MAJOR FIELD TESTS
Description of Test Reports

NOTE: Some of the tests do not include Assessment Indicators and some tests do not include Subscores.

■ Departmental Roster

Includes scale scores for all students tested, listed alphabetically by last name. Total scores are reported on a scale of 120-200; subscores (if the test has subscores) are reported on a scale of 20-100. For the MBA total scores are reported on a scale of 220-300.

■ Departmental Summary: Total Test and Subscores

Includes frequency distributions for total scores and subscores (if the test has subscores) showing the percent of examinees scoring below each decile. The departmental mean scale score and standard deviation are also shown. A Departmental Summary does not include data from answer sheets on which fewer than 50% of the questions on one or both parts of the test are unanswered.

Extremely small groups cannot yield meaningful data; therefore, the Departmental Summary Report requires a minimum of five answer sheets.

■ Departmental Summary: Group Assessment Indicators

Lists the mean (average) percent correct score and standard error of measurement for the group as a whole on each assessment indicator in the test. A graphic representation of each score relative to the possible range of scores on the test, with an error band of plus or minus two standard errors of measurement around the mean, is also provided. Assessment indicators are not reported for individual students.

Extremely small groups cannot yield meaningful data; therefore, the Assessment Indicator report requires a minimum of five answer sheets.

■ Departmental Demographic Summary

Student demographic information taken from the answer sheets is summarized for the group as a whole.

■ Additional Questions (Locally Written) Summary

If locally written questions were used, a summary for the group as a whole is provided showing the number and percent responding to each option choice to each question. Responses are not reported for individual students.

■ Individual Student Reports

Includes total score and subscores (if the test contains subscores) for each student tested. A band of plus or minus two standard errors of measurement around the score is provided along with interpretive information.

■ Subgroup Reports (roster and summary reports listed above for each subgroup, if requested)
GUIDE TO MFT COMPARATIVE DATA

Background and Development

The Major Field Tests are objective, end-of-program tests in many major disciplines. From their introduction in 1989, each of the tests is periodically reviewed for currency and each is completely revised at least once every five years. A note about the naming of the tests: during the first revisions of each original test, the Roman numeral II was added to the name distinguishes the new form from its predecessor. Beginning in 2000, that practice was discontinued. The name of the major field and the ETS "Form Code" of the test (printed on every test book) is now used to identify the tests.

Scores on these tests provide useful information for institutions seeking outcomes measures for self-study and for faculty in measuring the progress of their students and evaluating their curriculum. The Major Field Tests provide reliable data for individual and group measurement by assessing student learning in major fields of study.

### The Undergraduate Major Field Tests

<table>
<thead>
<tr>
<th>Biology</th>
<th>Criminal Justice</th>
<th>Literature in English</th>
<th>Political Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Economics</td>
<td>Mathematics</td>
<td>Psychology</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Education</td>
<td>Music Theory and History</td>
<td>Sociology</td>
</tr>
<tr>
<td>Computer Science</td>
<td>History</td>
<td>Physics</td>
<td></td>
</tr>
</tbody>
</table>

### The Graduate Major Field Test

Masters of Business Administration

Test Content

The content specifications for the Major Field Tests reflect the basic knowledge and understanding gained in the curriculum. They have been designed to assess the mastery of concepts, principles, and knowledge expected of students at the conclusion of a major in specific subject areas.

In addition to factual knowledge, the tests evaluate students' abilities to analyze and solve problems, understand relationships, and interpret material. They contain questions that call for information as well as questions that require interpretation of graphs, diagrams, and charts based on material related to the field.

Go to [www.ets.org/hea/mft/](http://www.ets.org/hea/mft/) for a brief outline of each test, showing the specific questions that contribute to each of the subscores and/or assessment indicators.
Test Construction

The Major Field Tests are constructed according to specifications developed and reviewed by committees of experts in each subject area. Selected experts in appropriate ETS subject matter areas review the specifications, each question, and the completed tests before the tests are made available for use. The test development process includes an extensive review of each test question to eliminate language, symbols, or content considered to be potentially offensive, inappropriate for any subgroups of the test-taking population, or serving to perpetuate any negative attitudes that may be conveyed to these subgroups.

