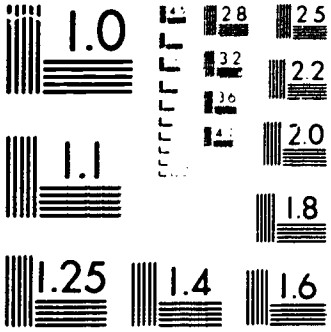
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6

DOCUMENT RESUME

ED 174 652 TM 009 501

AUTHOR Madaus, George; And Others
INSTITUTION Massachusetts State Dept. of Education, Boston. Bureau of Research and Assessment.
TITLE Development and Application of Criteria for Screening Commercial Standardized Tests for the Massachusetts Basic Skills Improvement Policy.
PUB DATE Apr 79
NOTE 86p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Achievement Tests; *Basic Skills; Check Lists; Educational Assessment; *Evaluation Criteria; Minimum Competency Testing; Rating Scales; Readability; Standardized Tests; *State Standards; Student Testing; Test Bias; *Test Reviews; *Test Selection; Test Validity
IDENTIFIERS Massachusetts

ABSTRACT

To permit local districts to exercise their option to use standardized tests for state-mandated evaluations, a test rating form was developed. A key concern was the percentage of state approved reading and mathematics skills measured by a test, rather than the test's value for other uses. Other content concerns were readability and cultural and sex bias. Most technical criteria were straightforward, but additional issues included extrapolation of grade equivalent scores and the consistency of mastery decisions resulting from the test. Mathematics, reading, and vocabulary tests only, from fourteen standardized achievement batteries were reviewed by three committees: technical, reading, and mathematics. Reading content ratings (representing content validity, readability, and bias) ranged from 0-8 of the maximum 10 points available; mathematics content ratings were either 2 or 5. Technical ratings for both subject areas were much higher--ranging from 29-50, or 52-89% of the possible score. It was recommended that bias and technical adequacy be considered independently of content validity. The first two issues should be handled by the state; content validity is the responsibility of the local district. (The rating form and basic skills objectives are appended.) (CP)

DEVELOPMENT AND APPLICATION OF CRITERIA FOR SCREENING COMMERCIAL STANDARDIZED TESTS

FOR THE

MASSACHUSETTS BASIC SKILLS IMPROVEMENT POLICY

A Report Prepared By

George Madaus
Peter Airasian
Ron Hambleton
Robert W. Consalvo
Lisanio R. Orlandi
Public Affairs Research Institute

for

Massachusetts Department of Education
Bureau of Research and Assessment
31 St. James Avenue
Boston, Massachusetts 02116

April, 1979

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The Department of Education and staff of the Public Affairs Research Institute wish to thank the members of the Review and Screening Committees for their cooperation and participation in the development and test review process.

Review Committee

Fred Andelman	Massachusetts Teachers Association	Boston
Pamela Almeida	Director of Reading	Reading
Tom Armstrong	Director of Mathematics	Foxboro
Robert Bray	Secondary School Principal	Attleboro
Jid Kamitian	Secondary School Principal	Holliston
Ann Keenan	Director of English	Braintree
Barbara Krysiak	Director of English	Waltham
Ben Levy	Massachusetts Teachers Association	Boston
Margarita Muniz	Bilingual Department	Boston
Ed Reidy	Project Competency	Fitchburg
Ginny Yardumian	School Committee Member	Malden

Screening Committee

Etta Anderson	Director, School Psychological Services Unit	Boston
Stephen Baker	Director of Measurement	Worcester
James Carabetta	Math Department Head (High School)	Palmer
June Bowman	School Psychometrist	Brockton
Karen Childs	School Psychologist	W. Springfield
Rose Dawkins	Reading Teacher (Junior High School)	Worcester
Anita Dodson	Reading Department Head (High School)	Acton-Boxborough
Rose Feinberg	Director of Language Arts	Lunenberg
Abilio Fernandes	Curriculum Resource Specialist (High School)	Fall River
Louise Forsyth	Coordinator of Testing	Quincy
Grace Kaczynski	Director of Pupil Personnel Services	Watertown
Mary McGauvran	Vice-President, University of Lowell	Lowell
Guy Parker	School Psychometrist	Gloucester
Peter Pullen	Math Teacher (High School)	Greenfield
Severina Rios	Bilingual Math Teacher (High School)	Worcester
Vincent Silluzio	Director of Research and Planning	Newton
Theodora Silvester	Director - North End Community Center	Springfield
Derrick Sudeall	Math Teacher (Middle School)	Boston
J. Bryan Sullivan	Math Department Head (High School)	Hudson
John Tsang	Bilingual Math Teacher (High School)	Boston
James Tynan	Acting Assistant Superintendent	Pittsfield
Julia Wan	Director of Science	Watertown
Carrie Weinick	Reading Teacher (Vocational High School)	Canton
Arnold Zins	Math Department Head (Regional High School)	W. Newbury

INTRODUCTION

On January 23, 1979, the State Board of Education approved regulations for implementation of the Basic Skills Improvement Policy. These regulations define the requirements for the basic skills improvement programs that public school districts will be establishing in the areas of reading, writing, mathematics, listening and speaking. The regulations were adopted after public hearings held in November, 1978. The Board of Education believes that these regulations will serve as a vehicle for public schools, with broad community participation, to establish sound minimum standards for basic skills and to examine their instructional programs in light of these standards.

The regulations require each public school district to evaluate each student's achievement of the minimum standards at least once in each of the early elementary, later elementary, and secondary school grade levels. By 1980-81 each district must evaluate student achievement of minimum standards in reading, writing and mathematics.

At the high school level each public school district has the option of using one or more of the following evaluation instruments to evaluate student achievement of minimum standards:

(a) Evaluation instruments available from State Department of Education;

(b) Commercially available evaluation instruments approved by the Department of Education; or

(c) Locally utilized or developed evaluation instruments approved by the Department of Education as being comparable to either (a) or (b) above.

This report deals with the Department's effort to permit local districts to exercise option (b), using commercially available evaluation instruments approved by the Department.

We were asked to identify and refine basic criteria which could be used by committees of public school personnel to screen commercial standardized tests in terms of their suitability for possible use by school districts as part of the Massachusetts Basic Skills Improvement Policy. Our initial task, therefore, was to develop a rating form and accompanying procedures that could be used with selected commercial standardized tests to arrive at a score or rating of the tests' adequacy for use in the Basic Skills Improvement Policy. In short, we were asked to build a procedurally simple, yet practical, means of assessing standardized commercial test adequacy within the context of the Massachusetts Basic Skills Improvement Policy. Further, after developing the rating form and directions for its use, we were asked to manage two screening sessions at which two committees composed of educators chosen by the Department of Education applied the rating form to commercial tests submitted by publishers. At the first screening session a committee of teachers and subject matter specialists considered each test in terms of content validity, readability and overall freedom from sexual, racial, ethnic or cultural bias and stereotyping. At the second screening session a committee of school district test directors and guidance counselors considered each test in terms of its technical adequacy.

This report details our work in developing the criteria on the rating form, describes the final criteria, and presents the results obtained when the criteria were applied by the committees appointed by the Department of Education.

It is important to note the limitations of our involvement in the process of arriving at a state approved list of commercial standardized achievement tests. We were not involved in any way with the actual approval decision. We presented the Department with the factual results of the screening process; it was then the responsibility of the Department to develop a process by which the factual results of the screening were used to arrive at a decision to approve or disapprove any particular test. In other words, the standards used in the approval decision were solely the responsibility of the Department.

One other important caveat should be noted at the outset. The commercially available standardized tests under review were not constructed nor intended for a use as specific as that inherent in the Basic Skills Improvement Policy. In short, none of the tests were specifically built to assess the approved list of 14 reading skills objectives or 38 mathematics skills objectives in Massachusetts. Instead, these tests were designed to measure objectives that are common to the most widely used curriculum or textbooks at a particular level in Mathematics and Reading. Our review, therefore, says nothing about the value or suitability of the tests for other uses. It is concerned only with the use of tests as part of the Basic Skills Improvement Policy. If after review a test has not met the Department's standards for inclusion on its approved list, this should in no way be construed to reflect on the test's suitability in terms of its originally intended use.

DEVELOPMENT OF THE SCREENING CRITERIA AND PROCEDURES

Background

Many test review forms have been developed over the years for use with norm-referenced tests (cf. CSE 1970*). These review forms have typically reflected issues and technical matters discussed in the AERA/APA/NCME** Test Standards. Unfortunately, the Test Standards do not apply too well to criterion-referenced tests. Therefore, there are relatively few review forms available that can be used with criterion-referenced tests. However, some progress toward establishing guidelines and review forms for criterion-referenced tests has been made (see, for example, Hambleton and Eignor, 1978).*** Our efforts have not been, however, to produce a comprehensive list of evaluative criteria, associated directions and review forms. Rather, we have attempted to provide a brief set of evaluative criteria suitable for use with both norm and criterion referenced tests being considered for possible use in the Commonwealth's new Basic Skills Improvement Policy.

To reiterate, the task we were confronted with was limited: develop a practical, easily used format for evaluating the adequacy of standardized tests for use in the Massachusetts Basic Skills Improvement Policy. In the face of this practical problem we have omitted many criteria common to other test assessment forms and added other criteria which are of particular import in the context of the Basic Skills Improvement Policy. It should be noted that many additional criteria could be appended to those contained in our final rating form. For example, the

* Center for the Study of Evaluation
** American Educational Research Association/American Psychological Association/ National Council on Measurement in Education.
***Journal of Educational Measurement, Winter, 1978, pp. 321-327.

matter of test cost is often addressed in review forms. This is not included in final rating form because, for this review, the State is concerned only with content and technical qualities of available tests. Individual districts may decide for themselves whether a test is too expensive. In the interests of administrative and interpretive simplicity we have focused on those test properties which we consider to be most important for judging a test's adequacy for use by Local Education Agencies (LEA's) in the Basic Skills Improvement Policy.

The rating form focuses upon two subject areas: reading and mathematics. For tests in each area, two general domains of adequacy are rated: (1) Content Adequacy and (2) Technical Adequacy. For purposes of rating any given test, we felt that these two domains should be rated by different panels, one familiar with subject matter content, the other with technical standards of test construction. The remainder of this section considers important aspects of the rating form in the content and technical areas respectively.

