DOCUMENT RESUME

ED 383 716 TM 023 137

AUTHOR Cook, Linda L.; And Others
TITLE Aligning Score Scales for Achievement Tests in Multiple Content Areas.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-30
PUB DATE Dec 90
NOTE 132p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Achievement Tests; Course Content; High Schools; *High School Students; Intellectual Disciplines; Reference Groups; Sampling; *Scaling; *Scores
IDENTIFIERS *College Board Achievement Tests

ABSTRACT

As a result of a recent College Board Admissions Testing Program Achievement Test scaling study, L. L. Cook and others recommended that the practice of sampling only high school juniors taking the achievement tests in June might be expanded to include sophomores and that a two-stage scaling procedure be evaluated. The two-stage procedure would include a first stage involving scaling tests on a within-cluster basis, but the second stage would involve taking the results of the first stage and following it with a second scaling in which scaled scores from the first stage would be used as input. This study experimentally evaluated these scaling recommendations, separating the Achievement Tests into a language test cluster and a nonlanguage test cluster. In addition, high school sophomores were added to the samples. Results indicate that addition of the sophomores does not improve the relationship between Achievement Test scores and the scaling covariates. There is evidence to suggest that the use of empirical values for the reference group and the use of a two-stage scaling procedure may improve the alignment of the Achievement Test scales. Seven tables and three figures present analysis results. (Contains nine references.) (SLD)

RESEARCH REPORT

Aligning Score Scales for Achievement Tests in Multiple Content Areas

Linda L. Cook
Daniel R. Eignor
Elizabeth B. Burton

Educational Testing Service
Princeton, New Jersey
December 1990

Aligning Score Scales for Achievement Tests in Multiple Content Areas[1,2]

Linda L. Cook[3]
Daniel R. Eignor
Elizabeth B. Burton

Educational Testing Service

August 1990

[1]A previous version of this paper was presented at the annual meeting of the American Educational Research Association, Boston, 1990.

[2]This study was supported by the College Board through Joint Staff Research and Development Committee funding.

[3]The authors would like to thank Karen Carroll for her programming contributions, Nancy Petersen and Larry Wightman for review comments, and Georgiana Thurston for typing the paper.

EXECUTIVE SUMMARY

Cook, with the collaboration of Angoff and Schmitt, recently carried out a College Board Admissions Testing Program (ATP) Achievement Test scaling study (see Cook, 1988). A goal of the Cook study was to provide several alternative scaling models for the Achievement Tests which would be based on the empirical evidence gathered in the study. It was intended that, in these models, components such as scaling covariates, scaling samples, and characteristics of the reference group would be varied. In addition, it was anticipated that not all Achievement tests studied would be amenable to similar treatment; and most likely the tests would be clustered by content area and alternative models would be specified for each cluster.

As a result of the analyses carried out by Cook, suggestions were made for constructing scaling samples as well as the reference population. For instance, it was suggested that the practice of sampling only high school juniors taking the Achievement Tests in June for scaling purposes might possibly be altered to also include high school sophomores. It was also suggested that a two stage scaling procedure be evaluated. The two stage scaling procedure would include a first stage which would involve scaling tests solely on a within-cluster basis; i.e., different covariates, different reference populations, and possibly even different sampling procedures would be used for each cluster. The second stage of the two-stage scaling procedure would involve taking the results of the first-stage scaling procedure just described and following it with a second scaling in which the scaled scores obtained from the first-stage scalings would be used as input.

The purpose of this study was to experimentally evaluate the scaling recommendations provided by Cook (1988). The Achievement Tests were separated into two clusters: a language test cluster and a non-language test cluster.

Tests within the non-language test cluster were then subjected to a first-stage scaling, using scores on SAT-V and SAT-M as covariates, where the means and standard deviations of the reference population for SAT-V and SAT-M were empirically determined using combined data from all the samples taking each of the non-language tests. For the language tests, a similar procedure, but making use of semesters of study in addition to SAT-V and SAT-M scores, was implemented; again, reference group means and standard deviations for SAT-V, SAT-M, and semesters of study were empirically determined from the combined data from all samples taking each of the language tests.

The language tests were also subjected to an additional first-stage scaling procedure. This procedure can be thought of as an alternative to the procedure specified above for the first-stage scaling of the language cluster. The procedure consisted of scaling all of the language tests to the French Test, using SAT-V and -M scaled scores and semesters of study as covariates.

Scaled scores resulting from the one first-stage scaling applied to the non-language tests and the two first-stage scalings of the language tests were then subjected to second-stage scalings, using scores on SAT-V and SAT-M as covariates, but using empirically derived reference group values based on data from the combined sample derived from the individual samples for each language and non-language Achievement Test included in the study.