After a new form of a test is introduced, and before score reports are released, each test undergoes a rigorous statistical analysis to see whether each question yields the expected result. Such an appraisal sometimes reveals that a question is not satisfactory. A question that proves to be ambiguous or otherwise unsuitable is not used in computing scores.

Statistical properties of each question, such as difficulty level and correlation with the total score, are on record or are computed when new or revised test forms are first administered, to help ensure that each question contributes meaningfully to the test results. For each test, the aim is to provide an instrument that measures the subject matter and skills a student should gain from his or her major area of study.

Scores

Three types of scores are provided for the Major Field Tests:

- **Individually Reliable Total Scores**
  
  Each test yields an individually reliable total score for each student. An individually reliable score is one with statistical properties such that decisions about individual students can be made based on the scores. The length of the test and content coverage are factors in determining score reliability. Total scores are reported on a scale of 120-200 for the undergraduate titles, and 220-300 for the graduate titles.

- **Individually Reliable Subscores** *
  
  In addition to the total score, subscores represent achievement in broad areas within the field reflecting students’ strengths or weaknesses by area within their major. Subscores are reported on a scale of 20-100.

- **Group Reliable Scores** *

  Assessment Indicator scores result from clustering test questions that pertain to a subfield within a major field of study. Assessment indicator scores are reported as mean (average) percent correct for a given group of students.

  By obtaining data on the performance of total groups of students, it is possible to report group scores by means of the reduced number of questions that constitute the assessment indicators. However, a minimum of five students is required for any test in order for assessment indicators to be reported. **Assessment indicator scores cannot be reported for individual students.**

  The assessment indicator approach to academic outcomes measurement increases an institution's ability to examine the performance of groups of students on various elements of the curriculum. In addition, a department receives a more precise interpretation of test results, including those areas that are covered by the test but are not part of the department's curriculum.

* Note that not every test reports subscores AND assessment indicators. Check the subject-specific test description at [www.ets.org/he/a/mft/](http://www.ets.org/he/a/mft/) for a summary of reported scores.
INSTRUCTIONS FOR USING COMPARATIVE DATA TABLES

Overview

This Guide provides assistance in using the tables to interpret scores from the Major Field Tests. Tables are arranged alphabetically by test. The tables report data for seniors only, from institutions testing five or more seniors. Access the Comparative Data Tables for each test from our web site at www.ets.org/hea/mft/compare.html.

The following tables are reported within each subject:

- Distributions for Individual Student Total Scores Means, and Standard Deviations
- Distributions for Individual Student Subscores (where appropriate), Means, and Standard Deviations
- Institutional Mean Score Distributions, Means, and Standard Deviations
- Institutional Mean Subscore (where appropriate) Distributions, Means, and Standard Deviations
- Institutional Summary Data for Assessment Indicators (where appropriate)

In addition, the following tables are based on test administrations from recent testing years beginning when the current form of each test was introduced.

- Correlations among Subscores
- Reliability Coefficients and Standard Errors of Measurement of Subscores
- Mean Scaled Scores and Standard Errors of Measurement for Male and Female Examinees
- Mean Scaled Scores and Standard Errors of Measurement for African American and White Examinees
- Speededness Data

Tables

**Distributions for Individual Students Total Scores and Subscores**

The *distributions* in these tables may be used to interpret individual student results by determining what percent of all those taking the test have scores below that of a particular student.

Tables for total scale scores (and subscores where appropriate) for individual students are provided separately for each test. Each table shows scaled score intervals for total and subscores separately. By looking up the scaled score under the column labeled "Total Score" and reading across the row to the corresponding number in the column headed "% at or Below" the percent of individuals scoring at or below any interval can be determined.

As discussed later in this Guide, distributions based on the students at your own institution may provide a better way to interpret individual results (see Local Comparative Data).
**Distributions for Institutional Mean Scores and Mean Subscores**

The *distributions* in these tables present the number of institutions at each mean score level. These tables provide a way to compare the total score and subscore means for your *institution* with those of all other participating *institutions*. Institutions with fewer than five examinees are excluded. These tables show the mean of means (or the average of the mean scores for those departments in the database) as well as the standard deviations of those means.