Content Issues

The key concern in judging the content adequacy of standardized tests for use in the Basic Skills Improvement Policy is the percent of the Massachusetts reading or mathematics skills measured by the test. That is what is the congruence between the items on a test and the behaviors implied by the Massachusetts basic skills objectives. Obviously, other things being equal, the greater the number of Massachusetts basic skills in reading or math judged by subject matter specialists to be reflected in a test, the more suitable that test is for use in the Basic Skills Improvement Policy.

It was, however, thought to be unlikely that every objective or every item on any standardized test would measure one of the Massachusetts basic skills in reading or mathematics. It was thought to be more likely, although clearly not a certainty, that the Massachusetts basic skill statements would comprise a subset of the total set of the objectives and items contained in any standardized test. The central concern, therefore, was whether a high percentage of a test's items overlapped with the State-approved list of objectives. The rating form awards a higher score to tests whose objectives and items are highly congruent with the Massachusetts basic skill statements than to tests with lower congruence.

A related issue which generated a great deal of discussion among the authors involved how many test items must be present for each basic skill before it could be said that that basic skill was measured by the test. Was one item per basic skill sufficient; three items; five items? This issue is of concern because of the differing ways districts might select to score pupils' performance on the test. If a skill by skill assessment were adopted, a single item per skill would be unlikely to be a representative sample of pupil behavior or reliable enough to make pass-fail decisions about pupil performance on that skill. In general, the greater the number of items tapping a particular skill, the greater the confidence in pass-fail decisions made about that skill. On the other hand, if a total score across all skills was used as the basis for pass-fail decisions about individual students, the number of items tapping any particular skill would be of less concern since the decisions would not be about performance on a single skill.

For reasons of convenience and practicality, the rating form was based on the least stringent situation likely to occur in district use of the tests, namely, aggregation across skill areas to arrive at a single score on which achievement/non-achievement decisions will be based. Consequently, the number of items tapping any particular skill in reading or mathematics was not considered to be a critical concern; so long as there was at least one item measuring a given basic skill, regardless of other areas measured on the test, and regardless of the fact that a larger number of items might measure one skill rather than another, the test met the criterion of assessing that skill. Districts which adopt a skill by skill scoring and pass-fail decision procedure should be advised of the difficulty of basing decisions about skill mastery on small samples of items. Initially, then, we felt that a test could be judged to possess content validity in terms of the Massachusetts basic skills objectives if the test contains at least one item measuring each skill.

We felt that if a test did not have one item for each basic skill, the test could be judged as inadequate in terms of assessing the Massachusetts basic skills objectives. However, after consultation with a Review Committee, made up of members from the Advisory Committee on Basic Skills and its subcommittees, it was decided instead that a given test would be rated in terms of the of the percentage of basic skills objectives measured by at least one test item. Subsequently, the five point rating scale based on the percentage of basic skills tapped by a test was devised. This change left the ultimate decision about the standard to be used to judge a test's content validity in terms of the Massachusetts Basic Skills Objectives to the Department, after consultation with the Review Committee.

The next issue that had to be dealt with concerned which items from which tests in a battery should be considered by raters in seeking to identify matches between tests and the approved list of basic skills objectives. In fact, this problem contained two issues--not only what items from a test should be considered but which tests from a test battery should be reviewed. On the latter point, it was decided to include only the reading, vocabulary and mathematics tests from a given battery. On the former point two techniques were possible. First, have the reviewer consider each item in terms of the list of skill objectives or, second, ask the test publishers to nominate those items which the publisher felt measured each skill and have the reviewers certify the publisher's nomination. The Review Committee decided on January 31 to adopt the latter approach: Content review involved having reviewers certify or check upon test publishers' nominated items as being measures of the skill objectives in question.

When publishers were asked to nominate items, some elected items on tests in the battery other than reading or math. For example, the Study Skills Test in some batteries had, in thepublisher's view, items which measured some state objectives on the list. Items on tests other than math and reading were not considered and this is a limitation in the content review process. If these items had been considered, the content match percentage for many batteries may have been higher.

A second issue in the content review involved a judgment about the readability level of the test. Time and resources to conduct a separate readability study were not available. The Review Committee, after considerable discussion, decided that in

the judgment of the subject matter specialists--all school personnel--the readability level should be appropriate for the lowest grade level for which the test was designed.

A third issue in rating content involved test bias. In actuality, the general rubric "bias" contains two distinct concerns. The first concern, which is properly called "bias", involves the inclusion of items on the test which, because of their characteristics, affect the score attained by different identifiable groups on the test. Items which are culturally, racially or sexually loaded in such a way that one group of test takers has an unfair advantage in answering the items correctly are biased in this meaning of the word. A second concern involves the inclusion of items which may be offensive to members of certain racial, ethnic or sex groups - in the sense that they stereotype characteristics of these groups - but which do not affect test performance per se. Items which continually show women in homemaking situations and men in occupational situations generally involve sex stereotyping which may be offensive, but which does not necessarily affect the performance of test takers on the items.

To confront the issue of bias, therefore, involves two types of judgments. The first concerns a judgment about the inclusion of items which disproportionately affect the performance of different groups. The second judgment concerns the inclusion of items which may be offensive to, but not necessarily related to, the performance of these groups. The first type of judgment should be made on the basis of empirical evidence regarding the performance of different groups on items initially judged to be biased. It is important to note that empirical evidence is crucial in this judgment, because studies have shown that items perceived by panels as being unfair to particular groups are not always so in the light of the actual test performance of the groups. Of

course, we could not expect the review committee to carry out an empirical investigation necessary to document perceived bias. (The rating form does not ignore this issue entirely. Questions are asked and points awarded in the technical section of the review on the basis of whether the test publisher, in the test manual, addresses this issue and describes how it was dealt with.)

Stereotyping in the items is more easily detected and is part of the Content review. Reviewers examined the test items to determine whether there was a consistent or overriding pattern of racial, ethnic or sexual stereotyping. A caveat is in order as regards examination of the items for offensive stereotyping: such stereotyping should be considered within the context of the full item set, not on an item by item basis. For example, the fact that one item portrays a woman in the kitchen or a minority group member in an unskilled occupation does not necessarily imply stereotyping. Some women do spend large amounts of time in the kitchen and some minority group members do hold unskilled jobs. At issue is whether members of such groups are consistently or predominantly portrayed in such circumstances relative to the way in which other groups are portrayed. If women are portrayed only in the home and men only on the job, then the test does involve stereotyping. The rating form made provisions for awarding points to tests free of stereotypical bias.

The Bureau of Equal Educational Opportunity in the Department of Education conducted a separate study on possible bias inherent in the tests nominated for review for inclusion in the Basic Skills Improvement Policy. In order to be approved by the department, tests had to pass BEEO Bias Review.

Technical issues

Most of the technical concerns covered by the test rating forms are fairly straightforward and self-explanatory. There are two issues, however, which warrant additional consideration: extrapolated grade equivalent scores and the consistency of mastery decisions resulting from the test. This section deals with these two issues in turn.

In setting standards of satisfactory performance on the tests used to assess basic skills in reading and mathematics, it seems likely that some districts would wish to base their cut-off score on the grade equivalent norms provided with their test. That is, at the high school level, some districts will set a score equal to a grade equivalent of, say 10.0, as defined in the test norms as the cut point which will differentiate pupils who pass the test from pupils who do not. It is commonly accepted that there are problems in using grade equivalent scores for test interpretation purposes. However, there is an additional problem in using grade equivalent scores as performance standards at the high school level. Simply put, many standardized tests base their high school grade equivalent norms on extrapolated data rather than on actual data gathered from a high school norming sample. The meaning of such extrapolated data in relation to the actual performance of high school pupils will not be clear and may seriously over or under estimate actual pupil performance. Since extrapolated grade equivalent norms are less meaningul than grade equivalent norms based on a norming sample of high school pupils, the initial drafts of the rating form penalized tests whose norms were based upon extrapolated grade equivalent norms. However, the Department may make this point moot by a decision not to allow districts to report results in terms of grade

equivalents. The department is currently finalizing its regulations for reporting and may require that districts report their results in terms of a percentage of items answered correctly. That is, the Department may decide to require reporting in terms of raw scores rather than any derived score. This explains the absence from the criteria of any technical items dealing with derived scores.

A second technical concern which arises when tests are used to make pass-fail decisions about individual pupils involves the consistency of decisions resulting from the test. Most standardized tests provide reliability information in the form of internal consistency, test-retest, or split half correlation coefficients. Indices such as these are related to the accuracy of the scores resulting from the test. In the context of the Massachusetts Basic Skills Improvement Program, where testing is used to classify students into two categories, pass or fail, it is not the accuracy or consistency of the pupil's score per se which is of primary concern, but rather the accuracy of the ultimate classification made on the basis of that score. It is true that the accuracy of classification will be related to the reliability of the test scores; in general, the more reliable the test scores, the fewer the errors of classification. However, there is no simple or direct procedure which enables one to derive the number of misclassifications likely to occur given a particular test score reliability value. Moreover, regardless of the reliability of a test, the absolute number of misclassifications will vary with the cut off score used to differentiate passing from failing students. The closer the cut off score is to the 50th percentile (or some equivalent derived score) in the test score distribution, the greater will be the number of students misclassified when that cut off score is applied.

Consequently, it is most desirable that the manual accompanying a standardized test address the issue of consistency of decisions resulting from the test in light of different performance standards or cut scores. However, it is unlikely that such information is provided in most test manuals, although points are awarded on the rating form for manuals that do. Since standard test score reliability indices afford an approximation of the consistency of the decisions which will result from the test, the rating form provides a graduated scale which awards a test a higher number of points for high reliability indices. We would recommend that publishers of approved tests be required to provide data on the reliability of classifications resulting from their tests within two years or suffer the loss of that approval.

The Rating Form and Procedures

The rating form and procedures for its use went through four drafts before the final version was finalized. The first draft constituted the initial ideas developed by Madaus, Hambleton and Airasian at two day-long meetings. As a result of this draft Hambleton produced a second draft which was in turn revised by the three authors. On January 31, the third draft was submitted for a review by the Review Committee selected from the larger State Advisory Committee on Basic Skills Improvement and its subcommittees. That meeting resulted in several policy decisions mentioned above. Further, several items were discarded (for example, one asking the rater for an overall rating of test adequacy) and several items added (for example, criteria dealing with readability and test bias).

