DOCUMENT RESUME

ED 251 483 TM 840 693

AUTHOR	Schrader, William B.
TITLE	The Graduate Management Admission Test: Technical Report on Test Development and Score Interpretation for GMAT Users.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Management Admission Council, Princeton, NJ.
PUB DATE	[84]
NOTE	24p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Administrator Education; Business Administration; *College Entrance Examinations; *Graduate Study; Higher Education; Item Analysis; *Scoring; Student Characteristics; *Test Construction; Test Format; Testing; Testing Programs; Test Interpretation; Test Items; Test Manuals; Test Reliability; Test Validity
IDENTIFIERS	*Graduate Management Admission Test

ABSTRACT

This report provides information on test development, test administration, and score interpretation for the Graduate Management Admission Test (GMAT). The GMAT, first administered in 1954, provides objective measures of an applicant's abilities for use in admissions decisions by graduate management schools. It is currently composed of five sections: (1) Reading Comprehension; (2) Problem Solving; (3) Practical Judgment; (4) Data Sufficiency; and (5) Usage. New test forms are developed systematically from a test item pool. New items are pretested at regular test administrations and evaluated empirically. Test specifications are the blueprints for assembling a final test form. Uniform test administration conditions are maintained through responsible supervision and careful test security procedures. In addition, several publications provide all examinees with GMAT test information and test taking strategies. The GMAT score scale has a mean of 500, a standard deviation of 100 for the base group, and a possible range from 200 to 800. Angoff's methods are used to equate scores. Continuing validity studies (predictive, content, and construct) provide an important basis for score interpretation information on reliability, standard error, descriptive statistics, and biographical data for examinees are also given to assist in score interpretation. Appendices contain sample GMAT test items and methods for calculating reliability coefficients and equating parameters. (BS)

THE GRADUATE MANAGEMENT ADMISSION TEST:

Technical Report on Test Development and Score Interpretation for GMAT Users

William B. Schrader

Prepared for the Graduate Management Admission Council
with the advice and assistance of
The Graduate Management Admission Council Selection Committee
and a number of colleagues at Educational Testing Service

CONTENTS

THE GRADUATE MANAGEMENT ADMISSION TEST:

Technical Report on Test Development and Score Interpretation For GMAT Users

I. INTRODUCTION

The GMAT was first administered in February 1954 to about 1,300 prospective students of graduate schools of business. Less than a year earlier, in March 1953, a conference at which 12 graduate schools of business were represented had agreed that a nationwide testing program in this area would be useful. There followed a period of vigorous activity. Two meetings of a policy committee formed to guide the new program were held. An important focus of this effort was the identification of suitable abilities to be measured by the test. All necessary steps for scoring, publicizing, developing, and administering the new test were worked out by ETS with the advice and approval of the Policy Committee. The test was called the Admission Test for Graduate Study in Business until 1976. In that year the test name was changed to Graduate Management Admission Test.

From the outset the program was guided by representatives of participating schools. The test, which was the focus of the program, was prepared by ETS test development staff members and was administered, under secure conditions, throughout the United States and in a number of foreign cities. Scoring, reporting, and various statistical services designed to aid in test development and score interpretation were provided. Finally, research aimed at improving program effectiveness was identified as an integral part of program activities. As the program developed over the years, services relevant to admission but not directly concerned with testing were initiated. This report, however, will be limited to matters directly related to the GMAT.

The incorporation, in 1970, of the Graduate Business Admission Council (now the Graduate Management Admission Council) defined explicitly the role of the Council with respect to the test and other program activities. The Council, which consists of representatives of 54 graduate schools of management, is both a service organization and a professional organization. As a service organization it seeks to improve the selection process for graduate management schools by developing and administering appropriate testing instruments, and informing schools and students as to the appropriate use of such instruments and other materials related to the selection process. In addition, it serves as a medium of information exchange between students and schools. As a professional organization it serves as a forum for interchange of ideas and information. The Council sponsors the GMAT; ETS consults with the Council on all matters of general policy affecting program activities that it conducts for the Council.

Purpose of GMAT

The purpose of the GMAT is to provide objective measures of an applicant's abilities for use by graduate management schools as one consideration in making admissions decisions. In order to make the test as useful as possible for this purpose, the test must measure abilities that are relevant to successful performance in graduate management school and that are developed by a wide range of educational experiences, it must be sufficiently long to provide a reasonably dependable measure, it must be administered under uniform, secure conditions, and it must be scored accurately. Finally, scores must be reported promptly in a convenient form and accompanied by materials to aid in their use. When these conditions are fulfilled, the test scores may be relied upon by admissions officers to supplement other data about applicants, particularly previous academic performance.

Evolution of Test Composition

The composition of the test with respect to the abilities measured and the relative weight given to each ability are the characteristics that define a particular test.

In planning the original 1954 form of the test, it seemed clear that both verbal and quantitative abilities were important, and that roughly equal weight should be given to each. Tests of these abilities were considered to be appropriate for students who had enrolled in different undergraduate programs. Tests of these abilities that had proved to be successful in other programs, that seemed appropriate on judgmental grounds, and that could be produced expeditiously were chosen for the 1954 test. The test consisted of four separately timed sections, as follows:

I. Verbal (25 minutes)
II. Quantitative (65 minutes)
III. Best Arguments (30 minutes)
IV. Quantitative Reading (55 minutes)

Beginning in 1955, the Total test score was supplemented by a Verbal part score, based on the Verbal and Best Arguments sections, and a Quantitative part score, based on the Quantitative section. Quantitative Reading items were not included in either part score. The new part scores provided users with information about an applicant's relative standing in verbal and quantitative ability.

The composition of the test was changed in several ways beginning in November 1961. Three new item types, Organization of Ideas, Directed Memory (later called Reading Recall) and Data Sufficiency were introduced, in part because research evidence indicated that they would increase the predictive effectiveness of the test (Pitcher, 1960). The Organization of Ideas section provided an objective measure of an examinee's ability to identify a logical structure within a set of statements. The Directed Memory section measured reading comprehension under conditions that prevented the examinee from referring back to the reading passages when answering questions based on the passages. The Data Sufficiency item type was a measure of quantitative ability based on the examinee's ability to analyze a mathematical problem without carrying out the actual solution. Quantitative Reading, Verbal, and Best Arguments were dropped from the test. Of the five parts, Quantitative and Data Sufficiency defined the Quantitative part score and the other three parts defined the Verbal part score. The test included the following sections.

I. Directed Memory (Reading Recall) (35 minutes)
II. Quantitative (75 minutes)
III. Organization of Ideas (20 minutes)
IV. Data Sufficiency (15 minutes)
V. Directed Memory (Reading Recall) (35 minutes)

In November 1966, Organization of Ideas was replaced by a 20-minute Verbal Omnibus section that included antonyms, analogies, and sentence completion items. Except for this change the basic structure of the test remained the same until 1972, when two 20-minute sections of Practical Business Judgment replaced 35 minutes of the time allocated to Reading Recall. Practical Business Judgmen tems were included in the Verbal part score. At t' same time, the Verbal Omnibus section was shortened from 20 minutes to 15 minutes.

In 1976 several changes were introduced in the composition of the test. Of the two sections that defined the Quantitative part score, the 75-minute Quantitative section that emphasized Data Interpretation items was replaced by a 40-minute Mathematics section that emphasized problem solving items, and the time allotment for Data Sufficiency was increased from 15 to 30 minutes. Several changes were made in the sections included in the Verbal part score. The 15-minute Verbal Omnibus section was replaced by a 15-minute Usage section. The new section called for the identification of errors in Standard Written English. It was introduced in recognition of the importance of written expression in management, a point that was highlighted in the findings of the Casserly and Campbell (1973) survey of skills and abilities needed by graduate students. A further change, introduced in 1977, substituted 30 minutes of Reading Comprehension for the 35 minutes of Reading Recall. It was judged that these item types measured very similar abilities, but that Reading Comprehension would present fewer complications in test administration

The changes introduced in 1976 and 1977 brought GMAT to its present composition, which is as follows

I. Reading Comprehension (30 minutes)
II. Problem Solving (40 minutes)
III. Practical Judgment (40 minutes)
IV. Data Sufficiency (30 minutes)
V. Usage (15 minutes)

A complete recent form of GMAT is published in the *1979-80 Guide to Graduate Management Education*. Appendix A of the present report gives sample items for each of the item types included in the test from 1954 to the present time.

This brief review of the evolution of the test suggests that changes have been gradual and relatively infrequent. Thus, there is a strong continuity within the test over the years, a condition that is highly desirable if scores earned at different times are to be treated as comparable.

II. DEVELOPING A NEW FORM OF GMAT

In an ongoing testing program, a systematic plan for developing new forms of the test is essential, particularly to minimize the possibility that examinees will have an opportunity to anticipate some of the questions included in the test. Ordinarily, a new form of the test should measure the same abilities and be at the same difficulty level as previous forms, but should be composed mainly or exclusively of items not included in any previous form. If changes are to be introduced between a new form and its predecessors, they should be made deliberately, not inadvertently, and gradually, not abruptly, in order to maintain comparability between scores earned at different test administrations over a period of several years. This discussion will be based on the more typical situation in which the new form is designed to match earlier forms as closely as possible.

Each new form of GMAT is composed of objective test items that call for the examinee to choose among five options of which only one is the best choice and is scored as correct. The first task in building a new form is to develop a supply of items that measure the abilities tested in the earlier forms and that are as free as possible of identifiable defects. It is also necessary to be able to compare the difficulty of new items in the pool with that of items included in previous tests. The new test form can then be matched with older test forms with respect to overall difficulty.

Developing an Item Pool

The indispensable first step in building an item pool is to create the first draft of an item. There are a number of item-writing rules, but their value is mainly in reducing the proportion of items that are found later to be defective. items do need to be compatible with other items of the same type when used in a test. a fact that item writers consider in developing a possible item. The actual production of items seems to depend mainly on a good understanding of the ability measured by a particular kind of item, perceptiveness in identifying tasks that will be suitable in difficulty for GMAT examinees, and fertility in devising incorrect responses that will attract the less able examinees. This is another way of saying that item writing is an art.

Once the item nas been drafted, however, a series of formal processes can be used to remedy apparent defects, particularly in the way in which the item is expressed. Obviously, faulty grammar, awkardness of expression, and inconsistencies in style can be corrected by competent editing. Review by one or two persons familiar with the item type may identify items that may be ambiguous, especially to examinees who are very high in the ability measured, and items that may inadvertently give clues, particularly to sophisticated test takers, concerning the correct answer. Finally, items are reviewed by persons who are sensitive to expressions that are objectionable to women or to minority groups; items are then revised to remove this kind of defect.

Pretest Item Analysis: Purpose

Items that survive the review processes are next pretested at a regular test administration. Because items do not necessarily work in the way that their authors intended, it is important to make an empirical evaluation of each item before it is permitted to contribute to an examinee's score. Pretesting also makes it possible to control the difficulty level of new test forms. These two purposes correspond to the two main kinds of statistical analyses applied to pretest data. First, the relationship between examinees' performance on each item and their total scores on all items of that particular type helps to identify items that need to be revised or, possibly, discarded. Second, an index of the difficulty of the items, when adjusted to take account of the ability level of the pretest group, is useful in controlling the level of difficulty of the test.