The results of the experimental first- and second-stage scalings were then compared to the results of two single-stage scaling procedures. One procedure used empirically determined estimates for the reference group values based on the combined sample of all examinees from each of the language and non-language test samples, i.e., the same reference group used for the second stage scaling described above. The second procedure used 500 and 100 as

reference population values for SAT-V and SAT-M. Hence, all that differed between the two single-stage scalings were the reference group values used in the scaling equations.

Finally, in an effort to assess the effects of including high school sophomores in the scaling samples from the June administration, sophomores were included in the scaling samples for Achievement Tests in Biology and Chemistry for the first- and second-stage experimental scalings that were carried out.

The results of the study related to the sampling question indicate that addition of high school sophomores to the samples does not improve the relationship between Achievement Test scores and the scaling covariates (at least as evaluated by the correlations between Achievement Test score and the covariates) and thus is probably not an appropriate change to consider. The results of the study related to the investigation of the two stage scaling procedure indicate there is some reasonable evidence to suggest that the use of empirical values for the reference group and use of the procedure that involves scaling the test in two stages may improve the alignment of the Achievement Test scales. A viable alternative involves application of the single-stage scaling procedure based on empirically determined estimates for the reference group values to the non-language tests and the two stage scaling to the language tests. This combination of procedures appears to provide a comparable degree of alignment of the scales as that provided by applying the two stage procedure to all tests, and should be somewhat easier to implement.

The results of this study should be considered tentative and further research should be carried out. Procedures that involve modeling and correcting for the selection bias present in the Achievement Test scores might

prove fruitful. In addition, use of non-linear scaling procedures that should provide improved alignment for scores throughout the entire scaled score range for the tests might also be desirable and should be further investigated.

Aligning Score Scales for Achievement Tests in Multiple Content Areas

Linda L. Cook
Daniel R. Eignor
Elizabeth B. Burton

INTRODUCTION

The College Board Admissions Testing Program (ATP) offers two varieties of tests: the Scholastic Aptitude Test (SAT) and the Achievement Tests. The SAT is a test of general verbal and mathematical developed ability that all examinees testing through ATP usually take. The Achievement Tests, on the other hand, are a battery of fourteen subject matter tests (fourteen tests when this study was done; a fifteenth, Modern Italian, was added in June 1990). Examinees testing at a particular date may take either one, two, or three of the fourteen tests. Moreover, the examinee is allowed to choose which of the tests he/she wants to take at the particular administration. Hence, the group of examinees taking any one of the Achievement Tests is a self-selected group, different from the self-selected group that may have chosen to take one of the other tests. Usually, however, score users wish to compare the scores of the groups of examinees who take the different Achievement Tests and, hence, some method of scaling, that aligns the scales of the various tests so that scores on the tests are on reasonably comparable scales, is necessary.

The desired outcome of any procedure used to align the scales of the Achievement Tests is fairly evident. According to Angoff (1968):

> "The purpose of this scaling is to ensure that a candidate who chooses to compete with more able candidates is not put at a disadvantage; that is, that a candidate who is average in a highly selected group of candidates will earn a higher scaled score than a candidate who is average in a less able group."

Procedures which may be used to scale the Achievement Tests to achieve comparability across tests are discussed in the next section of the paper.

These procedures form an extension of a subset of an overall set of procedures known as moderation procedures. Keeves (1988) has provided an excellent overview of moderation procedures and Cooney (1975, 1976) has described certain of these procedures in detail, along with their use in moderating examination marks in Australia. According to Keeves (1988):

> "Moderation is a procedure that was first employed at Oxford University to compare and equate levels of performance in the examinations conducted within the colleges of the university. The statistical procedures that have been developed to serve the purposes of equating levels of achievement on different examination papers have also come to be known as "moderation". . . The function of moderation in this situation is to establish and maintain comparable standards between different examinations in the same subject area that are conducted on different syllabuses or in different settings. A further use of moderation occurs when a total score must be calculated from examination marks in different subject areas."

Keeves (1988) goes on to point out:

> "Howard (alias used by Sir Cyril Burt) (1958) identified the key requirements of moderation procedures. They are that candidates should not be disadvantaged by the marking pattern of examiners nor by the candidatures [examinee cohorts] with whom they compete. In practice these requirements demand that the same mark on different examinations should imply the same level of performance relative to a common population."

As can be seen from the above quotes, the demands being made of moderation procedures in Australia are somewhat greater than those that are made with respect to ATP Achievement Tests. The expectation in Australia is that moderation will account for: 1) differences between subjects in the quality of students attracted; 2) differences between schools in the characteristics of students attending them; 3) differences between graders, both between and within schools, in the distributions of scores given; and 4) differences between courses of instruction studied by students. With the ATP

Achievement Tests, scores are objectively determined by scanning multiple-choice answer sheets and it is assumed that students taking the exams have had suitable preparation. Further, with our public education system, it is assumed, on average, that little self-selection of schools takes place. Moderation is necessary with the ATP Achievement Tests simply to account for differences between subjects in the quality of students attracted. Moderation is frequently used in Australia even when there are no differences between subjects in the quality of students attracted, i.e., when all subjects are compulsory. The procedures are then used, for instance, to account for differences between graders in subjectively assigned marks given to students.