**Institutional Summary Data for Assessment Indicators** (where appropriate)

The *assessment indicator* summary information in these tables includes the frequency distribution of departmental means for each assessment indicator. That is, for each assessment indicator, the score means (averages) falling into designated intervals of the score scale are tabulated for departments that tested five or more students in that subject area. These tables show the distribution of institutional rounded means as well as the average and standard deviations of those means.

Since assessment indicators are reported for groups of students only, it is important to recognize that only mean or average scores are involved, both in the reports sent to a department and in the comparative data in the Comparative Data tables.

**Correlations among Subscores**

The *correlations* among the subscores are reported in *Correlations Among Subscores*. Both the observed correlations and the correlations adjusted for unreliability are provided. The corrected correlation adjusts for the unreliability of the two measures and can be interpreted as the expected correlation between “true scores” on the two measures. See the section on Measurement Considerations for an explanation of statistical terms.

**Reliability Coefficients and Standard Errors of Measurement**

Score *reliabilities* and *standard errors of measurement* are reported for each total score and subscore (where appropriate) in *Reliability Coefficients and Standard Errors of Measurement*. See the section on Measurement Considerations for an explanation of statistical terms.

**Mean Scaled Scores and Standard Errors of Measurement for Male/Female Examinees, and African American/White Examinees**

Scores and *standard errors of measurement* for male and female examinees are reported as well as African American and White examinees are reported in Table *Mean Scaled Scores and Standard Errors of Measurement*. For each test, a minimum of 200 students in a subgroup is required for these statistics to be reported.

**Speededness Data**

*Speededness* or rate of completion information is provided in *Speededness Data*. The data include the percent of examinees completing all items in the test, the percent of examinees completing at least 75 percent of the items in the test, the variance of index of speededness, and the number of items reached by 80 percent of the examinees. See the section on Measurement Considerations for explanations of statistical terms.

**Comparative Rather Than Normative Data**

It is important to remember that the data in this Guide should be considered comparative rather than normative because the institutions included in the data do not represent proportionally the various types of higher education institutions. A list of the participating institutions, by test, is included on this website at [www.ets.org/hea/mft/compare](http://www.ets.org/hea/mft/compare). The gender and ethnic makeup of each comparative group, along with other demographic data, are given in the *Summary of Demographic Information*. 
MEASUREMENT CONSIDERATIONS

Accuracy of Test Scores

The accuracy of a test score as a measure of a student's ability, achievement, or progress in the area being assessed is limited by numerous factors, such as the particular questions, the number of questions the test contains, and the amount of time allowed for answering them. Moreover, students perform at different levels on different occasions for reasons quite unrelated to the characteristics of the test itself. Therefore, a test can at best provide no more than an estimate of a person's achievement in a content area.

The precision or accuracy of test scores is best described by two statistical terms: reliability and standard error of measurement (SEM).

**Reliability** Reliability is an indicator of the consistency or stability of test scores. It refers to the extent to which scores obtained on a specific form of an assessment, administered under one set of conditions, can be generalized to scores obtained on other forms of the assessment, administered under other conditions. Reliability can also be viewed as an indicator of the extent to which differences in test scores reflect true differences in the knowledge or ability being tested rather than random variation caused by such factors as the form of an assessment, the time of administration, or the scoring method.