The statistical analysis of each pretested item provides information about the relationship between performance on the item and a score based on a set of similar items in three different ways:

(a) The biserial correlation coefficient between the item and the total score on items of the same type,

(b) The mean score on items of the same type for examinees choosing each of the five options and for students who omit the item; and

(c) For students who rank in each fifth on the total scores, the number choosing each option or omitting the item.

Essentially, the biserial correlation coefficient is an objective index of the extent to which the examinees who answered the item correctly differ in average score from the remainder of the group. The biserial correlation coefficient, which is used for item analysis work at ETS, adjusts the result to take account of the percentage of students who give the correct answer.* This adjustment is considered to make the resulting correlation coefficients more nearly comparable for items at different difficulty levels. Experience in using item analysis results has indicated that items that have a correlation below .30 need to be reviewed with special care to try to find out why examinees who earn high total scores on the set of items do not perform appreciably better on the item than do examinees who earn low total scores.

Pretest Item Analysis: An Example

The detailed steps involved in using pretest data may be illustrated by discussing the example shown in Figure 1.

Information comparing the test score of examinees who give the correct answer with those who choose one of the incorrect answers receives special attention. If too many high-scoring students choose a wrong answer on the item, it often happens that the question is open to misinterpretation. The statistical results are intended only to supplement the search for flaws in the item based on a thoughtful scrutiny of the item by reviewers.

The biserial correlation coefficient for the sample item, shown in the lower right hand corner of the printout, was .54 for the group tested. This is well above the "danger-point" figure of .30. Thus, it is unlikely that the more detailed statistical results available for each option will reveal serious flaws in the item.

A second way of looking at the results is to consider the average test score (expressed on a scale to be described later) of those choosing each option. This analysis makes it possible to spot any wrong answer that seems to be attracting too many above-average students. On the sample item, the average total score on quantitative items for examinees who chose the correct option (designated by an asterisk) is 16.3; the highest average for an incorrect option is 12.5. The average for all 2,000 examinees is 13.0.

One feature of the item analysis procedure designed for the convenience of those who use the results is that the average total score is always set at 13.0 for the group on which the item analysis is based. The standard deviation of Total scores for the total item analysis group is always set at 4.0. Thus, it is easy to tell whether the examinees choosing a particular option are above or below the average of the total group, and by how much. The use of a uniform scale for the test score enables persons who work with item analysis results to get some idea of how well an item is working by looking at the pattern of average total scores for the various options.

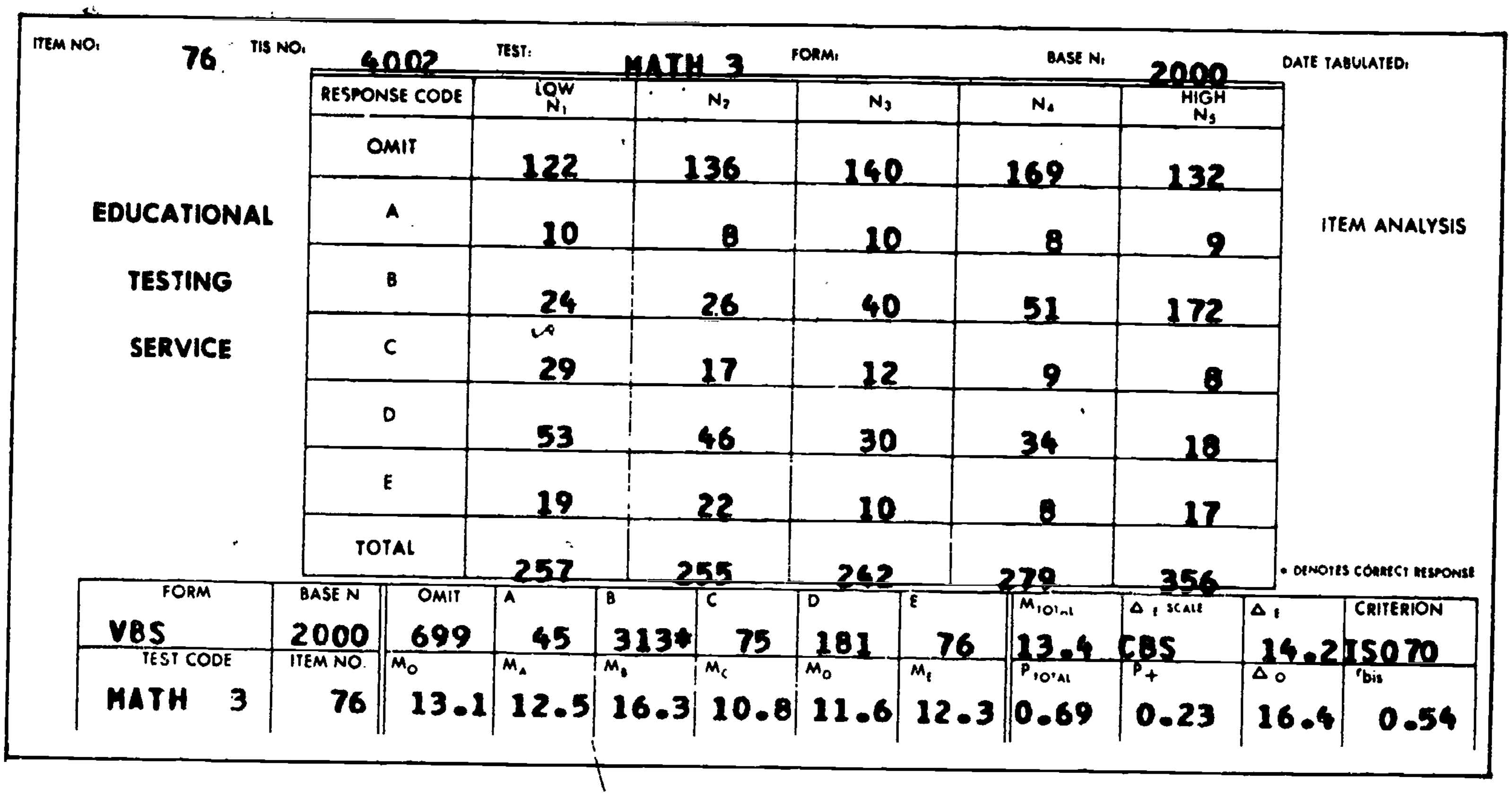
Along with the average test scores, the bottom section of the printout also shows how many examinees

*The biserial coefficient is described more fully in statistical textbooks (e.g. Guilford and Fruchter, 1973).

Figure 1. Item analysis results for a sample item.

A man has exactly enough fencing to enclose a rectangular region 3 times as long as it is wide. He discovers that, if he uses the same amount of fencing to enclose a square region, he can enclose 225 additional square feet. How many feet of fencing does he have?

(A) 30 (B) 120 (C) 150 (D) 675 (E) 900



ITEM NO: 76 TIS NO: 4002 TEST: MATH 3 FORM: BASE N: 2000 DATE TABULATED:

EDUCATIONAL TESTING SERVICE

ITEM ANALYSIS

RESPONSE CODE	LOW N_1	N_2	N_3	N_4	HIGH N_5
OMIT	122	136	140	169	132
A	10	8	10	8	9
B	24	26	40	51	172
C	29	17	12	9	8
D	53	46	30	34	18
E	19	22	10	8	17
TOTAL	257	255	242	279	356

* DENOTES CORRECT RESPONSE

FORM	BASE N	OMIT	A	B	C	D	E	M_{TOTAL}	Δ_E SCALE	Δ_E	CRITERION
VBS	2000	699	45	313*	75	181	76	13.4	CBS	14.2	IS070
TEST CODE	ITEM NO.	M_O	M_A	M_B	M_C	M_D	M_E	P_{TOTAL}	P_+	Δ_O	r_{bis}
MATH 3	76	13.1	12.5	16.3	10.8	11.6	12.3	0.69	0.23	16.4	0.54

chose each option. Of course, the item writer attempts to create wrong answers that will be attractive to examinees and thus to sharpen the differentiation between able and less able students on the ability measured by the item. The extent to which all options are attracting reasonable numbers of examinees tells the item writer how successful he or she has been in formulating effective wrong answers.

Besides the average score for each of the five possible responses, the printout also shows the average total score for examinees who omitted the item. An item is considered to be an "omit" only if the examinee has answered a subsequent item in the separately-timed part of the test under consideration. This definition attempts to distinguish between "omits" (i e., items that are considered but not answered) and items at the end of the test that the examinees may not even have read. In Figure 1, the group of examinees who omitted the sample item is fairly large and their average total score is slightly higher (13.1 vs 13.0) than that for the entire item analysis group. However, because the mean score of examinees who reached the item is 13.4, those who omitted it have a slightly lower score than all examinees who reached it. For this item, a considerable proportion of examinees at all five ability levels omitted it. On the whole, the proportion of omits on this item is larger than would be ideal, but is not sufficiently large to warrant revising it.

Finally, the numbers in the upper portion of the printout provide still greater detail on the relation between total score and responses. The right-hand column (headed "High N_5") shows the number of examinees in the top fifth on total score who gave each response, and the other four columns show the number of examinees in successively lower fifths who gave each response. It will be noted that the numbers for the correct response (B) increase consistently from 25 in the bottom fifth to 172 in the top fifth and that the numbers for the most popular incorrect response (D) show a downward trend. The figure for "TOTAL" shows the number who answered or omitted the item. Because there are exactly 400 examinees in each of the five groups, the difference between the figure reported for "TOTAL" and 400 shows how many did not reach this item. For the top fifth only 44 did not reach it; for the bottom fifth, 143 did not reach it.

A preliminary idea of how difficult an item is can be obtained by dividing the number of examinees who answered it correctly by the number of examinees who reached it. For the sample item, this figure (P_+) turned out to be .23. The figure in the box labeled "M_{TOTAL}" shows the average score on the total test for the examinees who reached the sample item. Because this average is 13.4, we may conclude that the examinees who reached this item were more able than the rest of the item analysis group. We need an estimate of how difficult the item would be for the total item analysis group. The measure of difficulty used for item analysis, called delta (Δ), provides such an estimate.