Keeves (1988) discusses two general classes of moderation procedures that have been used in Australia: 1) those procedures for moderation that involve attributes of a common stimulus task to which groups of students are required to respond, and 2) those procedures used for moderation that are concerned with the attributes of the groups of students with respect to a larger population of students. The first set of moderation procedures typically involve the use of a moderator test taken by all students. According to Cooney (1976):

> "With such a measure, scaling may then be achieved either by a modification of Pearson's bivariate adjustment method which leads to a linear transformation[1], or by the

---

[1]Keeves (1988) has presented simplified equations for bivariate adjustment, given in complete form by Cooney (1975), that make the moderation process easier to understand. If the joint distribution of the moderator test scores and the achievement test scores is bivariate normal and the marginal distributions are normal and if the moderator test has a significant correlation r with the achievement test, then the moderated score for student j on achievement test i (T_{ij}) may be expressed as

$$T_{ij} = M_{Yi} + r \frac{S_{Yi}}{S_x} (X_j - M_x)$$

where M_{Yi} is the mean score on achievement test i, X_j is the score of student j on the moderator test, M_x is the mean score on the moderator test, S_{Yi} is the standard deviation of the achievement measure Y_i, and S_x is the standard deviation of the moderator test X.

> equipercentile method. . . Both methods depend heavily on the use of a moderator variable which correlates uniformly and highly with the marks being scaled."

Keeves (1988) points out that the traditional procedure used to scale the ATP Achievement Tests, discussed in the next section of this paper, is an extension of the moderator variable method that involves two moderator tests (SAT-V and SAT-M) and for the language tests, a third moderator variable, semesters of study. Cooney (1975) points out that the bivariate adjustment procedure is seen as adequate only if there exists a moderator variable or set of variables which measure the attributes of performance in the courses of study and correlate highly and uniformly with the scores on the various subjects.

There are two moderation procedures that make use of characteristics of examinees in use in Australia. The first usually involves the situation where there is considerable overlap in the groups of students taking different subjects. According to Keeves (1988):

> "The most obvious achievement characteristic that is available for adjusting the level of performance of a student group to allow for differences in the quality of the candidatures is the performance of the students on the other subjects that they sat on the same occasion."

The second moderation procedure employs the characteristics of a student group given by the average performance of the group on a general ability test. According to Keeves (1988):

> "Superficially this procedure is similar to that associated with the use of a moderator test, and the similarity arises from the fact that for all examination subjects the correlation between the subject and the moderator variable is set at unity, to account for the differences in the magnitudes of the correlations which were seen to pose particular problems for that method. . . In practice this procedure. . . establishes a scholastic aptitude test scale on which the qualities of the candidatures of the different school and subject groups are measured."

To summarize, the scaling procedures discussed in the next section of this paper form an extension of a subset of the total set of procedures called moderation procedures that have been in use for some time in Australia--more particularly, these procedures are an extension of the subset of procedures that make use of a moderator test. The key to how effective moderation procedures perform in this context is related to how the moderator variables correlate with the scores to be scaled--the moderator variables should correlate highly and uniformly with the scores on the various subjects (Cooney, 1975).

BACKGROUND

When the College Board Achievement Tests were introduced for the first time in 1942 for operational use in admissions, the tests were initially scaled in such a way that the mean for the group choosing to take each test was set at 500 and the standard deviation at 100. That is to say, the average of each group of candidates taking its test was made to appear equal to the average performance of each of the other groups of candidates taking their tests. Similarly with their standard deviations. As a consequence of this scaling design, the score a candidate received on a test was clearly dependent on, among other things, the ability level of the group of examinees who took the particular test. For example, a candidate would appear more able if he or she took a test typically chosen by a group of less able examinees and would appear less able if he or she chose a test typically taken by high ability examinees. Any candidate who understood the design of the score scales, and wished to appear to be relatively knowledgeable in his/her field, could adopt the strategy of selecting the Achievement Test normally taken by the least able candidates. In order to remove this element of unfairness, a scaling

system was designed in the middle 1940's to adjust the scales for the several Achievement Tests to reflect the level and dispersion of ability of the candidates taking each test. With this system, a test typically taken by a more able group of candidates was made to yield an average scaled score higher than 500, and a test typically taken by a less able group of candidates was made to yield an average score lower than 500.