The suggestions from this meeting were incorporated into a fourth draft. The

draft was mailed to each member of the Review Committee and each member of the Advisory Committee on Basic Skills for their reactions. Comments from this review were incorporated into the final version of the review instruments and procedures.

The final set of review instruments, contained in Appendices A-F include:

(1) The Review Form itself. The first 8 questions ask for information about the reviewer and the test. Questions 9-11 dealt with three Content considerations: content coverage, readability level, and bias. Questions 13-31 were technical questions. (p. 40)

(2) Directions for conducting a content review. These directions tell the reviewer how to deal with each content question on the review form. (p. 45)

(3) Directions for conducting a technical review. These directions tell the reviewer how to deal with each technical question on the review form. (p. 48)

(4) A mathematics skills check list. This form was used for determining the match between the nominated items and the Massachusetts math skill objectives. (p. 51)

(5) A reading skills check list. This form was used for determining the match between the nominated items and the Massachusetts reading skill objectives. (p. 58)

(6) An evaluation summary sheet. Here the rating for each item was transferred from the rating form and a numerical value attributed to each rating. The total number of points for the content review and technical review were also calculated on this form. (p. 62)

TEST SCREENING

Process

The screening process was designed to be an objective review conducted by a neutral Screening Committee appointed by the Department of Education. Three sub-committees of the Screening Committee were established: Reading, Math and Technical. Each sub-committee consisted of nine members grouped into three teams of three persons. Committee members were selected from school systems and other education related agencies from across the state. Each Regional Council was invited to send one participant to serve on the Screening Committee. Minority and bilingual representatives were included on the Screening Committee. Members were assigned to each sub-committee by the Department of Education according to each person's area of expertise.

The makeup of each committee was:

Reading

Name	Position	Town
Rose Dawkins	Reading Teacher (Junior High School)	Worcester
Anita Dodson	Reading Department Head (High School)	Acton-Boxborough
Rose Feinberg	Director of Language Arts	Lunenberg

Reading (cont.)

Name	Position	Town
Mary McGauvran	Vice-President, University of Lowell	Lowell
Theodora Silvesta	Director - North End Community Center	Springfield
James Tynan	Acting Assistant Superintendent	Pittsfield
Carrie Weinick	Reading Teacher (Blue Hills Vocational High School)	Canton

Math

James Carabetta	Mathematics Department Head (High School)	Palmer
Abilio J. Fernandes	Curriculum Resource Specialist (High School)	Fall River
Peter Pullen	Mathematics Teacher (High School)	Greenfield
Severina Rios	Bilingual Mathematics Teacher (High School)	Worcester
Derrick Sudeall	Mathematics Teacher (Middle School)	Boston
J. Bryan Sullivan	Mathematics Department Head (High School)	Hudson
John Tsang	Bilingual Mathematics Teacher (High School)	Boston
Julia Wan	Director of Science	Watertown
Arnold Zins	Mathematics Department Head (Pentucket Regional High School)	Canton

Technical

Etta Anderson	Director, School Psychological Services Unit	Boston
Stephen Baker	Director of Measurement	Worcester
June Bowman	School Psychometrist	Brockton
Karen Childs	School Psychologist	W. Springfield
Louise Forsyth	Coordinator of Testing	Quincy

Technical (cont.)

Name	Position	Town
Grace Kazcynski	Director of Pupil Personnel	Watertown
Guy Parker	School Psychometrist	Gloucester
Vincent Silluzio	Director of Research and Planning	Newton

The Department of Education gathered data from school districts (by means of the October School System Summary Report) on the extent of standardized testing at the district level. Based on analysis of this information, fourteen standardized tests were identified as the most commonly used standardized group achievement tests in Massachusetts secondary schools (grades 7-12). Because these tests were used by an overwhelming majority (over 90%) of the school districts in Massachusetts, they were identified as those to be examined in an initial screening of commercial tests.

Test	Level	Publisher	Copyright Date
Basic Skills Assessment Program	17, 18, 19	Addison-Wesley/ETS	1977
California Achievement Tests	17, 18, 19	CTB/McGraw-Hill	1977
Cooperative English-Reading Comprehension		Addison-Wesley/ETS	1960
Cooperative Mathematics - Arithmetic		Addison-Wesley/ETS	1962
Comprehensive Tests of Basic Skills	3, 4	CTB/McGraw-Hill	1973
Gates-MacGinitie Reading Tests	E, F	Houghton-Mifflin	1978
Iowa Tests of Basic Skills	13, 14	Houghton-Mifflin	1978
Iowa Tests of Educational Development		Science Research Assoc.	1970
Metropolitan Achievement Test	Adv. 1 & 2	Psychological Corp.	1978
SRA Achievement Test	F, G, H	Science Research Assoc.	1978
Stanford Achievement Test	Advanced	Psychological Corp.	1972
Stanford Diagnostic Reading Test	Blue/Brown	Psychological Corp.	1974/1976
Stanford Test of Academic Skills (TASK)	IA, IIA	Psychological Corp.	1972
STEP Sequential Test of Educational Progress	Reading, I Math, I, IJ	Addison-Wesley/ETS	1979

Between December and February, the Department communicated with publishers of these tests requesting that they provide the following information:

(a) The nomination of specific test items which the publisher determined were measures of each of the state basic skills objectives.

(b) Copies of Technical Manuals and directions for administering and interpreting the tests.

(c) Copies of the tests and answer keys.

(d) Any other materials which might be useful for the review process.

The responses of test publishers varied. Some included all information as requested, nominating test items for specific objectives and providing complete sets of manuals. Others were not as specific, choosing instead to nominate the entire test, or parts of a test as a measure of the objectives. A few failed to send technical manuals.

By February 23, Department of Education staff packaged and mailed the tests and accompanying materials, along with directions and instruments for the test review procedure, to members of the screening committees. Each member was asked to review the test (either for content or technical standards depending on the sub-committee) and complete the evaluation instruments independently of other sub-committee members.

On March 7th, the Reading and Math Committees met at the Central Massachusetts Regional Office in West Boylston to arrive at a consensus report for each test.

Members of PARI were available for consultation and to assist in the resolution of differences. Time and resources did not permit a study of inter-rater or intra-rater reliability.

All members of the Math sub-committee were present. Two members of the Reading sub-committee did not come to the meeting, leaving two teams with two members.

The review process proceeded smoothly toward consensus. Team members reported no difficulties in using the evaluation instruments independently, and were able to reach a single summary rating with little difficulty. In fact, consensus was arrived at so easily that members chose to return to the tests for a second review (Round II), principally to accommodate the tests for which specific items were not nominated by publishers as measures of the Massachusetts reading and math basic skills. Moreover, some reviewers also opted to review other items not nominated by publishers which, in their opinion, measured the State's objectives, making it possible for a test to achieve a higher rating than in the first screening.

The technical sub-committee met on March 9th. Their task was similar to the content teams'--to review independent ratings and to arrive at a consensus on the technical aspects of the tests. Only one member of the Technical sub-committee failed to attend the meeting, so that two teams were complete with three members, and one had two members.

As with the Content teams, the screening proceeded satisfactorily. Again, team members reported no difficulties in utilizing the evaluation instruments.

Issues Arising from the Screening Process

While the screening process functioned well, as noted above, several issues did arise. First, not all test publishers nominated items to measure the State's specific objectives. As cited earlier, some merely identified the entire test as a measure of basic skills, or identified items measuring a broad range of skills (e.g., literal comprehension). This, of course, made it impossible for the reviewers to make judgments concerning the validity of the item(s) as a measurement of a specific objective. Fortunately, some teams completed the first screening soon enough to allow time for a second review, in which test items were matched to State objectives and rated accordingly. This modification in the screening process was adopted to advance the opportunity for selecting the best instruments.

A second issue involved the fact that only reading and math tests were reviewed to correspond to the reading and math competencies. It was obvious to all involved--the reviewers, the developers of the criteria, and Department of Education staff--that other sub-tests of a test battery (e.g., Study Skills, Science) might include items addressing the State's objectives. However, because of the logistical and financial realities faced by local districts in obtaining a total score for items which are selected from several sub-tests, and because of the impracticality of reviewing all sub-tests of a battery, the Department of Education limited the review to reading and math sub-tests.

A third issue related to content dealt with the acceptability of items which measured higher-order skills but assumed mastery of lower-order skills. The reviewers were instructed that the State's objectives were to be interpreted literally, and that items were to correspond directly to the basic skill statements, and not through a

higher-order skill. This decision was made jointly by PARI and Department of Education staff at the Content meeting. (Further, many raters found it difficult to separate their judgment of an item's difficulty from whether or not it was an appropriate measure of a minimal basic skill objective.)

The fourth content issue was raised by the content sub-committee members during their work. They noted that many of the State's competency objectives were vague, included more than one skill, and were subject to several interpretations. This, of course, affected their ability to interpret the objective literally and to rate corresponding items. This problem was resolved, in part, by team consensus. This difficulty involves an important issue, however, and one likely to be faced by local school districts whether in selecting or constructing a test.

The final concern was the only one involving the Technical sub-committee. As noted previously, technical data were not available because publishers frequently failed to send technical manuals or complete information. Because of this, technical information was missing, and reviewers were forced to rate some tests lower. Of particular note was the absence of data pertaining to cut points or mastery decisions. Of course, since the tests were, for the most part, standardized norm-referenced achievement tests, these kinds of data were not likely to be available. The absence of this particular information bears on the concern discussed earlier: many excellent tests were screened for uses other than for which they were intended.

Results

Ratings were obtained on both Content and Technical considerations. Points were assigned to each factor in accordance with the evaluation criteria. In addition, a Total Content Score and a Total Technical Score were calculated. It should be noted that the accuracy of the reviewers' data was checked, e.g., errors of addition, obvious incongruencies of reported findings with known test characteristics. Where problems of this kind emerged, the original screening instruments were re-examined to enhance the overall effort toward accuracy.

Reading Tests - Content Ratings

A summary of the results of the Content ratings for the reading tests is presented in Table 1.

Percent of Basic Skills Measured - In Round I of the review, 13 of the 25 reading tests were able to be screened, because only the publishers of these 13 tests nominated items to measure specific skills. The highest percentage of the 14 Massachusetts' reading skills measured by a test was 71%; eight other tests measured 50% or more of the State's objectives.