Delta is so defined that a higher value of delta means a more difficult item and thus a smaller percentage of examinees in the item analysis group who would answer it correctly. It also assumes that a given change in the delta value would have a greater effect on proportion correct for items in the middle of the difficulty range than for those at the extremes. The following table shows the proportion of the item analysis group who would be expected to give the correct answer for items having selected values of delta:

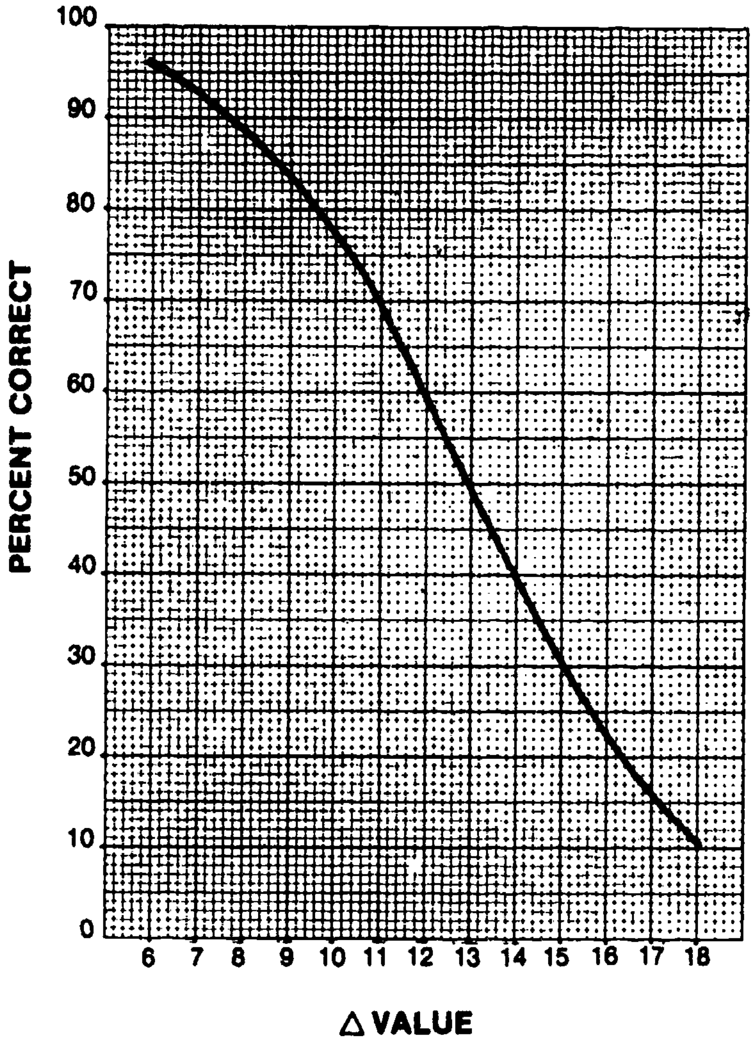
Δ Value	Proportion Correct
17	.16
15	.31
13	.50
11	.69
9	.84
7	.93

The delta scale for item difficulties is defined in terms of a normal curve having a mean of 13 and a standard deviation of 4. Then the percentage of the no*r*mal curve above a particular difficulty value is equal to the percentage of members of the item analysis group who answered the item correctly. Figure 2 illustrates this point.

Because test construction often requires precise measurements of difficulty level, it is necessary to take account of the fact that different item analysis groups differ from each other in ability level. A relatively simple way of adjusting for the difference between two groups is applicable provided that sufficient items (usually 20 or more) have been administered to both groups. Each of these common items will have two values of delta—one for each group. When the pairs of deltas are plotted on ordinary graph paper, they generally fall along a straight line. It is then possible to determine a linear equation relating the two sets of deltas. The resulting equation can be applied to transform a delta obtained on one group to the corresponding delta for the other group. The process of equating item difficulties for a new pretest in continuing programs is facilitated by the fact that any item that has previously been equated can be used in the set of 20 or more items needed for equating the new items.

A delta value calculated for a particular item analysis group is called an observed (or raw) delta (Δ_O). Because item analysis groups vary in ability level, the observed delta for an item will be higher if it was administered to a less able group than if it had been administered to a more able group. Observed deltas can be adjusted statistically so that they represent the difficulty level that each item would have had if it had been administered to a standard reference group. The adjusted delta for an item, called its equated delta (Δ_E), may be compared directly to equated deltas for other items, even though the item analyses were based on different groups.

Figure 2. Percent correct for various values of when all examinees have reached the item.



The sample item shown in Figure 1 has an observed delta of 16.4 and an equated delta of 14.2. Clearly, the item is more difficult for the group on which its item analysis was based than it would have been for the standard reference group to which equated deltas are referred. This result indicates that the group on which the item analysis was based was less able than the standard reference group.

Assembling the Final Test Form

If the items in the item pool are the building blocks from which a new test form is built, the test specifications are the blueprint that guides the construction of a test form for operational use in the testing program. The specifications state the number of items of a particular item type and the level and range of difficulty of the items to be included in each separately-timed section. In the usual case, the specifications will be designed so that the new test form will match recent previous forms in these respects.

For GMAT and other continuing testing programs, number of items in each separately timed section in relation to the time limits is regularly monitored for each new test form. Although several indicators are used for this purpose, the percentage answering the last item may serve to illustrate the principle. As a general guideline, it is considered that, if about 80% of the examinees attempt the last item, the time limits and number of items are reasonably consistent with each other. The results for this indicator for the two most recent forms for which results are available are as follows:

	% Reaching Last Item	
Section	Form A	Form B
Reading Comprehension	81.4	74.8
Problem Solving	20.5	39.2
Practical Business Judgment	88.2	83.0
Data Sufficiency	77.3	70.1
Usage	61.4	64.3
Practical Business Judgment	83.6	90.6

These results suggest that Usage and, to a greater extent, Problem Solving may include a few more items within the allotted time than would be optimal. Particularly for Problem Solving, however, the results may be affected by the tendency of some examinees not to tackle the more difficult items even when they have time to do so. To the extent that this occurs, the percentage attempting the last item cannot be regarded as a satisfactory indicator of the degree to which time allowed and number of items are suitably matched.

If a new item type is to be introduced, it is necessary to make a judgment concerning the number of items that can be completed by the great majority of examinees in the allotted time. This judgment is guided by experience with the item type in pretest studies or in other testing programs. It is also necessary to judge the appropriate level and range of item difficulties so that the new test will be appropriate for GMAT takers.

Once specifications are set, the tasks of selecting a set of items that will fulfill the desired specifications and of arranging the items in a suitable manner can be performed. Items are often arranged in an ascending order of difficulty, but other considerations such as grouping similar items may be given priority in determining the arrangement of items. Finally, as an essential test development step, suitable directions to the examinee must be provided.

The draft test is reviewed to insure that editorial and printing layout rules have been followed and that errors in the item or the designation of the correct answers have not been introduced. Each separately timed part is reviewed with respect to content balance and the test as a whole is reviewed from the viewpoint of how women and ethnic minorities are presented in items that refer to individuals or groups.

III. ADMINISTERING THE TEST

Maintaining Uniform Testing Conditions

Administering the test under uniform conditions is essential if scores earned by different examinees are to be strictly comparable. It is especially important that the time limits for each separately timed test section be uniform for all examinees, that the test directions be fully understood, and that distractions be held to a minimum. Because an examinee who had access to test items before the examination would gain an unfair advantage over other examinees, elaborate precautions are taken to keep test booklets secure before, during, and after the examinations. Examinees are not permitted to use any kinds of extraneous materials (e.g., dictionaries, calculators, notes) during the test and supervisors and proctors are cautioned to be alert to prevent copying. To insure that the person for whom scores are reported is actually the person who took the test, each examinee is asked to provide positive identification and this identification is checked by the supervisor or proctor before the examination begins. From a logical viewpoint, the goal of insuring that each examinee is tested under uniform conditions calls for thorough efforts to preserve test security and to prevent copying and impersonations.

The key to maintaining uniform conditions during the testing sessions is the selection of supervisors who have good judgment and who take a highly responsible attitude toward administering the tests. In addition, the *GMAT Supervisor's Manual* provides specific information on the many detailed tasks that supervisors need to perform. From the time when the examinees have been seated until the examinees are dismissed, each statement to be made in conducting the test is specified by the manual and is read verbatim by the person administering the test in a particular room. Finally, the manual includes a brief form on which the supervisor is asked to report any significant irregularities affecting individual candidates (e.g., illness, defective test materials) or affecting a group of candidates (e.g., mistiming). These Supervisor's Irregularity Reports identify any significant deviations from uniform testing conditions. Each reported deviation is given to an appropriate ETS staff member for action. Ordinarily, the action is based on guidelines or procedures established for handling various difficulties. For example, if a supervisor reports suspected copying, the irregularity is referred to the staff group concerned with test security. Again, if the supervisor discovers that a mistiming has occurred, the report is evaluated by a GMAT program direction staff member, possibly in consultation with test statisticans, to determine whether a special test administration may be needed, or whether some other solution is appropriate.

Because any breach of test security involves a risk that some examinees will gain an unfair advantage, the care that is taken within ETS and by the companies responsible for printing the tests to protect the security of the tests is an essential part of maintaining uniform test conditions. After the tests have been administered, further steps for detecting copying or impersonation may be performed, based on analyses of answer patterns or handwriting comparisons. These procedures are followed if a school questions the scores earned by one of its applicants or if a person repeating the test has shown an exceptionally large score gain. The procedures followed have been carefully designed both to protect the examinee whose score is bona fide and to avoid reporting a score that is not a fair representation of the examinee's ability because he or she has copied or has been impersonated.

Preparing the Examinee for the Test

Over and above the need for maintaining uniform conditions at the test administration, an obligation has been accepted to provide examinees with information about the test and how to approach it. In this way, the possible advantage of more sophisticated test-takers should be minimized. Consequently, the *Bulletin of Information* supplied to prospective examinees goes beyond providing the necessary information about the mechanics of registration and procedures for dealing with exceptional conditions that may arise. In addition, it provides a carefully-prepared set of sample items designed to give examinees a realistic idea of the kinds of items that they may expect to encounter on the actual examination. Moreover, there is a discussion of what GMAT measures and of the role it plays in admission to graduate management schools. This discussion should be helpful, for example, in reassuring examinees who have an exaggerated idea of the importance of test scores in admissions.

In recent years, the Council has published a complete sample test for use by examinees. The most recent publication is in the *1979-80 Guide to Graduate Management Education*. Moreover, both general test-taking suggestions and advice on how to deal with each item type have been provided. An examinee who works through the sample test and who adheres to the time limits for each section should have a very good idea of the demands that the actual test will make. Because an answer key is also provided, and a number of the items are discussed, the prospective examinee may obtain an idea of how well he or she has done, and may be alerted to the need for careful attention to all information given by an item in choosing an answer. The main value of the sample test may well be that working through it enables the examinees to adopt a more realistic attitude in approaching the actual test.

One point of test-taking strategy that has received special attention arises from the fact that GMAT scores are corrected for guessing; that is, a percentage of the number of wrong answers is subtracted from the number of right answers. This procedure is designed to discourage blind guessing. It is important, however, for examinees to understand that they should answer questions about which they have some information even if they are not sure of the correct answer. In order to emphasize this point, the person administering the examination reads the following statement to the examinees:

> "Although you have already read instructions about guessing, they are very important, and I have been asked to summarize them before you begin the test. Your GMAT scores will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. Therefore, it is unlikely that mere guessing will improve your scores significantly, and it does take time. However, if you have some knowledge of a question and can eliminate at least one of the answer choices as wrong, your chance of getting the right answer is improved, and it will be to your advantage to answer the question. If you know nothing at all about a particular question, it is probably better to skip it."

IV. FACILITATING SCORE INTERPRETATION

Defining the Score Scales

An important step in initiating a new testing program is the definition of the score scale in terms of which test performance is reported. The definition of the scale is particularly important when, as in GMAT, there is a continuing program of building new test forms composed mainly or entirely of new test items and yielding scores that are interchangeable with scores on earlier forms of the test. Any change in the definition of the score scale in a continuing program conflicts with strict comparability of scores from one test form to another and thus introduces confusion and possible error.