The operational scale definition initially adopted in the middle 1940's to achieve this result was that the candidate of average ability, relative to a hypothetical aggregate of all candidates taking the College Board tests, would, in theory earn a score of 500 regardless of the Achievement Test that he or she chose to take; also, that the dispersion of scaled scores for this hypothetical population would be defined with a standard deviation of 100 (see Wilks, 1961). Thus, higher ability Achievement Test groups would automatically have higher means and, correspondingly, lower ability Achievement Test groups would have lower means. This definition was implemented by defining "general ability" as measured by the verbal and mathematical Scholastic Aptitude Tests (SAT-V and SAT-M), respectively; and the degree to which the SAT-V and -M scores played a role in this operational definition was a direct function of the relevance of those tests for the particular Achievement Test in question, as measured by the correlation of the SAT scores with the scores on the particular Achievement Test. A further adjustment was later introduced into this system by adding semesters of study to the SAT-V and -M scores for scaling the foreign language tests. This adjustment was intended to account for the fact that some languages were typically studied for longer periods of time than were other languages.

In 1959, Professor Samuel Wilks of Princeton University was engaged to review the work of equating and scaling the SAT and the Achievement Tests (see Wilks, 1961). The scope of his review included not only an examination of the particular methods in use at the time, but also an examination of the system, its philosophy, and its mode of implementation. One of the questions under consideration in his review was that of the relative emphasis to be placed on the efforts to perpetuate the scales of the individual Achievement Tests by providing undisturbed form-to-form equivalence through the equating process versus the emphasis to be placed on the efforts to maintain the appropriate, up-to-date inter-relationships among the scales for the tests, through the scaling process. Wilks recommended that scaling should be the first order of business. Accordingly, a plan was instituted to rescale the tests each year, and to average the results of the rescaling with equating results for that year. This plan was implemented in 1964 and applied annually until 1972, with the results incorporated into the scores reported for the following year's cohort. The expectation was that the differences between the scaling and equating efforts in this time would have diminished to the vanishing point. A review of the rescaling efforts in 1972 revealed, however, that the scaling operation was not moving consistently in one direction, but fluctuated from one rescaling to another, sometimes by sizable amounts. Hence, from 1972 to 1978, rescaling was carried out bienially. In 1979 and 1980, the procedure was carried out annually. Rescaling was discontinued in 1980 because it was thought that the available methodology probably was not providing optimal results. Since 1980, only form-to-form equating has taken place in the program and the current scale for the Achievement Tests is the scale defined

in 1980 when rescaling was discontinued. However, since that time, a number of studies of the Achievement Test scaling process have been undertaken.

In the early 1980's, H. Braun and L. R. Tucker conducted studies (see Dorans, 1985) designed to investigate Achievement Test scaling that used both real and simulated data. These studies were undertaken to gain a better understanding of how operational decisions affected the outcomes of scaling. The effects of changing the definitions of the hypothetical reference group and of changing the definition of the samples of Achievement Test takers used for calculating the scaling equations, as well as the effects of various choices of covariates, were given particular attention. The results of the Braun and Tucker studies indicated that choice of the reference population impacts scaling results as does the relationship of the Achievement Test score with the scaling covariates.

Cook, with the collaboration of Angoff and Schmitt, (Cook, 1988) recently carried out an Achievement Test scaling study. The purpose of the Cook study was to explore the relationships between College Board Achievement Test scores and potential scaling covariates for various subgroups of the test taking population. It was speculated that such an exploration would lead to the following:

- The selection of additional scaling covariates that might provide improved scaling results for those tests that do not provide scores correlating highly with SAT-V and/or SAT-M scores;
- An improved specification of the characteristics of the sample of Achievement Test examinees that are used for the scaling of the tests, i.e., such a specification might possibly lead to Achievement Test scores that show a higher correlation with selected scaling covariates and;
- An improved specification of the reference group (population) used for the scaling. As Braun and Tucker pointed out, the characteristics of the hypothetical population differentially affect the scales of the tests. A change in specifications for this

population, from the traditionally specified SAT-V and SAT-M means and standard deviations of 500 and 100, respectively, might possibly provide improved scales for some of the tests.

The final goal of the Cook study was to provide several alternative scaling models for the Achievement Tests which would be based on the empirical evidence gathered in the study. It was intended that, in these models, components such as scaling covariates, scaling samples, and characteristics of the reference group would be varied. In addition, it was anticipated that not all Achievement tests studied would be amenable to similar treatment; and most likely the tests would be clustered by content area and alternative models would be specified for each cluster.

As a result of the analyses carried out by Cook, suggestions were made for constructing scaling samples as well as the reference population. It was also suggested that a two stage scaling procedure be evaluated. The two stage scaling procedure would include a first stage which would involve scaling tests solely on a within-cluster basis; i.e., different covariates, different reference populations, and possibly even different sampling procedures would be used for each cluster.

The second stage of the two-stage scaling procedure would involve taking the results of the first-stage scaling procedure just described and following it with a second scaling in which the scaled scores obtained from the first-stage scalings would be used as input. It should be mentioned that Cook did not identify, as a result of her analyses, additional covariates that could be used in either the first or second stages of the two stage scaling procedure that was suggested.