Round II of the review included all but four of the tests. Two of these reading tests were reviewed in Round I and not reviewed again in Round II, because of lack of time. Two tests were not reviewed because publishers did not nominate items and because of lack of time. Changes in the percent of basic reading skills measured on tests reviewed in Round I and Round II were slight, ranging from 0-8%. The range of percentages in Round II was 36-79%. Four of the tests reviewed in Round II measured 71% or more of the basic skills reading objectives. The highest percentage was 79%. The majority of tests (12 including the two tests reviewed only in

Table 1

RESULTS OF CONTENT SCREENING OF READING TESTS

Test	Percent of Basic Skills		Content Rating (5)[a]		Readability (2)[a]	Bias (3)[a]	Total Content (10)[a]	
	Round I	Round II	Round I	Round II			Round I	Round II
Basic Skills Assessment	71(10)[b]	79(11)	3	3	2	3	8	8
California Achievement Tests, 17	N[c]	57(8)	N	0	2	3	N	5
California Achievement Tests, 18	N	71(10)	N	3	2	3	N	8
California Achievement Tests, 19	N	57(8)	N	0	2	3	N	5
Cooperative English Test, Rdg. Comp.	57(8)	N	0	N	0	0	0	N
Comprehensive Tests of Basic Skills, Level 3	N	71(10)	N	3	2	3	N	8
Comprehensive Tests of Basic Skills, Level 4	N	50(7)	N	0	2	3	N	5
Gates-MacGinitie, Level E, Reading Comprehension	36(5)	43(6)	0	0	2	3	5	5
Gates-MacGinitie, Level E, Vocabulary	N	N	N	N	N	N	N	N
Gates-MacGinitie, Level F, Reading Comprehension	36(5)	43(6)	0	0	2	3	5	5
Gates-MacGinitie, Level F, Vocabulary	N	N	N	N	N	N	N	N
Iowa Tests of Basic Skills, Level 13	50(7)	57(8)	0	0	2	3	5	5
Iowa Tests of Basic Skills, Level 14	50(7)	57(8)	0	0	2	3	5	5
Iowa Tests of Educational Development	50(7)	N	0	N	0	0	0	N
Metropolitan Achievement Tests, Advanced 1	36(5)	50(7)	0	0	2	3	5	5

[a] Maximum possible score

[b] In parentheses is the number of Massachusetts basic skills reading objectives actually measured by the test

[c] Not nominated correctly by publisher, not reviewed by team

Table 1 (Cont.)

RESULTS OF CONTENT SCREENING OF READING TESTS

Test	Percent of Basic Skills		Content Rating (5)[a]		Readability (2)[a]	Bias (3)[a]	Total Content (10)[a]	
	Round I	Round II	Round I	Round II			Round I	Round II
Metropolitan Achievement Tests, Advanced 2	57 (8)[b]	57 (8)	0	0	0	3	3	3
SRA Achievement Series, Level F	57 (8)	57 (8)	0	0	2	3	5	5
SRA Achievement Series, Level G	57 (8)	57 (8)	0	0	2	3	5	5
SRA Achievement Series, Level H	57 (8)	57 (8)	0	0	2	3	5	5
Stanford Achievement Test, 1 & 2	N[c]	71 (10)	N	3	2	3	N	8
Stanford Diagnostic Reading Test, Brown Level	N	36 (5)	N	0	2	3	N	5
Stanford Diagnostic Reading Test, Blue Level	N	43 (6)	N	0	2	3	N	5
Stanford TASK, Level IA	N	43 (6)	N	0	2	0	N	2
Stanford TASK, Level IIA	N	43 (6)	N	0	0	0	N	0
STEP, Level I	43 (6)	43 (6)	0	0	2	3	5	5

[a] Maximum possible score

[b] In parentheses is the number of Massachusetts basic skills reading objectives actually measured by the test

[c] N= Not nominated correctly by publisher, not reviewed by team

Round I) measured 50% - 57% of the skills. Seven tests measured less than 50% of the basic reading skill objectives.

Content Rating - The Content rating reflects the percentage of basic reading skills measured by a test. No reading test achieved the two highest possible ratings (5 and 4); four tests received ratings of 3. Nineteen tests received a zero rating.

Readability - The reviewers judged that all but four of the reading tests had suitable reading levels for most students in the lowest grade covered by the tests.

Bias - All but four of the reading tests reviewed received a rating reflecting no overall sexual, racial, and/or ethnic content or stereotyping.

Total Content Rating - Content ratings ranged from 0-8 of the maximum 10 points available. (It should be noted that the Bias rating represents 30% of the Total Content score and Readability 20%, leaving 50% for item validity considerations.) No reading test achieved a maximum score. Three reading tests scored at zero. The highest score attained was 8. One test achieved an 8 on Round I, with four tests reaching 8 on Round II. Most tests (36% of Round I, and 56% of Round II) attained a rating of 5.

Reading Tests - Technical Ratings

Table 2 presents the Total Technical Rating for each reading test. Each Total rating was transformed to a percentage score based on the maximum possible score of 56. Additionally, because most reading tests provided no data concerning criterion-referenced reliability, a second

Table 2

SUMMARY OF RESULTS OF TECHNICAL SCREENING OF READING TESTS

Test	Total Technical Ratings[a]	Rating Expressed as Percent of Maximum	Rating as Percent of Maximum Not Including Criterion Referenced Considerations[b]
Basic Skills Assessment	43	77	84
California Achievement Tests, 17	46	82	90
California Achievement Tests, 18	46	82	90
California Achievement Tests, 19	46	82	90
Cooperative English Test, Reading Comprehension	38	68	75
Comprehensive Tests of Basic Skills, Level 3	44	79	86
Comprehensive Tests of Basic Skills, Level 4	43	77	84
Gates-MacGinitie, Level E, Reading Comprehension	38	68	75
Gates-MacGinitie, Level E, Vocabulary	38	68	75
Gates-MacGinitie, Level F, Reading Comprehension	38	68	75
Gates-MacGinitie, Level F, Vocabulary	38	68	75
Iowa Tests of Basic Skills, Level 13	49	88	96
Iowa Tests of Basic Skills, Level 14	49	88	96
Iowa Tests of Educational Development	29	52	57
Metropolitan Achievement Tests, Advanced 1	40	71	78
Metropolitan Achievement Tests, Advanced 2	41	73	80

[a] Maximum possible score= 56

[b] Maximum possible score= 51, excluding item 18 on Review Form (See Appendix, p. 42)

Table 2 (Cont.)

SUMMARY OF RESULTS OF TECHNICAL SCREENING OF READING TESTS

Test	Total Technical Ratings[a]	Rating Expressed as Percent of Maximum	Rating as Percent of Maximum Not Including Criterion Referenced Considerations[b]
SRA Achievement Series, Level F	37	66	73
SRA Achievement Series, Level G	37	66	73
SRA Achievement Series, Level H	39	70	73
Stanford Achievement Test, 1 & 2	44	79	86
Stanford Diagnostic Reading Test, Brown Level	46	82	90
Stanford Diagnostic Reading Test, Blue Level	38	68	75
Stanford TASK - Level IA	29	52	57
Stanford TASK, Level IA	31	55	61
STEP, Level I	50	89	98

[a] Maximum possible score= 56

[b] Maximum possible score= 51, excluding item 18 on Review Form (See Appendix, p. 42)

percentage score was derived based on a possible score of 51, which excluded points awarded for this factor.

In general, the Technical ratings for reading tests tended to be much higher than the Content ratings. Technical scores ranged from 29-50 or 52%-89% of the possible score. Seven tests scored above 80%; seven tests scored between 70% - 79%; eight were in the 60% - 69% range; and three tests were near 50%. When the ratings were calculated on a maximum score of 51 which eliminated ratings for criterion-referenced considerations, percentage scores rose 6-9%. This change raised seven tests to over 90%, with all save three of the tests scoring 70% or better.

A complete listing of the ratings for each technical consideration is presented in Appendix G. It is important to note that because many of the test publishers did not supply all of the materials required (e.g., technical manuals) for the review, often data were not available to permit a rating. In these cases, the criteria called for a zero rating. If the data were available, the technical scores of most of those tests would rise. The technical considerations which received a zero rating because information was not available are identified by an asterisk (*) in the Table in Appendix G. (p. 66)

Math Tests - Content Rating

A summary of the results of the Content ratings for the Math tests is presented in Table 3.

Percent of Basic Skills Measured - In Round I of the review, 16 of the tests for which publishers had properly nominated items were examined. Ratings indicated

Table 3

RESULTS OF CONTENT SCREENING OF MATH TESTS

Test	Percent of Basic Skills		Content Rating (5)[a]		Readability (2)[a]	Bias (3)[a]	Total Content (10)[a]	
	Round I	Round II	Round I	Round II			Round I	Round II
Basic Skills Assessment	55 (21)[b]	N	0	N	2	3	5	N
California Achievement Tests, 17	N[c]	55 (21)	N	0	2	3	N	5
California Achievement Tests, 18	N	50 (19)	N	0	2	3	N	5
California Achievement Tests, 19	N	45 (17)	N	0	2	3	N	5
Cooperative Mathematics, Arithmetic	39 (15)	45 (17)	0	0	2	3	5	5
Comprehensive Tests of Basic Skills, Level 3	45 (17)	45 (17)	0	0	2	3	5	5
Comprehensive Tests of Basic Skills, Level 4	32 (12)	32 (12)	0	0	2	3	5	5
Iowa Tests of Basic Skills, Level 13	39 (15)	N	0	N	2	3	5	N
Iowa Tests of Basic Skills, Level 14	50 (19)	N	0	N	2	3	5	N
Iowa Tests of Educational Development	21 (8)	21 (8)	0	0	2	0	2	2
Metropolitan Achievement Tests, Advanced 1	45 (17)	N	0	N	2	3	5	N
Metropolitan Achievement Tests, Advanced 2	21 (8)	N	0	N	2	3	5	N
SRA Achievement Series, Level F	45 (17)	45 (17)	0	0	2	3	5	5
SRA Achievement Series, Level G	39 (15)	39 (15)	0	0	2	3	5	5

[a] Maximum possible score

[b] In parentheses is the number of Massachusetts basic skills mathematics objectives actually measured by the test

[c] N= Not nominated correctly by publisher, not reviewed by team

Table 3 (Cont.)