Several widely-used tests (College Board Scholastic Aptitude Test, Law School Admission Test, Graduate Record Examinations Aptitude Test) have scales so defined that some reasonably appropriate group of examinees has a mean score of 500 and a standard deviation of scores of 100. A further element in the definition may prescribe that no reported score can be higher than 800 or lower than 200. The GMAT scale was so defined that it had a mean of 500 and a standard deviation of 100 for the base group and a possible range from 200 to 800. In establishing the GMAT scale, the base group that was used included all examinees tested in February, May, and August, 1954. In effect, this choice of the reference group assumed that the ability level of examinees in future years would not change so drastically as to require a revision of the definition of the scale. Many graduate management schools find the score range 650-800 important in admissions decisions and a substantial number of examinees score in the 200 to 350 range. There has been no apparent need to expand or contract the range of possible scores. Thus, the definition of the scale on the basis of 1954 examinees remains satisfactory. Of course, the scale value of 500 has not represented the average performance of examinees for many years. The current average score, now about 460, can be found only by consulting descriptive statistics on current examinees.

The choice of the 200 to 800 scale has the advantage that scores on GMAT cannot be confused with IQ's, percentage grades, or percentile ranks. It is true, however, that the use of the same numerical scale for GMAT as for other widely-used admissions tests may occasionally cause confusion if it is thought, for example, that a 500 on GMAT has the same meaning as a 500 on the Law School Admission Test. Both because the various tests measure somewhat different abilities and because the examinee groups used in defining the score scales were different for the different tests, this assumption of comparability is clearly unwarranted.

The score scales for the Verbal and Quantitative tests were established on the basis of the same group used for defining the total score scale. However, the score scales for these tests were set so that the base group would have a mean of 30 and a standard deviation of 8 for these scores. As part of the definition, it was decided that no scaled score for Verbal or Quantitative could be lower than 0 or higher than 60. The use of different units minimizes the risk that the part scores will be confused with the total scores and serves as a reminder to users that these tests were designed to supplement rather than to replace the total score.

Score Equating

In a program that offers tests at several administrations each year, and gives different test forms at each administration, a procedure that will permit scores on the different test forms to be used interchangeably is essential. Only in this way can admission officers confidently assume that scores earned at different test administrations are comparable.

If two different forms of the same test are to yield interchangeable scores, it is important that the composition of the two tests be as similar as possible. Even modest changes in the weight given to various abilities result in some loss of strict comparability. This fact does not preclude changes in the test but it does emphasize the need for considering changes carefully before introducing them. It is also highly desirable that the difficulty level of items in the new form be matched as closely as possible with the difficulty level of items in the previous form. It is this use of item difficulties in test construction that makes the precise determination of item difficulty indexes, discussed earlier in this report, so important.

Assuming that the new test form has been carefully matched with previous forms with respect to abilities measured and difficulty level, score equating completes the process of making scores fully interchangeable between the new form and previous forms.

The method of score equating described in this report was introduced in 1962, and has served as the basic method for score equating in GMAT since that time. It is expected that extensive modifications in GMAT equating procedures will be required as the result of legislation enacted by New York State requiring disclosure of test questions following each test administration. This section, accordingly, should be considered as describing how score comparability was maintained during the period from 1962 to the present time.

This method of equating calls for administering each new form with an old form so that the groups taking each form are substantially equal in ability. In the simplest application of this method, the old form and new form are alternated in each package of testbooks. When the number of examinees taking each form is large, it can safely be assumed that this process will produce groups that are closely matched in ability level. Because the old form has already been equated, it is possible to calculate the mean and standard deviation of reported scores on GMAT To. :l for the group taking it. Then, equating can be done by determining an equation that, when applied to raw scores on the new form, will yield a mean and standard deviation equal to the mean and standard deviation of reported scores on GMAT Total for the group taking the old form. The same procedure is used for equating Verbal and Quantitative scores for the new test form.

The development of the linear equation relating raw scores to reported scores on the new form can be described briefly. For the old form, the equation relating raw scores (X_O) to reported scores calls for multiplying the raw score by a constant (A_O) and adding a constant (B_O). We want to determine values of A_N and B_N for the new form so that the reported score for the new form and the old form will have the same mean and standard deviation for the two equating groups.

If the standard deviations of reported scores for the two equating groups are to be made equal, we may write:

$$A_N \sigma_N = A_O \sigma_O ,$$

where σ_N and σ_O are the raw score standard deviations of the new and old forms respectively. Then,

$$A_N = A_O \frac{\sigma_O}{\sigma_N} .$$

Suppose, for example, that the new form has a larger standard deviation than the old form. Then, $A_N$will be proportionately larger than A_O, and equating will compensate for this difference between the two test forms.

Similarly, if the mean reported scores for the two equating groups are to be equal,

$$A_N M_N + B_N = A_O M_O + B_O ,$$

where M_N and M_O are the mean raw scores for the new and old forms, respectively, so that

$$B_N = A_O M_O + B_O - A_N M_N .$$

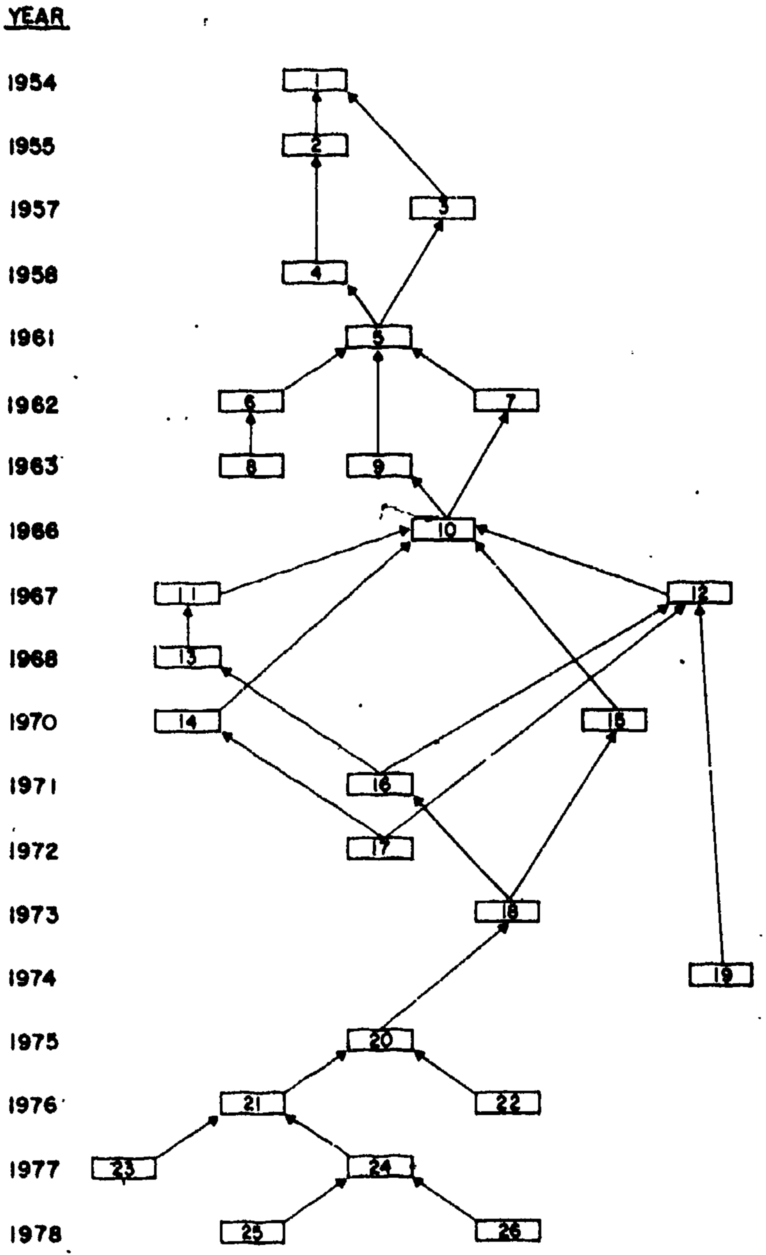
Thus the new multiplier (A_N) and the new additive constant (B_N) may readily be determined using the mean and standard deviation of raw scores on the new and old forms and the A_O and B_O values for the old form.

A simplified example of the way this equation works can be developed if we suppose that A_O is equal to A_N. Then, if the new form is easier than the old form, its mean raw score will be higher than the mean raw score for the old form, so that M_N will be larger than M_O. Because the product of A_N with M_N will be larger than the product of A_O with M_O, application of the equation will make B_N smaller than B_O. Thus, if standard deviations are equal, equating compensates for an easier test by reducing the size of the additive constant. Of course, under realistic conditions, the relative size of B_O and B_N will depend on both the relative size of A_O and A_N and on the relative size of M_O and M_N.

In some instances, three forms are packaged in a sequence, making it possible to equate one new form to two old forms and thus to check on the precision of equating, or to equate two new forms to one old form. A chart showing the linkages involved in maintaining the GMAT score scale is shown in Figure 3. It will be noted that each form shown has been linked directly or indirectly to the first form used in the program.

The basic equating method used for GMAT requires, as a practical matter, that both the old and the new forms be administered in the same testing room. On the rare occasions when this condition cannot be fulfilled, more complex procedures utilizing common items are necessary. A comprehensive discussion of equating methods can be found in Angoff (1971). Appendix B of this report provides a sample of the calculations involved in determining the linear equation relating raw scores to scaled scores.

Figure 3. Genealogical chart showing linkages between GMAT forms.



Validity Studies: Purpose and Background

From the beginning of the program, an important basis for score interpretation has been provided by validity studies. These studies provide objective evidence on the extent to which scores predict subsequent performance. These studies have generally used first year average grades as the measure of academic achievement and test scores and undergraduate average grades as predictors. Because undergraduate grades have a long history of acceptance as one factor in admissions, the relative validity of the test scores and previous grades is a point of special interest. A closely related question is the extent to which the use of test scores along with previous grades results in more effective prediction.