The theoretical concepts of “truth” and “error” are important in the study of reliability. Theoretically, if a student takes an infinite number of equivalent editions of a test, the scores obtained would vary but would cluster around the student's true score. The "true score" would be the average score over the infinite number of replications. A true score is not necessarily a correct score. A consistent or systematic error that occurs in every sample would be counted as truth. Error is best thought of as the random fluctuations from sample to sample. Error can be thought of as differences between a person’s "true score" and the obtained test scores (the sample scores). When a test is administered to a group, we assume that the variance in the distribution of scores is due partly to real differences in ability and partly to random errors (error variance). Reliability is the proportion of total variance attributed to true variance. That is,

\[
\sigma_{\text{total}}^2 = \sigma_{\text{true}}^2 + \sigma_{\text{error}}^2 \\
\text{Reliability} = \frac{\sigma_{\text{true}}^2}{\sigma_{\text{total}}^2}
\]

For example, if the reliability of a test is 0.95, then it can be interpreted to mean that 95% of the differences among the scores can be attributed to true differences in examinees' abilities rather than random fluctuations.

Since we can never know a person’s true score, we can never measure reliability directly. Instead, we can estimate reliability in various ways. It may be derived from comparisons of scores obtained by individuals on two parallel forms of a test (alternate-form reliability), from comparisons of scores obtained by individuals on repeated administrations of the same test (test-retest reliability), or from analysis of the consistency of the performance of individuals on items within a test (internal-consistency reliability). The numerical indices derived are called coefficients of reliability.

The coefficients of reliability as reported in this Guide are measures of internal-consistency of the test. The internal consistency coefficients look at the measurement of test takers across the items in a single test form. They summarize the relationship between each item and performance on the total test. Various versions of the internal consistency reliability coefficients may be used, depending on the type of items and the structure of the test. The coefficients used here were computed using the Kuder-Richardson Formula (K-R 20)

\[
\text{Reliability} = \frac{n}{n-1} \left[ \sigma_i^2 - \sum_{i=1}^{n} p_i q_i \right]
\]

\[
\text{Reliability} = \frac{n}{n-1} \left[ \frac{\sigma_i^2}{\sigma_i^2} \right]
\]
Where \( n \) = number of items in the test,
\[ p_i = \text{proportion of correct responses to item } i \]
\[ q_i = 1 - p_i \]
\[ \sigma^2 = \text{variance of total scores on the test defined as } \sum (X - \bar{X})^2 / N \]

Where \( X = \) total score for each individual examinee
\[ \bar{X} = \text{mean score} \]
\[ N = \text{number of examinees} \]

For the reported total score, the desired level of reliability is 0.90 or higher. The reliability of a score is directly related to the number of items contributing to the score. Thus, the reliability coefficients for the subscores are usually not as high as that of total score. The internal-consistency reliability estimates include only one source of measurement error: sampling of items. They do not include instability of examinees’ performance over time as a source of measurement error. The more homogeneous a test is (the greater the extent to which all of the items measure the same thing), the higher the K-R 20 reliability will be.

Reliability coefficients for the Major Field Tests are given in Reliability Coefficients and Standard Error of Measurements. While scores on tests with reliabilities less than 0.90 should be interpreted more cautiously, the reliabilities of all the tests offered by the Major Field Tests program are sufficient to provide a sound basis for analysis of data based on groups of students as part of a planning or evaluation process.

**Standard Errors of Measurement (SEM)** No test can be perfectly reliable because a test is only a sample from a larger population of possible questions and possible times and conditions of administration. There are differences between a "true score" and an obtained score. These differences are called errors of measurement. A way of characterizing the typical amount of error is called standard error of measurement (SEM). The SEM is a statistic that indicates the standard deviation of the differences between observed scores and their corresponding true scores.

The interpretation of the SEM is usually made in terms of a statement of probability that the score obtained by an individual is within a certain distance of his or her true score. Theoretically, we can say that the obtained score for an individual will fall between +1 and –1 SEM of the true score 68% of the time, or between +2 and –2 SEM of the true score 95% of the time. The practical problem is that we never know an individual’s true score. However, for large numbers of people with obtained score X, we can say that 68% of them are likely to have true scores between +1 and –1 SEM of X. That allows us to make practical use of the SEM. For example, the SEM of the SAT verbal test is about 30 points. We can say that 68% of the people who score at 500 have true scores no higher than 530 and no lower than 470.

The crucial use of the SEM is to treat each score as a band rather than as a point when using scores to make decisions about people. It is common practice to extend the band one SEM above the obtained score and one SEM below the obtained score.