PURPOSE

The purpose of this study was to experimentally evaluate the scaling recommendations provided by Cook (1988). The Achievement Tests were separated into two clusters: a language test cluster and a non-language test cluster. Tests within the non-language test cluster were then subjected to a first-stage scaling, using scores on SAT-V and SAT-M as covariates, where the means and standard deviations of the reference population for SAT-V and SAT-M were empirically determined using combined data from all the samples taking each of the non-language tests. For the language tests, a similar procedure, but making use of semesters of study in addition to SAT-V and SAT-M scores, was implemented; again, reference group means and standard deviations for SAT-V, SAT-M, and semesters of study were empirically determined from the combined data from all samples taking each of the language tests.

The language tests were also subjected to an additional first-stage scaling procedure. This procedure can be thought of as an alternative to the procedure specified above for the first-stage scaling of the language cluster. The procedure consisted of scaling all of the language tests to the French Test, using SAT-V and -M scaled scores and semesters of study as covariates.

The French Test was chosen as the base test not only because it has been, until recently, the largest volume language Achievement Test, but also because of certain properties of the French Test scale, i.e., for a number of French Test forms, the maximum raw score produces a scaled score of 800; this is true to a lesser extent for the other language Achievement Tests. In addition, scores on the French Test correlate more highly with SAT-V and -M scores than do scores on the other language tests except for Latin.

Scaled scores resulting from the one first-stage scaling applied to the non-language tests and the two first-stage scalings of the language tests were then subjected to second-stage scalings, using scores on SAT-V and SAT-M as covariates, but using empirically derived reference group values based on data from the combined sample derived from the individual samples for each language and non-language Achievement Test included in the study.

The results of the experimental first- and second-stage scalings were then compared to the results of two single-stage scaling procedures. One procedure used empirically determined estimates for the reference group values based on the combined sample of all examinees from each of the language and non-language test samples, i.e., the same reference group used for the second stage scaling described above. The second procedure used 500 and 100 as reference population values for SAT-V and SAT-M. Hence, essentially all that differed between the two single-stage scalings were the reference group values used in the scaling equations.

METHODOLOGY

Description of the Tests

The thirteen[1] Achievement Tests used in the study fall into five general subject areas:

English

- English Composition (two versions: all multiple-choice and multiple choice with essay)
- Literature

[1]The Achievement Test in Hebrew was excluded from this study because, at the time of the study, the test was undergoing redevelopment to make it more relevant to the current Hebrew test-taking population.

Foreign Languages

French
German
Latin
Spanish

History and Social Studies

American History and Social Studies
European History and World Cultures

Mathematics

Mathematics Level I
Mathematics Level II

Sciences

Biology
Chemistry
Physics

All the Achievement Tests take one hour of testing time, and consist entirely of multiple-choice questions except the English Composition Test with Essay, which consists of a 20 minute essay and 40 minutes of multiple-choice questions. The tests vary in content as well as in the number of multiple-choice items they contain. The approximate number of questions contained in each test is listed in the table below.

Test	Approximate Number of Questions
English Composition with Essay	70 multiple-choice items plus one essay
English Composition	85
Literature	60
French	85
German	80
Latin	70
Spanish	85
American History and Social Studies	95
European History and World Cultures	95
Mathematics Level I	50
Mathematics Level II	50
Biology	95
Chemistry	85
Physics	75

Scores for all Achievement Tests are reported on scales that range from 200 to 800.

The thirteen Achievement Tests were split into two clusters for certain

two clusters are:

Language Test Cluster	Non-language Test Cluster
French	English[1]
Spanish	Literature
German	American History and Social Studies
Latin	European History and World Cultures
	Mathematics Level I
	Mathematics Level II
	Biology
	Chemistry
	Physics

Description of the Samples

The Achievement Test data used to provide a base for the scalings in this study were obtained from the following Achievement Test administrations:

May 1987
June 1987[2]
November 1987
December 1987[2]
January 1988

For each of the administrations, examinees were selected as follows:

May 1987:	All juniors taking an Achievement Test (or Tests)
June 1987:	For all tests except Biology and Chemistry--all juniors taking the Achievement Test (or Tests) For Biology and Chemistry--all juniors and all sophomores taking the Achievement Test (or Tests)[3]

[1]The English Composition with Essay and the all-objective English Composition tests are placed on the same score scale via the score equating process. Scaled scores for both tests were used interchangeably in this study.

[2]The small volume tests, European History and World Cultures, German, and Latin are offered only at the December and June administrations.

[3]Sophomores were included in the scaling samples for Biology and Chemistry only for the experimental scalings carried out in this study in an attempt to produce improved scaling results. For the two single-stage scalings, sophomores were not included in the scaling samples.