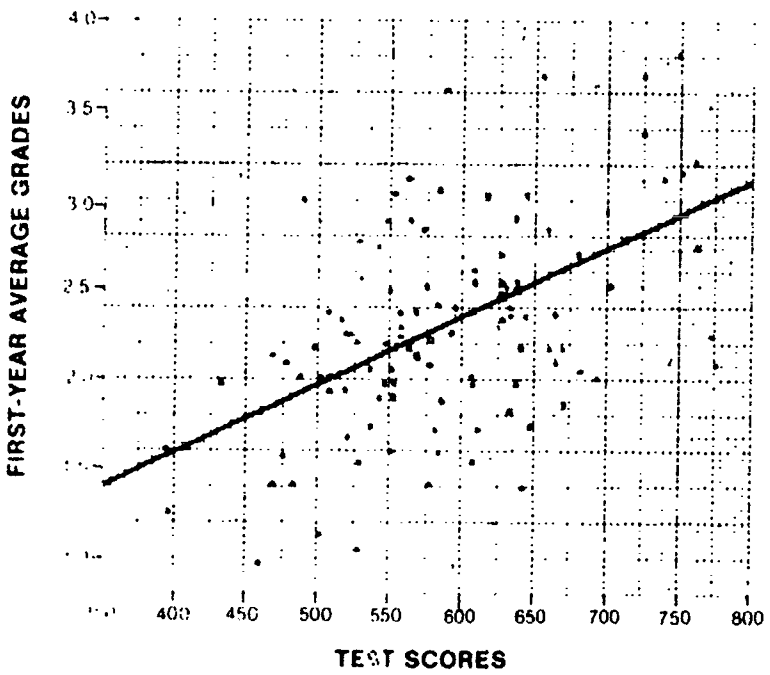
The first validity studies were initiated in 1955, as soon as the students tested in 1954 had earned first-year grades (Olsen, 1957). A second series of studies was initiated in 1958 as part of an effort to evaluate possible new item types for inclusion in the test (Pitcher, 1960). In 1963, all schools represented on the Council were invited to participate in validity studies, and 19 did so. During the three-year period 1967 through 1969-70, 67 graduate schools participated in a comprehensive program in which the study reports were supplemented by regional seminars at which admissions methods as well as study results were discussed (Pitcher, 1972). For a number of years subsequent to this major effort, the responsibility for conducting validity studies rested solely with the schools that use the tests. Recently, a validity study service has been instituted. The new service emphasizes flexibility by facilitating the use of additional predictors beyond test scores and undergraduate average grades, of additional measures of success beyond first-year grades, and of subgroups as well as the total student group. A manual for users of the service has been prepared by Powers and Evans (1977). This manual is included as one section of the *GMAC Handbook*. In 1977-78, 10 schools participated in a pilot study of the new service (Powers and Evans, 1978), and during 1978-79, 25 schools participated in the service. Nearly all participating schools have taken advantage of the options provided by the new system.

Validity Studies: Methods

Because validity study results are customarily expressed in terms of correlation coefficients, a brief discussion of this index should be useful. Figure 4 presents graphically the relationship between test scores and grades for a group of students. Consideration of the plotted points reveals a clear upward trend. The correlation coefficient provides a widely-accepted, objective method of summarizing a set of data of this kind. Essentially, the coefficient measures how closely a straight line fitted to the data describes all the points. When the trend is upward from left to right, the coefficient is given a plus sign; when the trend is downward, the coefficient is given a minus sign. If all the points fall on the line, the correlation is perfect and the coefficient is 1.0. When there is no upward or downward trend in the line, the coefficient is zero. Validity data generally show a clear upward trend with the plotting points scattered above and below the trend line, as in Figure 4. The correlation coefficient provides a convenient way of summarizing a complicated set of data in the form of a single number. Although many other ways of analyzing validity data are useful for various purposes, none has approached the correlation coefficient in general acceptance.

Figure 4. How the line of relation summarizes the main trend of the relationship between test scores and grades. (Each dot represents one student.)

Tne correlation coefficient for these data is about .50.



When results of validity studies are being considered, the question arises as to just how close a relationship a particular correlation represents. Table 1 has been prepared to provide a partial answer to this kind of question. In preparing this table, the group was divided into top fifth, middle three-fifths and low fifth on the predictor and on the measure of success. Then the probability that students standing at each level on the predictor will attain each level on the measure of success was determined using published tables of probabilities for the bivariate normal distribution (Schrader, 1965). For example, with a correlation of .50, a student in the top fifth on the predictor has 44 chances in 100 of scoring in the top fifth and only 4 chances in 100 of scoring in the bottom fifth on the measure of success. Consideration of the data presented in Table 1 provides a reasonable estimate of the strength of relationship represented by different levels of correlation coefficients.

Because the combined effectiveness of two or more predictors is often the primary concern in validity studies, the multiple correlation coefficient, which expresses the correlation between the measure of success and the best-weighted total of scores on two or more predictors, is a valuable tool. For this purpose the weights are determined statistically so that the correlation will be as high as possible for the set of data being analyzed. Thus, the multiple correlation coefficient of graduate school first-year grades with a combination of GMAT scores and undergraduate grades can be compared with the corresponding correlation coefficient based on undergraduate grades only.

Table 1

Relation Between Standing on Predictor and Standing on Criterion for Various Values of the Correlation Coefficient

Correlation Coefficient	Standing on Predictor	Per Cent of Students Standing In Each Criterion Group		
		Bottom Fifth	Middle Three-Fifths	Top Fifth
	Top Fifth	16	60	24
10	Middle Three-Fifths	20	60	20
	Bottom Fifth	24	60	16
	Top Fifth	13	59	28
20	Middle Three-Fifths	20	60	20
	Bottom Fifth	28	59	13
	Top Fifth	10	57	33
.30	Middle Three-Fifths	19	62	19
	Bottom Fifth	33	57	10
	Top Fifth	7	55	38
40	Middle Three-Fifths	18	64	18
	Bottom Fifth	38	55	7
	Top Fifth	4	52	44
50	Middle Three-Fifths	17	66	17
	Bottom Fifth	44	52	4
	Top Fifth	2	48	50
60	Middle Three-Fifths	16	68	16
	Bottom Fifth	50	48	2
	Top Fifth	1	43	56
70	Middle Three-Fifths	14	72	14
	Bottom Fifth	56	43	1
	Top Fifth	0 2	35 4	64 4
80	Middle Three-Fifths	11 8	76 4	11.8
	Bottom Fifth	64 4	35 4	0.2
	Top Fifth	(0 002)	25 2	74 8
90	Middle Three-Fifths	8 4	83 2	8.4
	Bottom Fifth	74 8	25 2	(0 002)

Validity Studies: Results

The primary concern in this summary of validity study results will be with the validity of undergraduate average grades used alone and with the validity of an optimally weighted total of undergraduate average grades, GMAT Verbal scores, and GMAT Quantitative scores. Median correlation coefficients for five studies bearing on this question will be shown, as follows:

(a) 1963-64 results for 17 schools that participated in both the 1963-64 studies and the 1967-70 studies;

(b) 1967-70 results for the same 17 schools;

(c) 1967-70 results for 69 studies conducted for 67 participating schools;

(d) 1977-78 results for 10 schools participating in the pilot study of the new Validity Study Service; and

(e) 1978-79 results for 25 participating schools.

In order to enhance the comparability of results between the earlier and more recent studies, the medians of the multiple correlation coefficients were calculated for the earlier studies, using results published in Pitcher's (1972) survey of pre-1972 studies.

Table 2 shows the results for the five comparisons. Perhaps the most striking finding is the fact that the multiple correlation obtained by the use of three predictors (undergraduate average grades, GMAT-Verbal, and GMAT-Quantitative) is substantially larger than the correlation of undergraduate average grades used alone. The increment in validity attributable to GMAT scores ranges from .16 for the 67 schools in the 1967-70 studies to .22 for the 10 schools in the 1977-78 studies. Except for the first two sets of coefficients shown in Table 2, the interpretation of the comparisons is obscured by the fact that the schools represented in the medians differ from one set of studies to another. The importance of this point is supported by the fact that 1963-64 and 1967-70 studies show virtually identical results when the comparison is limited to schools that participated in both studies. Under these circumstances, it is difficult to evaluate possible trends in the results, particularly because the median validity of undergraduate average grades ranges only from .23 to .29 and the multiple correlation ranges only from .39 to .48 in the five sets of studies.

Table 2

Median Validity Coefficients for Undergraduate Average Grades Separately and in Combination with GMAT Test Scores

		Correlation with First-Year Grades	
Years in which studies were done	Number of Schools	Undergraduate Average Grades	Undergraduate Average Grades, GMAT-Verbal, GMAT-Quantitative
1963-64	17*	29	47
1967 70	17*	28	46
1967 70	67	23	39
1977-78	10	23	45
1978 79	25	27	48

*[illegible] 1967-70 [illegible] studies.

Content and Construct Validity

Tests that are used extensively in admissions have traditionally been evaluated in terms of their effectiveness in predicting academic perform nce. At the same time, the abilities tested have been judged to be relevant to the kind of tasks that graduate management students are called upon to perform. Considerations of this kind have been implicit in the selection of item types to be tried out for possible inclusion and in decisions about what item types to introduce into GMAT. The survey of 19 graduate schools by Casserly and Campbell (1973) represented a more systematic effort to approximate a job analysis of graduate study. Their survey provided strong support for the relevance of the verbal and quantitative abilities measured by GMAT. Their finding that written English was important contributed to the decision to include the Usage section in current forms of GMAT.

In recent years the concept of construct validity has received increasing attention from testers. Construct validation calls for systematic efforts to find out what a test measures, studying the relation of test performance to other variables, and the development and testing of tentative theories that account for the observed results (Cronbach, 1971). The well-established principle that GMAT scores and undergraduate average grades supplement each other in the prediction of academic performance shows how much can be gained by using different measures jointly rather than in isolation. Studies of such factors as age, sex, ethnic group membership, undergraduate major field, and previous business experience may be regarded as steps toward developing construct validity. Also relevant are the long-range prediction studies by Harrell (1969) and by Crooks and Campbell (1974) and studies of factors affecting test peformance such as the speededness study done by Evans and Reilly (1972). Although construct validation presents formidable tasks, it represents a promising approach to better use of test scores.

Reliability of Test Scores

Because GMAT scores often play a significant role in decisions about individuals, high standards of reliability for these scores have been maintained since the program was begun. The importance of reliability for score interpretation arises from the fact that it measures the consistency of *individual* scores from one test form to another. Unless the test scores are highly reliable, an individual's relative standing, and hence his or her score, would show excessive fluctuations.

The logic of reliability is perhaps most readily understood if it is thought of as the correlation coefficient between scores on two forms of the same test. If the two forms are closely matched with respect to the abilities that they measure, and if they include a reasonably large number of questions, we would expect each examinee's relative standing to be quite similar on the two forms. Thus we would expect that, if we administered both forms to each member of a large group of examinees and calculated the correlation coefficient between the scores on the two forms, the resulting reliability coefficient would be relatively high. Rather than calculating reliability coefficients by this relatively cumbersome procedure, it is customary to use certain theoretical developments to estimate what the correlation would be between scores on a test form and a similar form. A more detailed account of the procedures used in calculating reliability coefficients is given in Appendix B.

The reliability coefficients for four recent forms of GMAT are as follows:

Form	Reliability Coefficient of:		
	GMAT Total	Verbal	Quantitative
A	.92	.90	.86
B	.92	.89	.87
C	.92	.90	.88
D	.93	90	.88

These coefficients are based on a cross section of examinees at regular administrations of GMAT. The reliability coefficients for GMAT Total are, as would be expected, somewhat higher than those for the two part scores. However, the part scores may be considered to have acceptable reliability, particularly if both are taken into account in making decisions about individuals. The extent of relationship represented by a correlation of .90 is shown in Table 1.

Standard Error of Measurement

Although the reliability coefficient is useful for judging whether a test yields sufficiently reliable scores to warrant using the scores as one element in important decisions. the standard error of measurement is more useful in judging how much individual scores would fluctuate from one form to another.