The Standard Errors of Measurement (SEM) for the Major Field Tests are provided in Mean Scaled Scores and Standard Errors of Measurement.

**Standard Error of the Mean** The accuracy of the mean scores reported for the assessment indicators can be interpreted in terms of the standard error of the mean. (In the following explanation it is assumed that the groups tested at each institution represent a sample and not the entire population at that institution.) If it were possible to choose an infinite number of samples of the same size from a given population and calculate the mean for each of these samples, a frequency distribution of means would result. The mean of this distribution would represent the true population mean. Its standard deviation is called the standard error of the mean. While the true population mean cannot be calculated based on a sample of examinees, it is possible to estimate the standard error of the mean from the observed data for a single group. The larger this single group, the smaller is the standard error of the mean.

Continuing with the hypothetical infinite sampling scheme described above, if an interval is defined by two standard errors of the mean above and below the mean for each sample observed, then one can be confident that 95 percent of the intervals defined in this way would capture the population mean. Thus, the error bands reported on the institutional reports for the assessment indicators reflect a 95 percent level of confidence that the population mean is located within the interval indicated.
**Speededness Data**

Information concerning the speededness or rate of completion of a test is provided in Speededness Data. Speededness data are based on the unanswered items after the last item answered by each student and do not take into account previous items that were omitted. Information in Speededness Data includes:

- Percent of examinees completing all items in the test
- Percent of examinees completing at least 75 percent of the items in the test
- Variance index of speededness (the ratio of the variance of not-reached items to the variance of the number right items)
- Number of items reached by 80 percent of the examinees

As a rule of thumb, a test is usually regarded as essentially unspeeded if at least 80 percent of the examinees reach the last question and if virtually everyone reaches at least three-quarters of the items. A variance index less than 0.15 may be taken to indicate an unspeeded test, and an index greater than 0.25 usually means that the test is clearly speeded. Value between 0.16 and 0.25 generally indicate moderate speededness. However, these are only arbitrary indices, and judgments of speededness should be made in the context of additional data and in the light of the purpose of the individual testing program.

**Local Comparative Data**

One approach that a department may decide to pursue is to develop its own set of tables based on local data accumulated over several years. For institutional purposes, such as studying student performance, planning courses, and gauging the progress of individual students, local comparative data are often more useful than the national data presented in this publication. Not only do local data focus on the specific group of students with whom the faculty is concerned, they also represent student performance in an educational setting with which the faculty is familiar. In developing local comparative data, the total group can be separated into subgroups that differ from one another in important respects. For example, local comparative data by subarea of concentration, major field, and depth of course experience may provide significant distinctions. Comparisons between transfer and non-transfer students, or between students anticipating graduate study and those who are not planning to attend graduate school, may also be made.

Departmental data can be extremely valuable for planning courses. In departments where annual enrollments are small, it probably would be necessary to accumulate data for several years to provide adequate reliability. However, some large departments may have enough students to develop local comparative data even within special fields by accumulating scores for only one or two years.
APPROPRIATE USE OF TESTS AND SCORES

There are a number of ways an institution can judge whether these tests are appropriate for its purposes and—if it requires students to take the examinations—whether it is using the scores properly. Some general guidelines related to use of the program's tests and services are discussed below. These guidelines summarize important considerations for appropriate use of the tests and provide examples of normally appropriate and inappropriate uses.

Appropriateness of Test Content

Tests that may be quite appropriate for one department in a specific discipline may be inappropriate for another. An important first step in evaluating a Major Field Test for use at your institution is a content review by appropriate faculty members to determine whether the content and coverage of the test are consistent with the content coverage expected of students majoring in that field at your institution. Departments can obtain a review copy of any of the tests by returning a signed copy of the Confidential Review Copy Request Form to ETS.

The tests cannot measure every discipline-related skill necessary for academic work, nor do they measure other factors important to academic success. Moreover, programs requiring skills other than, or in addition to, academic skills (such as in the performing arts) cannot be fully served by written multiple-choice tests. For these reasons, each department is advised to conduct a careful review of the test's content, possibly in conjunction with empirical studies to determine the test's suitability.