November 1987: All seniors taking the Achievement Test (or Tests)
December 1987: All seniors taking the Achievement Test (or Tests)
January 1988: All seniors taking the Achievement Test (or Tests)

Examinees needed to have SAT-V and SAT-M scores from at least one of the following seven administrations to be included in the samples.

April 1987	November 1987
May 1987	December 1987
June 1987	January 1988
October 1987	

It should be noted that examinees cannot, at present, take the SAT and the Achievement Tests at the same administration date.

Examinees taking the French, German, Spanish, and Latin tests were included in the sample only if they responded to the question on the background questionnaire having to do with semesters of study. Table 1 provides the background questionnaire which was in use with the French Test at the time data were collected for this study. The same questionnaire, with the appropriate name change, is used with the German and Spanish Tests while a very similar questionnaire is used with Latin. To be included in the sample for the French, German, and Spanish Tests, examinees had to have marked one of responses 3-8 to the background question. For the Latin Test, examinees had to have marked one of responses 2-8 to the background question for that test. For use in the scaling equations, French, German, and Spanish background responses 3, 4, 5, 6, 7, and 8 were recoded as 4, 5, 6, 7, 8, and 9 semesters of study, respectively. For Latin, background responses 2, 3, 4, 5, 6, 7, and 8 were recoded as 3, 4, 5, 6, 7, 8, and 9 semesters of study, respectively.

Insert Table 1 about here

Scalings

All scalings in this study were carried out using Tucker equations for two anchor and three anchor scalings described in detail by Angoff (Angoff, 1984, pp. 108-111 and pp. 132-133). These scalings differed, however, in how reference group means, variances, and covariances of SAT scores and semesters of study were established and in the number of stages used in the scalings--single or two-stage scalings.

The Tucker two-anchor scaling equations for estimating scaled score means ($\hat{M}$) and scaled score standard deviations ($\hat{S}$) are:

$$\hat{M}_x = M_x + b_{vx}(\mu_v - M_v) + b_{mx}(\mu_m - M_m) \quad (1)$$

$$\hat{S}^2_x = S^2_x + b^2_{vx}(\sigma^2_v - S^2_v) + b^2_{mx}(\sigma^2_m - S^2_m) + 2b_{vx}b_{mx}(\sigma_{vm} - C_{vm}) \; ; \quad (2)$$

and the Tucker three-anchor scaling equations are:

$$\hat{M}_x = M_x + b_{vx}(\mu_v - M_v) + b_{mx}(\mu_m - M_m) + b_{sx}(\mu_s - M_s) \quad (3)$$

$$\hat{S}^2_x = S^2_x + b^2_{vx}(\sigma^2_v - S^2_v) + b^2_{mx}(\sigma^2_m - S^2_m) + b^2_{sx}(\sigma^2_s - S^2_s) + 2b_{vx}b_{mx}(\sigma_{vm} - C_{vm}) + 2b_{vx}b_{sx}(\sigma_{vs} - C_{vs}) + 2b_{mx}b_{sx}(\sigma_{ms} - C_{ms}) \; ; \quad (4)$$

where M, S^2, C, and b represent the observed mean, variance, covariance, and partial regression coefficient, respectively, of the subscripted variables; μ, σ^2, and σ represent the reference group mean, variance, and covariance, respectively, of the subscripted variables; and x, v, m, and s represent Achievement Test scaled scores, SAT-V scaled scores, SAT-M scaled scores, and semesters of study, respectively.

Once estimates of Achievement Test means and standard deviations have been obtained using the two- or three-anchor scaling equations, these estimates are used to obtain linear scaling parameters as follows:

$$(X - \hat{M}_x)/\hat{S}_x = (T - 500)/100,$$

which yields a linear equation of the form $T = AX + B$, where

$$A = 100/\hat{S}_x \text{ and } B = 500\text{-}A(\hat{M}_x). \quad (5)$$

Single-Stage Scalings

Equations (1), (2), and (5) were used for the single-stage scalings of the nine non-language Achievement Tests, with the reference group values for μ_v, μ_m, σ_v, σ_m, and σ_{vm} set equal to 500, 500, 100, 100, and 6,000, respectively.

Equations (3), (4), and (5) were used for the single-stage scalings of the four language Achievement Tests. The reference group values for μ_v, μ_m, σ_v, σ_m, and σ_{vm} were again set equal to 500, 500, 100, 100, and 6,000, respectively. The reference group values for μ_s, σ_s, σ_{vs}, and σ_{ms} were set equal to the corresponding observed values in the combined sample of all language test-takers formed by pooling the four language samples.

Empirical Single-Stage Scalings

The empirical single-stage scalings of the thirteen Achievement Tests mirrored the single-stage scalings except that the reference group values for μ_v, μ_m, σ_v, σ_m, and σ_{vm} for each test were set equal to the corresponding observed values in the combined sample of all test-takers, formed by pooling all thirteen samples. As with the single-stage scaling for the language tests, μ_s, σ_s, σ_{vs}, and σ_{ms} were set equal to the corresponding values in the combined sample of all language test takers formed by pooling the four language samples.