If a person could take a large number of forms of the same test, we may assume that his or her scores on these forms would follow a normal distribution with a standard deviation equal to the standard error of measurement. The mean score on all the forms is called the individual's true score. Thus, if the GMAT Total score has a standard error of measurement of 30 and if an examinee has a true score of 500, we would estimate that approximately two-thirds of his or her observed scores would fall between 470 and 530 and 95 percent would fall between 440 and 560. There is no way of knowing, of course. whether the person's observed score on a particular occasion is higher or lower than his or her true score. The main value of the standard error of measurement is in providing some idea of how much variation in observed scores we would expect to find if a person took different forms of the same test.

The size of the standard error of measurement of scores on GMAT for four recent forms was as follows:

Form	Standard Error of Measurement of:		
	Total Score	Verbal	Quantitative
A	29	3	3
B	29	3	3
C	30	3	3
D	28	3	3

Standard Error of a Score Difference

When an examinee repeats a test, a number of factors, including the effect of practice, growth in ability during the interval between tests, and differences in motivation and anxiety may affect the differences in scores. It is possible, however, to estimate the probability that various score differences are attributable to standard error of measurement. For this purpose we need to know the standard error of the difference. If the two tests have equal standard errors of measurement, the standard error of the difference is simply the standard error of measurement times $\sqrt{2}$. Thus, if the standard error of measurement is 30, the standard error of the difference of two scores would be 42. Assuming that the differences are normally distributed, we could conclude that, if the person's true score remains the same, about two-thirds of the differences would be 42 points or less and 95 percent would be 84 points or less. The same reasoning can be followed in estimating the likelihood of various score differences for persons having the same true scores.

Table 3

Percentages of Candidates Tested from November 1975 through July 1978 Who Scored below Selected Total Test Scores

Score	Percentage below
700	99
675	98
650	97
625	94
600	90
575	85
550	79
525	71
500	63
475	53
450	44
425	36
400	28
375	21
350	16
325	11
300	08
275	05
250	03
225	02
Number of Candidates	457,103*
Mean	461
Standard Deviation	107

*Candidates included were self selected

The fact that the standard error of measurement and the standard error of the difference can be computed for test scores offers some aid in interpreting test scores, by indicating the limitations of scores even when tests are professionally constructed and accurately scored. Although standard errors of measurement are not available for non-test measures such as undergraduate average grades, it is plausible that they, too, would vary if the person had attended a different college, followed a different program of courses, or even had different instructors. In summary, although an individual's test scores are facts, inferences about the individual based on these facts are more realistic if account is taken of the fact that the reliability of the scores is less than perfect.

Descriptive Statistics on Test Scores

Although 500 was the mean score for examinees in 1954, the mean of all GMAT takers for November 1975 through June 1978 is 461. Thus, an examinee or admissions officer whc wishes to know how well a score compares with those earned by the whole group of examinees needs information on the distribution of scores for current applicants to graduate management schools who take GMAT. This kind of information is provided in the *Guide to the Use of GMAT Scores* prepared for admissions officers and admissions committees and in *GMAT Candidate Score Interpretation Guide*. The most recent table for GMAT Total from the *1978-79 Guide to the Use of GMAT Scores* is shown in Table 3. Users of this information are reminded that the group includes only those prospective graduate students of management who take the GMAT, and is thus self-selected. From the viewpoint of examinees, these descriptive statistics provide a rough idea of how well they have performed on the test. Because graduate management schools differ substantially in the score level of applicants that they attract and of students whom they enroll, examinees are urged, in the *Bulletin of Information*, to talk with a placement or counseling officer in their undergraduate college. By considering the student's test scores in relation to his or her college record, and by drawing on experience with the success of students with various credentials in gaining admission to various graduate management schools, the placement or counseling officer can often provide a more meaningful interpretation of the scores than is provided by national statistics.

Biographical Information for GMAT Examinees

In recent years, GMAT examinees have been asked to answer nine biographical data questions. For most of these questions, their answers are transmitted as part of the GMAT report to schools that they designate. Two questions, concerned with self-reported language fluency and with population subgroup membership, are asked solely for research purposes, and the responses given by an individual to those questions are never reported. Results for a few questions will be reported here because they help to describe the total group of examinees. Unless otherwise noted, the results are based on 457,730 questionnaires completed in 1975-76, 1976-77, and 1977-78. (Because test repeaters completed a questionnaire each time they were tested, the number of individuals included in the sample is appreciably less than the number of questionnaires analyzed.)

Table 4

Representation of Undergraduate Major Fields in GMAT Examinee Group, 1975-1978

Major Field	Percent of Examinees
Business and Commerce (39.9%)	
Accounting	13.9
Management	9.3
Marketing	5.1
Finance	4.1
Business Education	1.5
Industrial Relations	0.5
Hotel Administration	0.2
Other Business and Commerce	5.3
Social Science (25.2%)	
Economics	8.0
Psychology	3.7
Political Science	3.5
History	3.4
Education	2.0
Sociology	1.9
Government	0.5
Other Social Science	2.1
Science (21.5%)	
Engineering	10.3
Mathematics	3.0
Biological Sciences	2.9
Chemistry	1.5
Computer Science	0.9
Physics	0.6
Architecture	0.4
Statistics	0.2
Other Science	1.7
Humanities (7.1%)	
English	3.0
Foreign Language	1.7
Fine Arts	1.0
Philosophy	0.9
Other Humanities	0.5
Other Major (6.2%)	

Of the total examinee group, 2.5% did not respond to the question on undergraduate major field

Graduate students in management are drawn from a wide spectrum of undergraduate major fields, as shown in Table 4. Roughly two-fifths of the examinees majored in Business and Commerce, about one-fourth majored in Social Sciences, and over one-fifth majored in Science. When major fields are considered separately, Accounting (13.9%), Engineering (10.3%), Management (9.3%), and Economics (8.0%) are most heavily represented in the examinee group. Indeed, these four fields account for more than two-fifths of all examinees. Another background item of special interest concerns the number of months of full-time work experience reported by the examinees. Results for the total group may be summarized as follows:

No response or zero months	16.7%
1-12 months	19.9%
13-24 months	14.6%
25-60 months	23.9%
61 or more months	24.9%

In this group nearly one-half (48.8%) had more than two years of full-time work experience and nearly one-fourth (24.9%) had more than five years of work experience.

Table 5

Representation of Male and Female U.S. Citizens who Reported Membership in Various Population Subgroups in the GMAT Examinee Group, 1975-78

Population Subgroup	Percentage of Examinee Group
American Indian (0.3%)	
Male	0.2
Female	0.1
Black/Negro/Afro-American (6.5%)	
Male	3.8
Female	2.7
Caucasian/White (87.4%)	
Male	64.1
Female	23.3
Mexican American/Chicano (0.7%)	
Male	0.6
Female	0.1
Oriental/Asian (3.0%)	
Male	2.1
Female	0.9
Puerto Rican (0.4%)	
Male	0 3
Female	0.1
Other (1.7%)	
Male	1.3
Female	0.4

Of the total group of examinees 15 5% reported that they were not U S Citizens 4 9% reported that they did not care to respnnd to the question on population membership and 4 8% did not respond to the question

Many GMAT examinees plan to pursue their graduate studies on a part-time basis. In the 1975-78 group, 46.9% reported that they planned to enroll full-time, 41.1% reported that they planned to enroll part-time, and 12.0% were undecided.

Representation of various population subgroups in the 1975-78 examinee group is shown in Table 5. It is a matter of some interest that the six population groups other than Caucasian/White constitute 12.6%, or about one-eighth, of the examinees who are United States citizens and who reported their population group membership.

A separate tabulation of men and women, based on all but two members of the total 1975-78 sample, showed that 74.3% were male and 25.7% were female. (These results differ slightly from the totals for males and females for the sample on which Table 5 was based.) Evidence from other tabulations shows a rising percentage of female examinees: for the 1973-75 group, the percentage was 18.2 and for the 1978-79 group it was 31.5.

The great majority of GMAT examinees takes the examination after college graduation. A tabulation of responses for 1977-78 examinees showed the following distribution of years of graduation (or expected graduation):

1979	2.7%
1978	25.4%
1973-77	49.6%
1972 or earlier	22.3%

In this group, well over two-thirds (71.9%) of examinees were tested after they had completed college, and more than one-fifth (22.3%) were tested more than five years after completing college.

REFERENCES

Angoff, W. H. "Scores, Norms, and Equivalent Scores." In Thorndike, R. L., editor, *Educational Measurement, Second Edition*. Washington: American Council on Education, 1971.

Casserly, P. L., & Campbell, J. T. *A Survey of Skills and Abilities Needed for Graduate Study in Business*. ATGSB Research and Development Brief Number 9. Princeton, N.J.: Educational Testing Service, 1973.

Cronbach, L. J. "Test Validation." In Thorndike, R. L. editor, *Educational Measurement, Second Edition*. Washington: American Council on Education, 1971.

Crooks, L. A., & Campbell, J. T. *Career Progress of MBAs: An Exploratory Study Six Years After Graduation*. Program Report PR-74-8. Princeton, N.J.: Educational Testing Service, 1974.

Evans, F. T., & Reilly, R. R. *The Test Speededness Study: A Study of Test Speededness as a Potential Source of Bias in the Quantitative Score of the Admission Test for Graduate Study in Business* ATGSB Research and Development Brief Number 8. Princeton, N.J.: Educational Testing Service, 1972.

Guilford, J. P., & Fruchter, B. *Fundamental Statistics in Psychology and Education, Fifth Edition*. New York: McGraw Hill, 1973.

Harrell, T. W. *Predicting Job Success of MBA Graduates*. ATGSB Research and Development Brief Number 1. Princeton, N.J.: Educational Testing Service, 1969.

Olsen, M. *The Admission Test for Graduate Study in Business As A Predictor of First-Year Grades in Business Schools, 1954-55*. Statistical Report SR-57-3. Princeton, N.J.: Educational Testing Service, 1957.

Pitcher, B. *The Effectiveness of Tests of Directed Memory, Organization of Ideas and Data Sufficiency for Predicting Business School Grades*. Statistical Report SR-60-33. Princeton, N.J.: Educational Testing Service, 1960.

Pitcher, B., & Winterbottom, J. A. *The Admission Test for Graduate Study in Business as a Predictor of First-Year Grades in Business School, 1962-63*. Statistical Report SR-65-21. Princeton, N.J.: Educational Testing Service, 1965.

Pitcher, B. *Report of Validity Studies Carried Out by ETS for Graduate Schools of Business, 1954-1970*. Statistical Report SR-72-30, Princeton, N.J.: Educational Testing Service, 1972.

Powers, D. E. & Evans, F. R. *Designing Your Validity Study: A Manual for the Graduate Management Admission Council Validity Study Service*. Program Report 77-14. Princeton, N.J.: Educational Testing Service, 1977.

Powers, D. E., & Evans, F. R. *Relationships of Preadmission Measures to Academic Success in Graduate Management Education*. Research Bulletin RB-78-11. (prepublication draft) Princeton, N.J.: Educational Testing Service, 1978.

Schrader, W. B. A Taxonomy of Expectancy Tables. *Journal of Educational Measurement*. 1965, 2, 29-35.

APPENDIX A
Sample Items For Item Types Used In GMAT

This appendix includes examples of each item type that has been included in GMAT since the test was first administered in 1954. For each item type, the month and year of the administration at which it was introduced is stated. For item types not currently in use, the month and year of the first administration at which it was replaced is stated.