RESULTS OF CONTENT SCREENING OF MATH TESTS

Test	Percent of Basic Skills		Content Rating (5)[a]		Readability (2)[a]	Bias (3)[a]	Total Content (10)[a]	
	Round I	Round II	Round I	Round II			Round I	Round II
SRA Achievement Series, Level H	53(20)[b]	53 (20)	0	0	2	3	5	5
Stanford Achievement Test, (Tests 3, 4, 5)	N[c]	N	N	N	N	N	N	N
Stanford TASK- Level IA	34 (13)	37 (14)	0	0	2	0	2	2
Stanford TASK- Level IIA	26 (10)	26 (10)	0	0	2	0	2	2
STEP, Level I, Basic Concepts	50 (19)	53 (20)	0	0	2	3	5	5
STEP, Level IJ, Mathematics Computation	24 (9)	26 (10)	0	0	2	3	5	5

[a] Maximum possible score

[b] In parentheses is the number of Massachusetts basic skills mathematics objectives actually measured by the test

[c] N= Not nominated correctly by publisher, not reviewed by team

that the tests measured from 21%-55% of the State's 38 basic mathematical skills. Most tests were below 50%; only four were at or above 50%.

In Round II of the review, three of the four tests not examined in Round I were screened. Eleven of the tests reviewed in Round I were screened a second time in Round II, with only very slight changes, if any, in the rating of the percent of basic math skills measured. Five tests were screened in Round I, but not again in Round II, because of lack of time. One test was not screened in either round, because the publisher did not nominate items and because of lack of time.

If one examines ratings for Round II combined with ratings for tests screened only in Round I, the range of percent of basic skills measured by the tests was again 21%-55%. Most tests were still below 50%; six tests were at or above 50%. Five tests fell in the 40%-49% range, and eight tests were below 40%.

Content Rating - The content rating reflects the percentage of basic skills measured by the test. Every math test screened failed to achieve the minimum percentage of basic skills measured (60%) required to receive a rating above zero.

Readability - Every math test was judged to have a reading level suitable for most students in the lowest grade covered by the test.

Bias - All but three of the math tests received a rating reflecting no overall sexual, cultural, racial, and/or ethnic content or stereotyping.

Total Content Rating - Total Content ratings achieved by the tests were either 2 or 5 points of the maximum 10 points available. Sixteen tests received a score of 5, and three a score of 2. (Again, the reader is reminded that the Bias rating represents three of the possible 10 points, Readability, 2 points, and Content rating, 5 points.)

Math Tests - Technical Ratings

Table 4 presents the Total Technical Rating for each math test. Each total rating was transformed to a percentage score based on the maximum possible score of 56. Additionally, because most math tests provided no data concerning criterion-referenced reliability, a percentage score was derived based on a possible score of 51, which excluded points awarded for this factor from the total. As in the case of the Reading tests, the Technical ratings for Math tests tended to be much higher than the Content ratings. Technical scores ranged from 29-50 or 52%-89% of the possible score. Eight math tests were above 80%; six were between 70%-80%; three were in the 60%-69% range; and three math tests were near 50%. When the ratings were calculated on a maximum score of 51 which eliminated ratings for criterion-referenced considerations, percentage scores rose 5-9%. This change raised eight tests over the 90% level, with 17 of the 20 tests being above the 70% level.

A complete listing of the math ratings for each technical consideration is presented in Appendix G. It is important to note that because many of the test publishers did not supply all of the materials required (e.g., technical manuals) for the review, often data were not available to permit a rating. In these cases, the criteria called for a zero rating. If the data were available, the technical scores of most of those tests would rise. The technical considerations which received a zero rating because information was not available are identified by an asterisk (*) in the Table in Appendix G. (p. 68)

Table 4

SUMMARY OF RESULTS OF TECHNICAL SCREENING OF MATH TESTS

Test	Total Technical Ratings[a]	Rating Expressed as Percent of Maximum	Rating as Percent of Maximum Not Including Criterion Referenced Considerations[b]
Basic Skills Assessment	43	77	84
California Achievement Tests, 17	46	82	90
California Achievement Tests, 18	46	82	90
California Achievement Tests, 19	48	86	94
Cooperative Mathematics, Arithmetic	29	52	57
Comprehensive Tests of Basic Skills, Level 3	43	77	84
Comprehensive Tests of Basic Skills, Level 4	46	82	90
Iowa Tests of Basic Skills, Level 13	49	88	96
Iowa Tests of Basic Skills, Level 14	49	88	96
Iowa Tests of Educational Development	36	64	71
Metropolitan Achievement Tests, Advanced 1	41	73	80
Metropolitan Achievement Tests, Advanced 2	41	73	80
SRA Achievement Series, Level F	37	66	73
SRA Achievement Series, Level G	37	66	73
SRA Achievement Series, Level H	39	70	76

[a] Maximum possible score= 56

[b] Maximum possible score= 51, excluding item 18 on Review Form (See Appendix, p. 42)

Table 4 (Cont.)

SUMMARY OF RESULTS OF TECHNICAL SCREENING OF MATH TESTS

Test	Total Technical Ratings[a]	Rating Expressed as Percent of Maximum	Rating as Percent of Maximum Not Including Criterion Referenced Considerations[b]
Stanford Achievement Test, (Tests 3, 4, 5)	44	79	86
Stanford TASK, Level IA	31	55	61
Stanford TASK, Level IIA	31	55	61
STEP, Level I, Basic Concepts	50	89	98
STEP, Level II, Mathematics Computation	50	89	98

[a] Maximum possible score= 56

[b] Maximum possible score= 51, excluding item 18 on Review Form (See Appendix, p. 42)

CONCLUDING RECOMMENDATIONS

The process of reviewing commercially available standardized tests can conceptually be separated into three components. The first component of the review considered the content validity of the tests. That is, it asked how well the items in the test match the State's defined reading and math basic skills. The rating form called for a minimum of one item per basic skill in order to attain a match between the test content and any one of the 14 reading and 38 math basic skills. From the local districts' point of view where the districts' basic skills differ from the State's, there is the supplementary question of whether a test item reflects skills that are part of the districts' own Basic Skills Improvement Program. The State's content review did not consider this validity issue. Further, in undertaking the content review, because of time and resource constraints, the State Department limited review to the mathematics, reading and vocabulary tests from each test battery. Therefore, the judged content match between the basic skills and test items might have been higher if the entire test battery had been considered. The second component of the review considered the issue of whether the items in the test overall were free of offensive sexual, cultural, racial, and/or ethnic content and/or stereotyping.

The final component of the review was technical, concentrating on aspects of test development such as: item selection, item characteristics, item writing, reliability, norms, directions, test format, etc.

In determining whether or not a test should be approved, each component of the review process needs to be considered. Any test which lacks content validity should not be approved. Any test which is offensive in terms of racial, ethnic, cultural or sexual stereotyping should not be approved. Finally, any test which is technically deficient should not be approved.

The issues of bias and technical adequacy can and should be considered independently of whether the test is content valid for use in a particular local district. That is if a test does not meet minimal technical standards or is offensive in terms of sexual, racial, or ethnic bias, it should not be used by any district. Thus, these two issues clearly fall within the Department purview in terms of test approval.

The issue of content validity, however, is idiosyncratic to the district that wishes to use a test. If a district can show, after considering the entire battery rather than simply the reading and math tests, that there is a match between the test items and the locally endorsed basic skills of which the State basic skills are a subset, then it may be said that the test has content validity for that system. (Implied in this local review is a local decision on item difficulty relative to districts' definitions of minimum standards.)

The implication of the preceding discussion is that a two stage approval process be employed. In the first stage the State arrives at a list of tests that meet minimal technical standards and are free of bias. In the second stage the State approves of the procedure used by an LEA to analyze the content of tests in relation to the state's objectives. After a grace period the district would also have to show that all of the state skills are covered by the test they wish to employ. The grace period permits districts to use present tests that meet technical and bias

standards, but for which there is not a perfect match between the state skills and test items. This grace period permits new tests to be reviewed and gives publishers time if they wish to tailor tests for use in the Massachusetts Basic Skills Improvement Policy.

This two stage approval process overcomes several difficulties inherent in the Department's attempt to approve tests in terms of all three of the review components: content validity, technical standards and bias. The first difficulty is that the State review did not take entire batteries into consideration. The second difficulty relates to how overall approval would be determined. There are two possibilities. First, the test must be approved in terms of all three criteria. Failure to meet one criterion would result in the test not being approved. The second possibility is to arrive at some overall, or total cut score or standard, that the tests must meet. The problem here is that this permits weaknesses in one area to be compensated for by strengths in other areas. For example, the content section has 5 points associated with it; 7 if one includes points for readability. The bias section of the review has 3 points associated with it. (A separate bias review was also carried out by the State and need not concern us in this discussion.) The technical section of the review had 56 points associated with it. Clearly, the technical component would be most heavily weighted in any attempt to use some overall total score. In fact, the technical component would completely swamp the other two components. Consider two tests A and B. Test A receives 7 points for content match (none in fact did), 3 points for bias and 46 points for technical characteristics for a total of 56 points. Test B receives 0 points for content, 0 for bias and 56 points for technical characteristics. A total of 56 points. Both tests have the same score but are quantitatively quite different.

A final difficulty involves the issue of the State's attempt to determine content validity for each test for each district, particularly when the State was unable to consider the entire test battery in its content review.

In summary we would recommend the following:

(1) Each test be considered for approval by the State Department in terms of technical adequacy.

(2) Each test be considered for approval by the State Department in terms of bias and stereotyping.

(3) To be included on the State-approved list the tests must meet both the criteria in Steps 1 and 2. Failure on either of the first two steps disqualifies a test.

(4) Each district must review test items on the test(s) they wish to use, and match items to State objectives. However, no specific requirement for a set percentage of content coverage would be required for the present.

(5) A "grace period" be given to enable publishers to develop instruments which would assess <u>all</u> of the state objectives.

(6) After this "grace period" a test would have to measure all State objectives in order to be approved.

APPENDICES A-F

TEST REVIEW INSTRUMENTS

Standardized Achievement Test

- Review Form[1] -

1. Reviewer _____ Date of Review _____

3. Test Name _____

4. Test Publisher _____

5. Publication Date _____

6. Levels (Circle Grade Levels Covered by the Test):

K 1 2 3 4 5 6 7 8 9 10 11 12

7. Which form of the test is being reviewed? _____

8. Is the test being reviewed for Reading Skills or Math Skills? (Circle one)

Reading Math

If you are doing a content review, begin with Question 9.