First-Stage Scalings

The first-stage scalings of each of the thirteen Achievement Tests mirrored the empirical single-stage scalings except that the reference group

values for μ_v, μ_m, σ_v, σ_m, and σ_{vm} for the nine non-language tests were set equal to the corresponding values in the combined sample of all non-language test-takers (formed by pooling the nine non-language samples), and the reference group values for the four language tests were set equal to the corresponding values in the combined sample of all language test-takers.

First-Stage Scalings of Language Tests to French Test Scale

Equations (3) and (4) were used to perform the first-stage scaling of the German, Latin, and Spanish Tests to the French Test scale. For each of these three tests, the reference group values for μ_v, μ_m, μ_s, σ_v, σ_m, σ_s, σ_{vm}, σ_{vs}, and σ_{ms} were set equal to the corresponding values in the combined sample formed by pooling the specific language sample (either German, Latin, or Spanish) with the French sample. Linear scaling parameters for each of the tests were then derived as follows:

$$A = S_f/\hat{S}_x \text{ and } B = M_f - A(\hat{M}_x), \tag{6}$$

where M_f and S_f represent the scaled score mean and standard deviation, respectively, for the French Test sample. It should be noted that French Test scores were not rescaled in applying this procedure.

Second-Stage Scalings

All second-stage scalings for each of the thirteen Achievement Tests were carried out using Equations (1), (2), and (5) applied to the scaled scores for each test derived from the first-stage scaling. Consequently, the second-stage scaling was done twice for the language tests, once using the first-stage scaling results based on the empirically derived reference group (to be referred to as Second-Stage Scaling A) and once using the first-stage scaling to French results (to be referred to as Second-Stage Scaling B). In all

second-stage scalings, the reference group values for μ_v, μ_m, σ_v, σ_m, and σ_{vm} were set equal to the corresponding observed values in the combined sample of all test-takers, formed by pooling all thirteen Achievement Test samples.

Data Used in Scalings

Tables 2 and 3 summarize the data used as input for the various scalings performed. Table 2 contains scaled score summary statistics for the reference groups or populations and Table 3 contains scaled score summary statistics used for the Achievement Test scaling samples.

Insert Tables 2 and 3 about here

As seen in Table 2 and explained previously, the reference group values for SAT-V and SAT-M for the second stage scaling and empirically based single-stage scaling are identical. It is also apparent, from examination of the reference group values for SAT-V and SAT-M given in Table 2 for the first stage scaling of the non-language tests, that these values are very similar to those used for the second stage scaling and the empirically based single stage scaling. This is due to the fact that the non-language tests dominate the aggregate used to obtain the reference group values for these two scalings. The reason this occurs is: 1) the non-language tests outnumber the language tests (there are nine non-language tests compared to the four foreign language tests); and, 2) in general, the non-language tests are the larger volume tests and, therefore, they have a greater impact on the aggregate statistics than do the language tests. It should also be noted that the language and non-language test groups are reasonably similar in ability, as measured by SAT-V and SAT-M scores.

Summary of Scalings Performed

As a result of application of the scaling procedures, linear scaling parameters and "new" scaled scores for each of the thirteen Achievement Tests were created as follows:

1. Single-stage scaling -- a single set of linear scaling parameters for each of the thirteen tests.
2. Empirical single-stage scaling -- a single set of linear scaling parameters for each of the thirteen tests.
3. First-stage scaling -- A single set of first-stage scaling parameters for each of the non-language tests and two sets of first-stage results for each of the language tests, one set based on using an empirically derived reference group and the other set based on scaling the Spanish, German, and Latin Tests to the French Test scale.
4. Second-stage scaling -- A single set of second-stage scaling parameters for each of the non-language tests and two sets of second-stage results for each of the language tests.

Evaluation of Results

The results of the analyses were evaluated by comparing Achievement Test scaled scores obtained from the application of the experimental scaling procedures. The results were compared in several ways. First, the results were compared to determine if the rank orderings of the Achievement Test scaled score means were similar to what was to be expected, given the ability levels (as measured by SAT-V and SAT-M scaled scores) of the groups taking the tests.

Second, the results of the experimental scalings were evaluated by examining the Achievement Test scaled score means obtained from application of the experimental scaling parameters after conditioning on SAT-V and SAT-M scaled scores and semesters of study. The assumption underlying the conditioning procedure is that if the scales of the Achievement Tests are aligned, scaled score means on the tests will be somewhat similar for groups at the same ability level, as measured by the scaling covariates.

Results of the study were also evaluated by examining the relationship between Achievement Test scaled score means for pairs of Achievement Tests where each pair was taken by some reasonably sized group of examinees. Again, the assumption was that if the test scales were aligned, the scaled score means obtained for each pair of tests taken by each group of examinees would be reasonably similar.

