should:

1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.
2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended purposes.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).
4. Describe the process of test development. Explain how the content and skills to be tested were selected.
5. Provide evidence that the test meets its intended purpose(s).
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.
8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations.

Test Users Should:

1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.
2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.
3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Become familiar with how and when the test was developed and tried out.
5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.
6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.
7. Ascertain whether the test content and norms group(s) or comparison group(s) are appropriate for the intended test takers.
8. Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

*Many of the statements in the Code refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the test development process should be designed to help ensure that the completed tests will be in compliance with the Code.

B Interpreting Scores

Test developers should help users interpret scores correctly.

Test users should interpret scores correctly.

Test Developers Should:

9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.
10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.
11. Warn users to avoid specific, reasonably anticipated misuses of test scores.
12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.
13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).

Test Users Should:

9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.
10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.
11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.
12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.
13. Obtain evidence to help show that the test is meeting its intended purpose(s).

C Striving for Fairness

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Developers Should:

14. Review and revise test questions and related materials to avoid potentially insensitive content or language.
15. Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test Users Should:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
15. Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test.
16. When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

D Informing Test Takers

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

Test Developers or Test Users Should:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.

18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers.

Under some circumstances, test developers have direct control of tests and test scores. Under other circumstances, test users have such control. Whichever group has direct control of tests and test scores should take the steps described below.

Test Developers or Test Users Should:

19. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.

20. Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.

21. Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.

Note: The membership of the Working Group that developed the Code of Fair Testing Practices in Education and of the Joint Committee on Testing Practices that guided the Working Group was as follows:

Theodore P. Bartell
John R. Bergan
Esther E. Diamond
Richard P. Duran
Lorraine D. Eyde
Raymond D. Fowler
John J. Fremer
(Co-chair, JCTP and Chair, Code Working Group)
Edmund W. Gordon
Jo-Ida C. Hansen
James B. Lingwall
George F. Madaus
(Co-chair, JCTP)
Kevin L. Moreland
Jo-Ellen V. Perez
Robert J. Solomon
John T. Stewart
Carol Kehr Tittle
(Co-chair, JCTP)
Nicholas A. Vacc
Michael J. Zieky
Debra Boltas and Wayne Camara of the American Psychological Association served as staff liaisons

Additional copies of the Code may be obtained from the National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, D.C. 20036. Single copies are free.