For the item type designated as Verbal, three subtypes (Analogies, Antonyms, and Sentence Completion) are shown. For the Quantitative item type, two subtypes (Data Interpretation and Problem Solving) are shown.

Characteristically, item types that involve reading a passage or interpreting a tabular or graphic presentation include five multiple choice items. For most samples given in this appendix, however, only one of the multiple choice items is shown.

The correct response for each sample item is shown, either by starring the correct response or, in a few instances, by showing the correct response in parentheses.

Verbal

Introduced: 2/54
Replaced: 2/61
Restored: 2/66
Replaced: 10/76

Analogies

Directions: In each of the following questions, a related pair of words or phrases is followed by five lettered pairs of words or phrases. Select the lettered pair which best expresses a relationship similar to that expressed in the original pair.

Astronomy : Astrology

(A) chemistry : alchemy* (B) biology : botany
(C) religion : mythology (D) geography : geology
(E) medicine : magic

Antonyms

Directions: Each question below consists of a word printed in capital letters, followed by five words or phrases lettered A through E. Choose the lettered word or phrase which is most nearly opposite in meaning to the word in capital letters.

Since some of the questions require you to distinguish fine shades of meaning, be sure to consider all the choices before deciding which one is best.

DOUR: **(A) blithe* (B) talkative (C) inflexible**
(D) nest (E) modish

Sentence completion

Directions: Each of the sentences below has one or more blank spaces, each blank indicating that a word has been omitted. Beneath the sentence are five lettered words or sets of words. You are to choose the one word or set of words which, when inserted in the sentence, best fits in with the meaning of the sentence as a whole.

The manufacture of cupboards and doors, bathtubs and cooking stoves, taking place as it does in factories, should be unaffected by ....; but since the articles are parts of buildings and there is no demand for them unless buildings are going up, they too are .... in activity.

(A) price . . sluggish (B) cost . . expensive
(C) weather . . seasonal* (D) methodology . . regulated
(E) policies . . unstable

Quantitative

Introduced: 2/54
(Currently in use)

Data Interpretation

Directions: In this section solve each problem, using any available space on the page for scratch work. Then indicate the one correct answer in the appropriate space on the answer sheet.

Per Cent of the Total Value of U.S. Lend-Lease Supplies Received by U.S. Allies

	1st Year	2nd Year
Britain	68%	38%
Russia	5%	30%
All others	27%	32%
Total Value of Supplies (in billions of dollars)	2	8

What per cent of the total value of lend-lease supplies for both years was received by Russia and Britain combined?

(A) 31 (B) 44 (C) 69* (D) 70.5 (E) 141

Problem Solving

If the length of a rectangle is increased by 10 per cent and the width by 40 per cent, by what per cent is the area increased?

(A) 4 (B) 15.4 (C) 50 (D) 54* (E) 400

Best Arguments

Introduced: 2/54
Replaced: 11/61

Directions: The questions in this part are based on situations which involve some sort of dispute or disagreement. In most of the questions you will be asked to evaluate the arguments which might be offered by the disputants; some questions will require you to analyze the situations in other ways. You are to assume that these disputes are being brought before an intelligent lay arbitrator (not a court of law) for decision; the questions, therefore, will not involve any legal precedents or technicalities. You are to evaluate the situations objectively in terms of ordinary concepts of reasonableness and fair play and base your answers on a logical analysis of the facts and arguments as they are presented to you.

Bruce Bond, a broker, one morning overhears a famous financier say, "The price of American Beartrap stock will go sky-high within two weeks." Later that day Pete Goodfellow, an old friend to whom Bond owes many favors, calls on Bond to ask for advice about investments. He emphasizes that he wants to buy some stock on which he can make money quickly because he is in a tight financial spot. Bond says that American Beartrap is the best buy he knows at the moment. When Goodfellow protests that he has never heard of American Beartrap, Bonds replies that the basis of his recommendation is information received from a reliable source. Goodfellow accepts the advice and invests heavily. Within two weeks American Beartrap stock has become virtually worthless, Goodfellow's entire investment is lost, and Goodfellow is ruined financially. Goodfellow thereupon accuses Bond of causing his financial downfall.

Which one of the following arguments best supports Goodfellow's accusation?

(A) Bond should not have presumed to give Goodfellow any advice.
(B) Goodfellow naturally believed that Bond wanted to help him.
(C) Bond had misrepresented his knowledge of the situation.*
(D) Bond should have cautioned Goodfellow not to invest too heavily in the stock.
(E) Bond had taken advantage of Goodfellow's obvious lack of knowledge of financial matters.

Quantitative Reading

Introduced: 2/54
Replaced. 11/61

There have been many suggestions that in an emergency the professional schools, particularly medical schools, accelerate their programs, thus graduating more trained men and women. If more doctors are to be trained we must have more of the three essentials of such a process—teachers, students, and equipment—or we must utilize those which we have to greater effect. But objections have been made to asking students and faculty to work through the four quarters of the year, and the plan herewith submitted, recognizing these objections, attempts instead a fuller utilization of the third essential, equipment and supplies, to realize effectively the objectives of an accelerated program.

The proposed plan, which is essentially the more frequent admission of freshman classes, is designed for those schools which operate on the quarter, as opposed to the semester, plan. Following this plan such a school could graduate four classes in three years by the admission of a freshman class every nine months.

An illustration from the table below may serve to clarify the proposal. In accordance with the plan, one class, indicated on the table by the letter A, would enter medical school as freshmen in the Summer Quarter of 1951, continue in school through three consecutive quarters, and go on vacation during the Spring Quarter of 1952. Students in class B would enter as freshmen in the Spring Quarter of 1952, continue in school through the Summer and Autumn Quarters of 1952, and go on vacation during the Winter Quarter, at which time students in class C would begin their freshman year. As can be seen from the table, there would always be a freshman, sophomore, junior, and senior class in school.

Quarter Plan Organization for Class Acceleration

	Summer	Autumn	Winter	Spring
1951·52	A_1	$A_2X_4Y_7Z_{10}$	$A_3X_5Y_8Z_{11}$	$B_1X_6Y_9Z_{12}$
1952-53	A_4B_2	$A_5B_3X_7Y_{10}$	$A_6C_1X_8Y_{11}$	$B_4C_2X_9Y_{12}$
1953·54	$A_7B_5C_3$	$A_8B_6D_1X_{10}$	$A_9C_4D_2X_{11}$	$B_7C_5D_3X_{12}$
1954-55	$A_{10}B_8C_6E_1$	$A_{11}B_9D_4E_2$	$A_{12}C_7D_5E_3$	$B_{10}C_8D_6F_1$
1955·56	$B_{11}C_9E_4F_2$	$B_{12}D_7E_5F_3$	$C_{10}D_8E_6G_1$	$C_{11}D_9F_4G_2$
1956·57	$C_{12}E_7F_5G_3$	$D_{10}E_8F_6H_1$	$D_{11}E_9G_4H_2$	$D_{12}F_7G_5H_3$
1957·58	$E_{10}F_8G_6I_1$	$E_{11}F_9H_4I_2$	$E_{12}G_7H_5I_3$	$F_{10}G_8H_6J_1$
1958·59	$F_{11}G_9I_4J_2$	$F_{12}H_7I_5J_3$	$G_{10}H_8I_6K_1$	$G_{11}H_9J_4K_2$

Classes X, Y, and Z are included in the plan

(A) as the second, third, and fourth new classes
(B) as convenient symbols to take up the lapse before the admission of B
(C) as illustrations of the classes which work through four quarters a year
(D) to show how classes already formed fit into the new plan*
(E) to indicate the necessity for a summer recess

Organization of Ideas

Introduced: 11/61
Replaced: 11/66

Directions: Each set of questions in this section consists of a number of statements. Most of these statements refer to the same subject or idea. The statements can be classified as follows:

(A) the central idea to which most of the statements are related;
(B) main supporting ideas, which are general points directly related to the central idea;
(C) illustrative facts or detailed statements, which document the main supporting idea;
(D) statements irrelevant to the central idea.

The sentences do not make up one complete paragraph. They may be regarded as the components of a sentence-outline for a brief essay. The outline might, for example, have the following form:

A contains the central idea
B contains a main supporting idea
C presents an illustrative fact
B contains a main supporting idea
C presents an illustrative fact
C presents an illustrative fact

Classify each of the following sentences in accordance with the system described above.

(B) The Roman roads connected all parts of the empire with Rome.
(B) The Roman roads were so well built that some of them remain today.
(A) One of the greatest achievements of the Romans was their extensive and durable system of roads.
(D) Wealthy travelers in Roman times used horse-drawn coaches.
(C) Along Roman roads caravans would bring to Rome luxuries from Alexandria and the East.

Directed Memory—Reading Recall

Introduced: 11/61
Replaced: 10/77

Directions: In the test you will be given a period of time for the study of several extended prose passages. Then, without looking back at the passages, you will answer questions based on their contents. The following exercise is much shorter than those appearing on the test, but it illustrates the general nature of the passages and the questions. Remember, though, that on the test you will not be allowed to refer back to the passages.

SAMPLE PASSAGE:

Soon after the First World War began, public attention was concentrated on the spectacular activities of the submarine, and the question was raised more pointedly than ever whether or not the day of the battleship had ended. Naval men conceded the importance of the U-boat and recognized the need for defense against it, but they still placed their confidence in big guns and big ships. The German naval victory at Coronel, off Chile, and the British victories at the Falkland Islands and in the North Sea convinced the experts that fortune still favored superior guns (even though speed played an important part in these battles); and, as long as British dreadnoughts kept the German High Seas Fleet immobilized, the battleship remained in the eyes of naval men the key to naval power.

Public attention was focused on the submarine because

(A) it had immobilized the German High Seas Fleet
(B) it had played a major role in the British victories at the Falkland Islands and in the North Sea
(C) it had taken the place of the battleship
(D) of its spectacular activities*
(E) of its superior speed

Data Sufficiency

Introduced. 11/61
(currently in use)

Directions: Each of the questions below is followed by two statements labeled (1) and (2), in which certain data are given. In these questions you do not actually have to compute an answer, but rather you have to decide whether the data given in the statements are sufficient for answering the question. Using the data given in the statements plus your knowledge of mathematics and everyday facts (such as the number of days in July), you are to blacken space

A If statement (1) ALONE is sufficient, but statement (2) alone is not sufficient to answer the question asked;
B If statement (2) ALONE is sufficient, but statement (1) alone is not sufficient to answer the question asked;
C If BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient;
D If EACH statement ALONE is sufficient to answer the question asked;
E If statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

(E) In a four-volume work, what is the weight of the third volume?

(1) The four-volume work weighs 5 pounds.
(2) The first three volumes together weigh 6 pounds.

Practical Business Judgment:

Introduced: 11/72
(currently in use)

Directions: The passage in this section is followed by two sets of questions, data evaluation and data application. In the first set, data evaluation, you will be required to classify certain of the facts presented in the passage on the basis of their importance, as illustrated in the following example.

(This sample passage is much shorter than passages appearing in the test, but it is representative of data evaluation material.)