If you are doing a technical review, begin with Question 13.

CONTENT CONSIDERATIONS

9. How many of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills are measured by at least one item on the test?

____ No. of Skills
____ % of Skills

10. Overall, is the reading level of the items reviewed suitable for most of the students in the lowest grade covered by this test? (Cf. Question 6 above).

YES NO

---

[1]This review form was prepared by Ron Hambleton, George Madaus and Peter Airasian to meet specifications required by the Commonwealth of Massachusetts for use in conjunction with the Massachusetts Basic Skills Improvement Policy.

11. Overall, are the test items free of offensive sexual, cultural racial, and/or ethnic content and/or stereotyping. YES NO

12. If you answered "NO" to question 11, please explain the reasons for your answer, including the type(s) of bias and the item number of any items of concern.

_____

_____

_____

_____

_____

_____

This is the end of the Content Review

TECHNICAL CONSIDERATIONS

13. How many alternate forms of this test are available? ____ No. of forms

14. Is there a Technical Manual which includes information about the test regarding the following ten topics:

a. Item Review Methods . . . . . YES NO

b. Item Analysis . . . . . YES NO

c. Average Item Difficulty . . . . . YES NO

d. Internal Consistency Reliability . . . . . YES NO

e. Test/Retest Reliability . . . . . YES NO

f. Parallel Form Reliability . . . . . YES NO

g. Standard Error of Measurement . . . . . YES NO

h. Content Validity . . . . . YES NO

i. Norms . . . . . YES NO

j. Procedures for screening items for offensive sexual, cultural, racial, and/or ethnic content, and/or stereotyping.. YES NO

15.	How many of the items reviewed meet the standard rules of item writing?	____ No. of items reviewed ____ No. of acceptable items ____ % of acceptable items		
16.	Were item analysis results used to identify "defective" test items?	YES	NO	INA*
17.	Are data bearing on the consistency of mastery decisions (for one or more performance standards or cut-off scores) reported in the Technical Manual?	YES	NO	
18.	Is the consistency of mastery decisions (for one or more cut-off scores) reported in the Technical Manual equal to or above 90%?	YES	NO	INA
19.	Do standard indices of internal consistency reliability reported on the total reading score or total mathematics score reach or exceed .90?	YES	NO	INA
20.	Do standard indices of test-retest or parallel form reliability as reported on the total reading score or total mathematics score reach or exceed .90?	YES	NO	INA
21.	If parallel-forms of the Test are available, do both forms (or multiple-forms, if available) measure equally well the content spanned by the skills included in the Test? (In other words, do the multiple-forms of the Test have equivalent content validity?)	YES	NO	INA
22.	Are the test score norms based on data that is no more than five years old?	YES	NO	INA
23.	Were the norm groups of sufficient size (i.e., at least 300 students)?	YES	NO	INA
24.	Were the samples of students used in the norming study representative of students in the grades for which this test is intended? (Cf. Question 6)	YES	NO	INA
25.	Were the samples of students used in the norming study representative of important strata within the society (i.e., rural pupils, minority group pupils, pupils in large city schools, etc.)	YES	NO	INA

*INA - Information not available

26. Are the test administration directions suitable for students in the lowest grade covered by the test? (Cf. Question 6) YES NO

If "NO", please explain _____

_____

_____

_____

27. Do the test administration directions address the matter of time limits? YES NO

If "NO", please explain _____

_____

_____

_____

28. Do the test administration directions indicate to the student how to handle the problem of guessing? YES NO

If "NO", please explain _____

_____

_____

_____

29. Is the layout or format of the test booklet convenient for students in the lowest grade covered by the test? (cf Question 6) YES NO

If "NO", please explain _____

_____

_____

_____

30. Is the layout or format of the answer sheet convenient for students in the lowest grade covered by the test? (Cf. Question 6) YES NO

If "NO", please explain _____

_____

_____

_____

31. Does the test include practice questions? YES NO

This is the end of the Technical review

Directions for Test Reviewers
- Content Review -

The content review you are about to undertake involves three principal tasks:

a. Deciding whether each of the test items the publisher has nominated as measuring each of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills Policy in fact is appropriate indicators of the skill in question.

b. Deciding whether overall the reading level of the items on the test is suitable for the majority of students in the lowest grade covered by the test.

c. Deciding whether overall the test is free of offensive sexual, cultural, racial or ethnic content and/or stereotyping.

You are asked to make a determination on each of these points by completing the enclosed Review form. Three people will review each test and will meet to arrive at a composite rating for each test. A separate technical review of each test is also being carried out.

To begin the review you should have the following materials in front of you:

a. A copy of the reading or math tests to be reviewed.

b. A list of the test items which the test publisher feels correspond to each of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills Policy.

c. A skills checklist which lists the fourteen reading skills (blue color) or thirty-eight math skills (yellow color).

d. A Standardized Achievement Test Review Form.

e. A Standardized Achievement Test Evaluation Summary Sheet (pink color).

Step A. - Complete the "Basic Information" section of the Standardized Achievement Test Review Form (Questions 1 - 8).

Fill out the background information section on the Skills Checklist and on the Test Evaluation Summary Sheet.

Step B. - Read carefully through the list of skills included in the Skills Checklist.

Read carefully through all the test items on the reading or mathematics test under review.

Step C. - Question 9 on the Review Form

For each skill listed on the Skills Checklist read each item which the publisher has nominated as a measure of that skill. If you agree that the item is a valid indicator of the skill in question, list the item number in the space provided. Once you have finished with a skill, count up the number of items nominated by the publisher which you feel are valid indicators of the skills and place the total number in the blank space provided on the Skills Checklist.

If at least one item nominated by the publisher is a valid indicator of the skill in question you should place a "✓" beside the Commonwealth's skill listed on the Skills Checklist in the box provided.

After you have completed your review of each of the nominated questions for each of the fourteen reading skills or thirty-eight mathematics skills, add up the total number of acceptable items across all the skills and place your total in the space provided at the end of the check list. Next in the space provided write the total number of items on the reading or math test reviewed.

Finally count up the number of "✓" marks (i.e., each skill that has at least one item you feel is a valid indicator of that skill). Place the total number of "✓" in the space provided in Question 9 on the Review Form. Calculate the percent of skills measured by at least one test item. For example, suppose 8 of the Commonwealth's 14 reading skills are measured by at least one item on a Test. You would write "57" in the space provided beside Question 9 for percent of skills included in the test.

Step D. - Question 10 on the Review Form

This item is self-explanatory. Make your decision on the basis of your reading of all the items on the test. For example if the test is designed for 7th, 8th, and 9th graders (indicated in Question 6) the reading level should be appropriate for 7th graders.

Step E. - Questions 11 and 12

Question 11 - After reading through all the items on the test, decide whether overall the test is free of offensive sexual, cultural, racial, and/or ethnic content and/or stereotyping. You should examine all test items to determine whether there is a consistent or overriding pattern of racial, ethnic, cultural, or sexual stereotyping and/or offensive content. Your judgment should be made within the context of the total test. The fact that one or two items portray a woman in the kitchen or a minority group member in an unskilled occupation does not necessarily imply stereotyping. Some women do spend time in the kitchen and some minority group members do hold unskilled jobs. At issue is whether members of such groups are consistently or predominantly portrayed in such circumstances relative to the way in which other groups are portrayed.

Question 12 - Self-explanatory.

Step F. - Transfer the information from the Review Form to the Test Evaluation Summary Sheet.

Thank you your time and effort.

Directions for Test Reviewers

- Technical Review -

The technical review you are about to undertake involves making judgments about certain technical characteristics of tests which are being considered for possible inclusion on a State-approved list of standardized commercial tests. Local school districts may use a test on the list to assess basic skills in reading and mathematics at the secondary level (grades 7-12).

Three people will review each test and will meet to arrive at a composite rating for each test. A separate content review of each test is also being carried out to assess the test's content validity relative to the Massachusetts Basic Skills Policy.

- -

To begin the review you should have the following materials in front of you:

a. Copies of the test to be reviewed.

b. Copies of the Technical Manual for each test.

c. A Standardized Achievement Test Review Form.

d. A Standardized Achievement Test Evaluation Summary Sheet (pink color).

Step A - Complete the "Basic Information" section of the Standardized Achievement Test Review Form, Questions 1 - 8.

Fill out the background information section on the Test Evaluation Summary Sheet.

Step B - Read carefully through the test booklets and the Test's Technical Manual.

Step C - THE TECHNICAL REVIEW BEGINS AT QUESTION 13. Complete each of the following questions on the Review Form:

Questions 13 and 14 - Self-explanatory

Question 15 - Read the technical aid, "Multiple-Choice Item Writing Principles" on page 3, and then randomly select and review 25% of the test items to determine the percent of these test items which do not violate any of the standard rules of multiple-choice item writing. Write the number of items reviewed, the number of acceptable items and the percent of item reviewed which are acceptable in the spaces provided beside Question 15 on the review form.

Question 16 - Check to be sure that item difficulties and item discrimination indices were used in any item analyses. (In constructing criterion-referenced tests, however, the latter is a more important and useful statistic.

INA means Information Not Available.

Questions 17 and 18 - Check for the proportion of agreement in decision-making across parallel-form or retest administrations. Alternately, check to see if the statistic, k, is reported. It reflects the proportion of agreement over and above agreement which is due to chance alone.

Questions 19 and 20 - The test manual will most likely report numerous reliability indices. In general, do these indices reach or exceed .90?

Question 21 - Check to see if the content validity of two (or more) forms is the same. Often the Technical Manual will discuss content emphases and summarize the relevant information in charts or tables. If this information is not satisfactory the parallel forms will be reviewed separately another time by another review committee.

Questions 22 and 23 - Self-explanatory.

Questions 24 and 25 - Check to see if charts are produced to show the representation of any norms groups. Do they look reasonable?

Questions 26 to 31 - These five questions are self-explanatory.

Step D - Transfer the information from the Review Form to the Test Evaluation Summary Sheet.