Group and Individual Assessment

A key purpose of the Major Field Tests is to provide information for colleges and universities to use in curriculum evaluation, departmental self-study, and end-of-major outcomes assessment. Major Field Test summary data for a department's group of students can be an important part of the information available to a department or program in its self-evaluation. Test scores, however, should always be used in the context of other sources of information; test scores should never be the only criterion that is used when making decisions about programs or individuals. In particular, a Major Field Test score should not be used as a graduation requirement without careful study to ensure the validity of the test.

As data accumulate on national and local populations, the data may also be useful for individual assessment. If they wish to use the Major Field Test results as part of an assessment of individual students, departments are advised to collect data and evaluate their experience using the test before making the decision to use individual scores.

Criterion or passing scores for individual students should not be adopted unless a careful standard setting procedure has been employed. Decisions about individual students should never be based on Major Field Test scores alone. For guidance in setting local standards, the following publication is available: Passing Scores, A Manual for Setting Standards of Performance on Educational and Occupational Tests, by Samuel A. Livingston and Michael Zieky. Copies of this booklet may be obtained by contacting the Higher Education Assessment area at 1-800-745-0269

Students with Disabilities

The Major Field Tests can be obtained in large print formats, on an audiocassette, or in Braille form for test takers with disabilities. Allow 6 weeks for preparation of a test in one of these formats. In addition, allowing additional time for test takers with certain disabilities is at the discretion of the test administrator. The scores of such test takers (if they are seniors) are included in MFT Comparative Data.
Appropriate Use of Scores

Those using Major Field Test scores should be familiar with the statistical concepts of testing, such as percentile ranks, validity, and reliability. Use the following guidelines in interpreting your test scores and using the results.

- Scores are provided for groups as small as five students; however, care must be taken in interpreting the results from very small groups. For example, the reliability of the scores for small groups would be appropriate for evaluating curriculum but not appropriate for teacher evaluation or for group-to-group comparisons.

- Consider factors that influence performance when comparing groups. Differences may occur if your group represents a different selection of students than the comparative group, for example.

- When the number of students is very small, keep in mind that the mean score can be affected greatly by one or two students with either very high or very low scores.

- In comparing your group to the comparative group, look at assessment indicator results as well as total scores to determine if significant differences in group performance might be due to curricular choices as well as student achievement.

Administrators should periodically review the ways in which they are using the tests at their own institutions to confirm their continued appropriateness. ETS Higher Education Assessment program staff will be pleased to provide suggestions and advice regarding the use of the Major Field Tests.

Background Information Questions

Background information is collected on each answer sheet for the purpose of gathering data in group form about students' backgrounds, academic preparation, etc. Answers to these questions do not affect a student's test scores but can help to define comparative groups. Responses to these questions are summarized and reported as part of the scoring services.

Confidentiality of Information

All Major Field Test data are considered confidential and are reported only to the institution involved. Data aggregated across institutions are provided as comparative data; individually identifiable information is available only to the institution involved. Information about an institution gathered through the testing program will not be released in any form attributable to or identifiable with the institution unless ETS has received written authorization to do so. Answer sheets are kept for a period of twelve months and then destroyed. Score report data are kept on file at ETS for five years and then destroyed. ETS protects the confidentiality of all information about individuals as well as institutions.

The confidentiality of student information and scores should be recognized and maintained. We suggest that institutions obtain a general written authorization from students to the effect that certain faculty members and others who are directly concerned with the students' education will have access to students' scores.
SUMMARIES OF DEMOGRAPHIC INFORMATION

There is a demographic summary of the individuals taking the tests for the most recent five-year period. This table shows Gender, Ethnic Subgroup, Transfer Status, Enrollment Status, English as Best Language, Undergraduate GPA, and Major GPA. The total number of tests taken in each subject area and the proportions of test takers in each subcategory are also presented.

Because the Major Field Tests are revised on a five-year cycle, the sample of test takers included in the Comparative Data Guide will increase for several years and then sharply decline, as a revised test is included.