SAMPLE PASSAGE

Fred North, a prospering hardware dealer in Hillidale, Connecticut, felt that he needed more store space to accommodate a new line of farm equipment and repair parts that he intended to carry. A number of New York City commuters had recently purchased tracts of land in the environs of Hillidale and there had taken up farming on a small scale. Mr. North, foreseeing a potential increase in farming in that area, wanted to expand his business to cater to this market. North felt that the most feasible and appealing recourse open to him would be to purchase the adjoining store owned by Mike Johnson, who used the premises for his small grocery store. Johnson's business had been on the decline for over a year since the advent of a large supermarket in the town. North felt that Johnson

would be willing to sell the property at reasonable terms, and this was important since North, after the purchase of the new merchandise, would have little capital available to invest in the expansion of his store.

Consider each item separately in terms of the passage and choose

A if the item is a MAJOR OBJECTIVE in making the decision, that is, one of the outcomes or results sought by the decision-maker;

B if the item is a MAJOR FACTOR in making the decision, that is, a consideration, explicitly mentioned in the passage, that is basic in determining the decision;

C if the item is a MINOR FACTOR in making the decision, that is, a secondary consideration that affects the criteria tangentially, relating to a Major Factor rather than to an Objective;

D if the item is a MAJOR ASSUMPTION in making the decision, that is, a supposition or projection made by the decision-maker before weighing the variables;

E if the item is an UNIMPORTANT ISSUE in making the decision, that is, a factor that is insignificant or not immediately relevant to the situation.

SAMPLE DATA EVALUATION QUESTIONS

(D) Increase in farming in the Hilldale area
(A) Acquisition of property for expanding store
(B) Cost of Johnson's property
(C) State of Johnson's grocery business
(E) Quality of the farm equipment North intends to sell

A second set of questions, data application, requires judgments based on a comparison of the available alternatives in terms of the relevant criteria, in order to attain the objectives stated in the passage.

Each of the following questions relates to the passage. For each question, choose the best answer

SAMPLE DATA APPLICATION QUESTIONS

I Potential demand for farm equipment in the Hilldale area
II Desire to undermine Mike Johnson's business
III Higher profit margin on farm equipment than on hardware goods

(A) I only* (B) III only (C) I and II only
(D) II and III only (E) I, II, and III

Usage

Introduced: 10/76
(currently in use)

Directions: The following sentences contain problems in grammar, usage, diction (choice of words), and idiom Some sentences are correct. No sentence contains more than one error.

You will find that the error, if there is one, is underlined and lettered. Assume that all other elements of the sentence are correct and cannot be changed. In choosing answers, follow the requirements of standard written English

If there is an error, select the *one underlined part* that must be changed in order to make the sentence correct, and blacken the corresponding space on the answer sheet.

If there is no error, mark answer space E.

(C) He spoke bluntly and angrily to we spectators. No error
A B C D E

(A) He works every day so that he would become financially independent in his old age. No error
A B C D E

Reading Comprehension

Introduced: 10/77
(currently in use)

Directions: Each passage in this group is followed by questions based on its content. After reading a passage, choose the best answer to each question and blacken the corresponding space on the answer sheet. Answer all questions following a passage on the basis of what is stated or implied in that passage.

One sample reading passage follows. (It is much shorter than passages in the test, but it illustrates their general nature.)

SAMPLE PASSAGE

Not until the mid-1960's did any agriculturally based unions in the Southwest show promise of sustained operation. Mexican Americans were involved in efforts during the 1920's and 1930's to establish farm labor unions, but although these efforts resulted in dramatic and partially successful strikes, they were episodic and without organizational continuity.

The migratory work pattern compounded the problem of labor organization. A dispersed population in motion is not an easy target for organizational appeals. Industrial unionism had the advantage of mobilizing workers who were more concentrated in workplace and residence. It was easier to organize a work force whose members filed in and out of the workplace at fixed locations and at fixed times and who were exposed to daily contact with organizers. In addition, the multiple-employer structure of farm work, partially a function of labor mobility, made it less likely that union gains in one area could be transferred to another.

QUESTION ON READING PASSAGE

According to the passage, when did the first efforts of Mexican Americans to form agricultural unions take place?

(A) At the turn of the century
(B) During the 1920's*
(C) During the 1930's
(D) Immediately after the Second World War
(E) During the 1960's

APPENDIX B
Calculating Reliability Coefficients and Equating Parameters

Calculating Reliability Coefficeints

For the relatively unspeeded homogeneous subtests of the type that constitute GMAT, the reliability coefficient for each subtest can readily be calculated. All necessary data for these calculations are readily obtained when the test is given at a regular test administration. Once the reliability of each subtest has been determined, the reliability of GMAT Total, Verbal, and Quantitative scores can also be determined by well established methods.

The method used for calculating subtest reliabilities is based on the assumption that the intercorrelations of all items included in the subtest are substantially uniform, as would be expected to happen if all items are measuring the same basic ability. For example, if each item on a reading comprehension test were correlated with each other item, the whole set of intercorrelations would be expected to be essentially uniform. Under these conditions, convenient formulas for calculating the reliability coefficient can be derived.

The particular formula used in calculating reliability coefficients for GMAT subtests (called Kuder-Richardson Formula No. 20) has been adapted for use with formula-scored tests. To calculate the reliability of a GMAT part, the following items of information are needed:

(a) the number of items (called "n"),
(b) the number of right answers for each item (called "R"),
(c) the number of wrong answers for each item (called "W"),
(d) the fraction by which answers are multiplied before subtracting from the number of right answers in calculating formula scores (called "k"),
(e) the standard deviation of formula scores (called "σ_T"), and
(f) the number of examinees in the sample (called "N").

In describing the process of calculating a reliability coefficient, formulas for using the item analysis results may be considered first. In these formulas, the symbol "Σ" means that the results are to be summed over all items. The formulas are:

$$A = \frac{N\Sigma R - \Sigma R^2}{N^2} .$$

$$B = \frac{N\Sigma W - \Sigma W^2}{N^2} .$$

$$C = \frac{\Sigma RW}{N^2} .$$

It may be useful to express the first formula in words In calculating A, we begin by summing the number of right answers for all items and multiplying the results by the number of examinees. From that result, we subtract the sum of the squares of the number of right answers for all items. The resulting number is then divided by the square of the number of examinees.

The values calculated by these three formulas are then substituted in the following formula:

$$\text{Reliability} = \frac{N}{N-1}\left(1 - \frac{A + k^2B + 2kC}{\sigma_T^2}\right).$$

The following numerical sample illustrates the calculation of the reliability of the Reading Comprehension subtest. Because the test is composed of 25 five-choice items, n is equal to 25, and k is equal to .25. The basic data for the calculations are obtained from the computer printout, as follows:

$N = 2.080$
$\Sigma R = 31,456$
$SR^2 = 42,480.212$
$\Sigma W = 15,961$
$\Sigma W^2 = 12,427,015$
$\Sigma RW = 17,731,905$
$\sigma_T = 5.3847$

From these figures, we can compute values for A, B, C, as follows:

$$A = \frac{(2,080)\,(31,456) - 42,480,212}{(2,080)^2} = 5.3042 .$$

$$B = \frac{(2,080)\,(15,961) - 12,427\ 015}{(2,080)^2} = 4.8012$$

$$C = \frac{17,731,905}{(2,080)^2} = 4.0985 .$$

The reliability for the Reading Comprehension part may now be calculated, as follows:

$$\text{Reliability} = \frac{25}{24}\left[1 - \frac{5.3042 + (.25)^2(4.8012) + 2(.25)(4.0985)}{(5.3847)^2}\right].$$

The resulting value for the reliability of this part is .767. Similar calculations provide reliability coefficients for each of the other parts.

Once the reliability of each part has been calculated, it becomes possible to calculate the reliability of the total score. For this purpose, we need the standard error of measurement of each part. The standard error of measurement can be defined by the following formula:

$$\text{Standard Error of Measurement} = \sigma\sqrt{1 - \text{reliability}}.$$

It can be shown that the reliability of the total score equals:

$$1 - \frac{\text{Sum of squared standard errors of measurement}}{\text{Squared standard deviation of total score}}$$

Table B1 illustrates the calculation of the reliability of the Total score for a recent form of GMAT. A similar procedure is followed in calculating the reliability of the Verbal and Quantitative part scores.

Table B1

Calculation of Reliability Coefficient of GMAT Total Score

I. Basic Data

Part	Reliability of Part	Standard Deviation*	Standard Error of Measurement**	Squared Standard Error of Measurement
Reading Comprehension	.7667	5.3847	2.6008	6.7642
Problem Solving	.7489	4.6941	2.3520	5.5319
Practical Judgment	.6219	3.9185	2.4094	5.8052
Data Sufficiency	.8029	6.1773	2.7428	7.5230
Usage	.7626	5.5644	2.7111	7.3501
Practical Judgment	.5889	3.8463	2.4663	6.0826
Total of Squared Standard Errors of Measurement				39.0570

*The standard deviation of total scores is 21 8836

**These values were calculated using more decimal places than are shown in the table

II. Calculations

Reliability =

$$1 - \frac{\text{Sum of squared standard errors of estimate}}{\text{Squared standard deviation}} .$$

Reliability =

$$1 - \frac{39.0570}{(21.8836)^2} = 1 - \frac{39.0570}{478.8919} = .918.$$

Calculating Equating Parameters

The basic method of equating used for GMAT scores from 1962 to the present time makes use of the well known statistical principle that if a very large group is divided at random into two or more subgroups, the resulting subgroups will be quite similar in every characteristic. In applying this method we administer a form for which scaled scores are already known for one random subgroup and we administer the new form to a different random subgroup. Assuming that the two random subgroups are equal in the ability measured by GMAT, we attribute differences in scores between the two forms to a difference in the difficulty of the two tasks.

In order to relate raw scores on a new form to the GMAT scale, we need to know:

(a) the equation relating raw scores to scaled scores for the old form;

(b) the mean and standard deviation of raw scores for the random subgroup of examinees who took the old form; and

(c) the mean and standard deviation of raw scores for the random subgroup of examinees who took the new form.

The following data were available for 9,850 examinees who took the old form and 9,795 examinees who took the new form:

	Old Form	New Form
Mean	65.765279	63.586932
Stanard Deviation	22.795999	23.601557

Equating files show that the equation for converting raw scores to scaled scores for the old form has a multiplier (A_O) of 4.69 and an additive constant (B_O) of 167.0300.

To solve for the multiplier for the new form, we use the formula

$$A_N = A_O \left(\frac{\sigma_O}{\sigma_N}\right)$$

$$A_N = 4.6369 \left(\frac{22.795999}{23.601557}\right)$$

$$A_N = 4.478635 .$$

This value agrees with the computer determined value of 4.4786.

To solve for the additive constant for the new form, we use the formula[3]

$$B_N = A_O M_O - A_N M_N + B_O .$$

so that

$$\begin{aligned} B_N &= (4.6369)(65.765279) - (4.478635)(63.586932) + 167.0300 \\ &= 304.9470 - 284.7827 + 167.0300 \\ &= 187.1943. \end{aligned}$$

As it turned out, the value obtained when B_N was calculated by the computer was also 187.1943, although the sample calculation did not reproduce in detail the calculations performed by the computer.