THANK YOU FOR YOUR TIME AND EFFORT

Multiple-Choice Item Writing Principles

1. Is the item stem clearly written for the intended group of students?
2. Is the item stem free of irrelevant material?
3. Is a single problem clearly defined in the item stem?
4. Are the answer choices clearly written for the intended group of students?
5. Are the answer choices free of irrelevant material?
6. Is there a correct answer or a clearly best answer?
7. Have words like "always," "none," or "all" been removed?
8. Are likely student mistakes used to prepare incorrect answers?
9. Is "all of the above" avoided as an answer choice?
10. Are the answer choices arranged in a logical sequence (if one exists)?
11. Was the correct answer randomly positioned among the available answer choices?
12. Are all repetitious words or expressions removed from the answer choices and included in the item stem?
13. Are all of the answer choices of approximately the same length?
14. Do the item stem and answer choices follow standard rules of punctuation and grammar?
15. Are all negatives underlined?
16. Are grammatical cues between the item stem and the answer choices, which might give the correct answer away, removed?
17. Are letters used in front of the possible answer choices to identify them?
18. Have expressions like "which of the following is not" been avoided?

- Mathematics Skills Checklist -

Reviewer _____ Date of Review _____

Test Name _____

Place a "✔" beside those skills which are measured by the test.

Mathematics Skills

a. Number and Numeration Concepts

1. Recognize number symbols (17, eighteen), whole numbers (34), fractions (1/2), decimals (3.75), and powers of 10 (10^2). ☐

   List the number of each item which you feel is a measure of this skill.

   _____ ☐

   _____

   Total number of items for this skill _____

2. Identify odd and even numbers. ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

3. Put numbers in numerical order. ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

4. Recognize equivalent fractions $\left[\frac{2}{4} = \frac{1}{2}\right]$ ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

b. Arithmetic Computations

1. Add, subtract, multiply, and divide whole numbers (4069 + 81 + 123, 254 x 17, 16.300 - 100). ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

2. Add and subtract mixed numbers $\left[\begin{array}{r} 2\frac{1}{2} \\ -\ 1\frac{1}{4} \\ \hline \end{array}\right]$ ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

3. Multiply whole numbers or money by fractions. (halves, quarters, thirds). ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

4. Add, subtract, multiply, and divide decimal numbers like money. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

5. Change a fraction to a decimal (1/4 to .25).

   List the number of each item which you feel is a measure of this skill. ☐

   _____

   _____

   Total number of items for this skill _____

6. Find a percent of a number in situations such as simple interest, discounts, commissions, and taxes. ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

7. Use ratio and proportion (mixtures, recipes, scale drawings). ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

8. Use simple formulas (A = l x w). ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

c. Estimation and Approximation

1. Round off numbers to a specified place. ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

2. Total number of items for this skill _____

   Approximate the answer to a computation problem (including discounts and percentages). ☐

   List the number of each item which you feel is a measure of this skill.

   _____

   _____

   Total number of items for this skill _____

3. Estimate length, weight/mass, capacity, time, temperature, area, and volume.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

4. Estimate with money.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

d. Measurement and Geometry

1. Choose an appropriate unit of measurement in the U. S. customary system (for example, feet, pounds, and gallons).

List the number of each item which you feel is a measure of this skill.

_____

_____

2. Total number of items for this skill _____

Choose an appropriate unit of measurement in the metric system (for example, meters, kilograms, and liters).

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

3. Choose an appropriate measurement instrument involving both U.S. customary and metric units.

List the number of each item which you feel is a measure of this skill.

_____

_____

4. Total number of items for this skill _____

Convert common measurements within the same system.

_____

_____

Total number of items for this skill _____

5. Read a scale drawing.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

6. Use a map to compute highway distances.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

7. Relate total cost and cost per unit.

_____

_____

Total number of items for this skill _____

8. Compute by using temperature.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

9. Compute by using time.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

10. Identify right angles and parallel, perpendicular, and intersecting lines like those in a street map.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

11. Recognize that an object has the shape of a square, rectangle, triangle, or parallelogram.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

12. Identify the radius, diameter, and center of a circle. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

13. Recognize that an object has the shape of a cube, cylinder, or sphere. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

14. Find the perimeter of a triangle, square, and rectangle. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

15. Find the area of a triangle, square, and rectangle. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

16. Find the volume of a cube or other rectangular solid. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

c. Graphs and Tables

1 Read a table. ☐
List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

2. Interpret a bar graph.

List the number of each item which you feel is a measure of this skill. ☐

_____

_____

Total number of items for this skill _____

3. Interpret a circle graph.

List the number of each item which you feel is a measure of this skill. ☐

_____

_____

Total number of items for this skill _____

4. Interpret a line graph.

List the number of each item which you feel is a measure of this skill. ☐

_____

_____

Total number of items for this skill _____

f. Prediction of Events and Statistics

1. Understand probabilities like those used in weather forecasting or lotteries (the chance something will or will not happen). ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

2. Find and use averages (mean and median) for a group of numbers. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

SUMMARY

Total number of acceptable items over the 38 skills. ☐

Total number of check marks ("✓") over the 38 skills. ☐

Total number of items on the math test itself. ☐

APPENDIX E

- Reading Skills Checklist -

Reviewer _____ Date of Review_____

Test Name _____

Place a "✓" beside those skills which are measured by the Test.

Reading Skills

a. Basic Word Meaning

1. Identify the meaning of commonly used words within a sentence that does not provide clues to the meaning of the word. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill_____

2. Identify the meaning of a word within a sentence that provides clues to the meaning of the word. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

b. Literal Comprehension

1. Identify the meaning of a written phrase, clause, sentence, or paragraph. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill 74

2. Demonstrate the ability to follow directions. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

3. Identify the main idea, supporting details and conclusion of a paragraph. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

4. Recognize the sequence of events or ideas in a written passage. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

5. Identify information on a chart, map, or graph. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

c. Interpretive Comprehension

1. Draw conclusions implied in a paragraph or passage.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

2. Identify cause and effect relationships implied in a paragraph or passage.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

3. Predict an outcome implied in a paragraph or passage.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

d. Evaluative Comprehension

1. Identify a statement as fact or opinion.

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

2. Identify the writer's purpose in a paragraph or passage written to inform or persuade.

_____

_____

Total number of items for this skill _____

e. Locating information

1. Use the parts of a book.

List the number of each item which you feel is a measure of this skill. ☐

_____

_____

Total number of items for this skill _____

2. Locate information in a variety of sources. ☐

List the number of each item which you feel is a measure of this skill.

_____

_____

Total number of items for this skill _____

SUMMARY

Total number of acceptable items for all 14 skills. ☐

Total number of check marks ("✓") over the 14 skills. ☐

Total number of items on the reading test itself. ☐

Standardized Achievement Test Evaluation Summary Sheet

Reviewer _____ Date of Review _____

Test Name _____

Check one - Reading _____ Math _____

Fill in your ratings, determine the points, and write in the score for each question in the space provided.

CONTENT CONSIDERATIONS

Question	Rating	Point System	Score
9	___%	90-100%-5 points 80- 89%-4 points 70- 79%-3 points 60- 69%-1 point < 60%-0 points	____
10	___	Yes - 2 points No - 0 points	___
11	___	Yes - 3 points No - 0 points	___
12		No points	

TOTAL CONTENT POINTS . . . . . ☐

TECHNICAL CONSIDERATIONS

Question	Rating	Point System	Score
13		No points	
14	a____ b____ c____ d____ e____ f____ g____ h____ i____ j____	Yes - 1 point No - 0 points for each item "a" through "j"	a____ b____ c____ d____ e____ f____ g____ h____ i____ j____

Question	Rating	Point System	Score
15	___%	90-100%-5 points 80- 89%-4 points 70- 79%-3 points < 70%-0 points	___
16	___	Yes - 3 points No - 0 points INA*- 0 points	___
17	___	Yes - 1 point No - 0 points	___
18	___	Yes - 5 points No - 0 points INA - 0 points	___
19	___	Yes - 5 points .80-.89-3 points .70-.79-1 point less than .70-0 points INA - 0 points	___
20	___	Yes - 5 points .80-.89-3 points .70-.79-1 point less than .70-0 points INA - 0 points	
21	___	No points However if No or INA then alternative forms of the test are subject to a separate review at another time	
22	___	Yes - 2 points No - 0 points INA - 0 points	___
23	___	Yes - 2 points No - 0 points INA - 0 points	___

*INA - "Information not available"

Question	Rating	Point System	Score
24	___	Yes - 3 points No - 0 points INA - 0 points	___
25	___	Yes - 3 points No - 0 points INA - 0 points	___
26	___	Yes - 2 points No - 0 points INA - 0 points	___
27	___	Yes - 2 points No - 0 points	___
28	___	Yes - 2 points No - 0 points	___
29	___	Yes - 2 points No - 0 points	___
30	___	Yes - 2 points No - 0 points	___
31	___	Yes - 2 points No - 0 points	___
		TOTAL TECHNICAL POINTS	☐

APPENDIX G:

COMPLETE TECHNICAL RATINGS OF

READING AND MATH TESTS

Ratings on Each Item of Technical Screening of Reading Tests

TEST / ITEM	14a Item Review Methods	b Item Analysis	c Ave. Item Diff.	d Intern. Consist Reliab.	e Test/Retest Reliab.	f Parallel Forms Reliab.	g Stand. Err. Meas.	h Content Validity	i Norms	j Proc. on Offens. Rems	15 Item Writing Rules	16 Item Analysis	17 Data on Consist. of Mastery	18 Consist. of Mastery above .90	19 Internal Consist. above .90	20 Test/Retest or Par. Forms above .90	21 Content Valid. of Multiple Forms	22 Norms within Five Years	23 Size of Norm Group	24 Grade Rep. of Norms	25 Strata Rep. of Norms	26 Grade Suit. of Norms	27 Directions on Time Limits	28 Directions on Guessing	29 Grade Suit. of Test Format	30 Convenience of Ans. Sh.	31 Practice Questions	Total Score
Basic Skills Assessment	1	1	1	1	0	0	1	1	1	0	5 94%	3	1	0*	5	0*		2	2	3	3	2	2	2	2	2	2	43
California Achievement Tests - 17	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	3		2	2	3	3	2	2	0	2	2	2	46
California Achievement Tests - 18	1	1	1	1	1	1	1	1	1	1	5 94%	3	0	0*	5	3		2	2	3	3	2	2	0	2	2	2	46
California Achievement Tests - 19	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	3		2	2	3	3	2	2	0	2	2	2	46
Comprehensive Tests of Basic Skills - Level 3	1	1	1	1	1	1	1	1	1	1	5 95%	3	0	0*	5	3		0	2	3	3	2	2	2	0	2	2	44
Comprehensive Tests of Basic Skills - Level 4	1	1	1	1	1	1	1	1	1	1	4 85%	3	0	0*	5	3		0	2	3	3	2	2	2	0	2	2	43
Cooperative English Tests Reading Comprehension	1	1	1	0	1	1	1	1	1	0	5 95%	3	0*	0*	0*	5		0	2	3	0*	2	2	2	2	2	2	38
Gates-MacGinitie - Level E, Reading Comprehension	1	0*	0*	1	0*	1	1	0*	1	1	5 91%	0*	0	0*	5	0*		2	2	3	3	2	2	2	2	2	2	38
Gates-MacGinitie - Level E, Vocabulary	1	0*	0*	1	0*	1	1	0*	1	1	5 100%	0*	0	0*	5	0*		2	2	3	3	2	2	2	2	2	2	38
Gates-MacGinitie - Level F, Reading Comprehension	1	0*	0*	1	0*	1	1	0*	1	1	5 100%	0*	0	0*	5	0		2	2	3	3	2	2	2	2	2	2	38
Gates-MacGinitie - Level F, Vocabulary	1	0*	0*	1	0*	1	1	0*	1	1	5 91%	0*	0	0*	5	0*		2	2	3	3	2	2	2	2	2	2	38
Iowa Tests of Basic Skills - Level 13, Reading	1	1	1	1	1	1	1	1	1	1	5 100%	3	1	0*	5	5		2	2	3	3	2	2	0	2	2	2	49
Iowa Tests of Basic Skills - Level 13, Vocabulary	1	1	1	1	1	1	1	1	1	1	5 100%	3	1	0*	5	5		2	2	3	3	2	2	0	2	2	2	49
Iowa Tests of Basic Skills - Level 14, Reading	1	1	1	1	1	1	1	1	1	1	5 90%	3	1	0*	5	5		2	2	3	3	2	2	0	2	2	2	49
Iowa Tests of Basic Skills - Level 14, Vocabulary	1	1	1	1	1	1	1	1	1	1	5 100%	3	1	0*	5	5		2	2	3	3	2	2	0	2	2	2	49

* Information not available.

Ratings on Each Item of Technical Screening of Reading Tests

TEST / ITEM	14a Item Review Methods	b Item Analysis	c Ave. Item Diff.	d Intern. Consist Reliab.	e Test/Retest Reliab.	f Parallel Forms Reliab.	g Stand. Err. Meas.	h Content Validity	i Norms	j Proc. on Offens. Items	15 Item Writing Rules	16 Item Analysis	17 Data on Consist. of Mastery	18 Consist. of Mastery above .90	19 Internal Consist. above .90	20 Test/Retest or Par. Forms above .90	21 Content Valid. of Multiple Forms	22 Norms within Five Years	23 Size of Norm Group	24 Grade Rep. of Norms	25 Strata Rep. of Norms	26 Grade Suit. of Norms	27 Directions on Time Limits	28 Directions on Guessing	29 Grade Suit. of Test Format	30 Convenience of Ans. Sh.	31 Practice Questions	Total Score
Iowa Tests of Educational Development	1	1	1	1	1	1	1	1	1	1	4 80%	0	0	0	0	5		0	2	3	3	0	0	0	2	0	2	29
Metropolitan Achievement Tests - Advanced 1	1	1	1	0	1	0	1	1	1	1	4 86%	3	0	0*	5	0*		2	2	3	3	2	2	0	2	2	2	41
Metropolitan Achievement Tests - Advanced 2	1	1	1	1	0	0	1	1	1	1	5 92%	3	0	0*	5	0*		2	2	3	3	2	2	0	2	2	2	41
SRA Achievement Series - Level F	1	1	1	1	0	0	1	1	1	1	5 95%	3	0	0*	5	0*		2	2	3	3	2	0	0	2	0	2	37
SRA Achievement Series - Level G	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	0*	5	0*		2	2	3	3	2	0	0	2	0	2	37
SRA Achievement Series - Level H	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	0*	5	0*		2	2	3	3	2	0	0	2	2	2	39
Stanford Achievement Test - 1 & 2	1	.	1	1	1	1	1	1	1	1	5 90%	3	1	0*	5	0*		0	2	3	3	2	2	2	2	2	2	44
Stanford Diagnostic Reading Test - Blue Level	1	1	1	1	0	1	1	1	.	0	4 85%	3	1	0*	3	3		0	2	3	3	2	2	0	2	0*	2	38
Stanford Diagnostic Reading Test - Brown Level	1	1	1	1	0	1	1	1	1	0	5 93%	3	1	5	5	3		2	2	3	3	2	2	0	2	0*	0	46
Stanford Task - Level I A	1	1	1	1	0	0	1	1	1	0	5 90%	3	0	0*	5	0*		0	2	3	0*	0	2	0	2	0*	0	29
Stanford Task - Level II A	1	1	1	1	0	0	1	1	1	0	5 90%	3	0	0*	5	0*		0	2	3	0	2	2	0	2	0*	0	31
STEP - Level I		1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	5		2	2	3	3	2	2	2	2	2	2	50

* Information not available

Ratings on Each Item of Technical Screening of Math Tests

TEST \ ITEM	14a Item Review Methods	b Item Analysis	c Ave. Item Diff.	d Interm. Consist Reliab.	e Test/Retest Reliab.	f Parallel Forms Reliab.	g Stand. Err. Meas.	h Content Validity	i Norms	j Proc. on Offens. Items	15 Item Writing Rules	16 Item Analysis	17 Data on Consist. of Mastery	18 Consist. of Mastery above.90	19 Internal Consist. above.90	20 Test/Retest or Par. Forms above.90	21 Content Valid. of Multiple Forms Norms within Five Years	22 Size of Norm Group	23 Grade Rep. of Norms	24 Strata Rep. of Norms	25 Grade Suit. of Norms	26 Directions on Time Limits	27 Directions on Guessing	28 Grade Suit. of Test Format	29 Convenience of Ans. Sh.	30 Practice Questions	31 Total Score
Basic Skills Assessment	1	1	1	1	0	0	1	1	1	0	5 94%	3	1	0*	5	0*	2	2	3	3	2	2	2	2	2	2	43
California Achievement Tests 17	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	3	2	2	3	3	2	2	0	2	2	2	46
California Achievement Tests 18	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	3	2	2	3	3	2	2	0	2	2	2	46
California Achievement Tests 19	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	5	2	2	3	3	2	2	0	2	2	2	48
Comprehensive Test of Basic Skills - Level 3	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	0	0	2	3	3	2	2	2	2	2	2	43
Comprehensive Test of Basic Skills - Level 4	1	1	1	1	1	1	1	1	1	1	5 100%	3	0	0*	5	3	0	2	3	3	2	2	2	2	2	2	46
Cooperative Mathematics - Arithmetic	1	1	0	1	0	1	1	1	1	0	5 92%	0*	0	0*	3	0*	0	2	0*	0	2	2	2	2	2	2	29
Iowa Tests of Basic Skills - Level 13	1	1	1	1	1	1	1	1	1	1	5 100%	3	1	0*	5	5	2	2	3	3	2	2	0	2	2	2	49
Iowa Tests of Basic Skills - Level 14	1	1	1	1	1	1	1	1	1	1	5 100%	3	1	0*	5	5	2	2	3	3	2	2	0	2	2	2	49
Iowa Tests of Educational Development	1	1	1	1	1	1	1	1	1	1	3 79%	3	0	0*	5	5	0	2	3	3	0*	0	0	0	0	2	36
Metropolitan Achievement Tests Advanced 1	1	1	1	1	0	1	1	1	0	1	5 100%	3	0	0*	5	0*	2	2	3	3	2	2	0	2	2	2	41
Metropolitan Achievement Tests Advanced 2	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	0*	5	0*	2	2	3	3	2	2	0	2	2	2	41
SRA Achievement Series-Level F	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	0*	5	0*	2	2	3	3	2	0	0	2	0	2	37

* Information not available

TEST	ITEM	14a Item Review Methods	b Item Analysis	c Ave. Item Diff.	d Interm. Consist Reliab.	e Test/Retest Reliab.	f Parallel Forms Reliab.	g Stand. Err. Meas.	h Content Validity	i Norms	j Prop. on Offens. Items	15 Item Writing Rules	16 Item Analysis	17 Data on Consist. of Mastery	18 Consist. of Mastery above 90	19 Internal Consist. above 90	20 Test/Retest or Par. Forms above 90	21 Content Valid. of Multiple Forms	22 Norms within Five Years	23 Size of Norm Group	24 Grade Rep. of Norms	25 Strata Rep. of Norms	26 Grade Suit. of Norms	27 Directions on Time Limits	28 Directions on Guessing	29 Grade Suit. of Test Format	30 Concentrac of Ans. Sh.	31 Practice Questions	Total Score
SPA	Achievement series- Level C	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	.	5	0*		2	2	3	3	2	0	0	2	0	2	37
SA	Achievement series- Level D	1	1	1	1	0	0	1	1	1	1	5 100%	3	0	0	5	0*		2	2	3	3	2	0	0	.	2	2	39
anford	Achievement Test (Tests 3, 4, 5)	1	1	1	1	1	1	1	1	1	1	5 90%	3	1	.	5	0*			2	3	3	2	2	2	2	2	2	44
stanford	Test - Level IA	1	1	1	1	0	0	1	1	1	0	5 90%	3	0	0*	5	0*		0	2	3	0*	2	2	0	2	0*	0	31
stanford	Test - Level IA	1	1	1	1	0	0	1	1	1	0	5 100%	3	0	0*	5	0*			2	3	0*	2	2	0	2	0*	0	31
SRA	Level 1 Basic Concepts	1	1	1	1	1	1	1	1	1	1	5 96%	3	0	0*	5	5		2	2	3	3	2	2	2	2	2	2	50
SRA	Level 5 Mathematic Computation	1	1	1	1	1	1	1	1	1	1	5 94%	3	0	0*	5	5		2	2	3	3	2	2	2	2	2	2	50

* Information not available