RESULTS

The results of the analyses conducted for this study are summarized in Tables 4-7 and Figures 1a-3d. The information provided in Table 4 summarizes the results of applying the linear parameters obtained from the four experimental scaling procedures used for the non-language tests and the six experimental procedures used for the language tests to values at fifty point intervals on the current Achievement Test scales, which range from 200 to 800. The scale values in the left column are labeled current scale and indicate the existing scale values for each of the tests prior to application of any of the experimental scaling results.

Insert Table 4 about here

Examination of the data presented for the English Composition Test in Table 4 indicates that the results of the first and second stage scalings and the empirically based single-stage scaling are almost identical. All three of these scaling procedures result in scaled scores that are somewhat lower than scaled scores on the current scale.

Only the single-stage scaling results shown in Table 4 for the English Composition Test provide scaled scores that differ from those provided by the other three experimental procedures. The major difference between this and the other procedures is the way in which the reference population is defined. For the single-stage scaling procedure, the reference population is defined to have a scaled score mean of SAT-V and SAT-M scores of 500 and a standard deviation of scaled scores of 100 for both tests. For the other three procedures, empirical values obtained by aggregating across samples taking the actual tests were used.

The results shown in Table 4 for the Literature Test, American History Test, and European History Test are quite similar to those obtained for the English Composition Test. For all of these tests, the results of the first and second stage scalings and the empirically based single-stage scaling are quite similar. For all tests, the results of these three procedures provide scaled scores that are similar and somewhat lower than those provided by the single-stage scaling procedure. As was the case for the English Composition Test, the results of the single-stage procedure applied to the American History and European History Tests provided scaled scores somewhat higher at the upper end of the score range and somewhat lower at the lower end of the score range than those associated with the current scale. For the Literature Test, scaled scores obtained from application of the single-stage procedure were almost identical to scores on the current scale.

The results of the experimental scalings of the Mathematics Level I and II Tests, displayed in Table 4, are somewhat different from those obtained for the previously discussed tests in that the single-stage scaling for the Level II Test does not provide results that are similar to those obtained for the other tests. As can be seen, results obtained for the Mathematics Level I Test are similar to the other tests evaluated so far in that the scaled scores from the single-stage results are somewhat higher at the upper end of the score range and lower at the lower end of the range than scaled scores on the current scale. On the other hand, the single-stage results obtained for the Mathematics Level II Test are consistently lower than the current scale throughout the entire scaled score range.

Examination of the information provided in Table 4 for the Biology, Chemistry, and Physics Tests shows that the scaled scores provided by all the experimental procedures, with the exception of the single-stage scaling procedure, are similar to those obtained for the other tests discussed so far. The single-stage scaling procedure used with the Biology Test provides scaled scores that are somewhat lower at the top and higher at the bottom of the scale score range when compared to scaled scores on the current scale. Results of the single-stage scaling procedure used with the Chemistry and Physics Tests provide scaled scores that are slightly higher at the top of the scaled score range and also somewhat higher at the bottom of the range when compared with scaled scores on the current scale.

The scaling results for the French Test provided in Table 4 indicate that four of the experimental procedures, first stage scaling, second stage scalings A and B, and the empirically based single-stage scaling, all yield quite similar results. The results of these four procedures all provide

scores that are lower throughout the scaled score range than scaled scores on the current scale. The results of scaling to the French Test are, of course, identical to the current scale. The results of the single-stage scaling provide scaled scores that are similar in the upper end of the score range and somewhat higher in the lower end of the score range than scaled scores on the current scale.

The results for the experimental scaling procedures used with the German Test, which are summarized in Table 4, indicate that the two second stage scaling procedures and the empirically based single-stage procedure all provide results that are quite similar. These procedures provide scaled scores that are somewhat lower than scaled scores on the current scale. Scaling to the French Test and single-stage scaling provide fairly similar results when applied to the German Test. Both of these procedures provide results that are slightly higher at the top of the scaled score range and somewhat lower at the bottom of the range than scaled scores on the current scale.

The Latin Test results presented in Table 4 show that the two second stage scaling procedures provide almost identical results; i. e., scaled scores that are lower at the top of the scale and almost the same at the bottom of the scale as scaled scores on the current scale. In addition, the results of the first stage scaling procedure and the empirically based single-stage scaling procedure are very similar, providing scaled scores that are lower at both the top and bottom of the scaled score range than scaled scores on the current scale. The results of scaling to the French Test and the single-stage scaling procedures are different from each other and from the results of the other procedures. The results of scaling to the French Test

provide scaled scores that are lower at the bottom and the top than scaled scores on the current scale. The results of the single-stage scaling procedure provide scores that are somewhat lower at the top and higher at the bottom than scaled scores on the current scale.

Data provided in Table 4 for the Spanish Test indicate that the results of all the scaling procedures, with the exception of scaling to the French Test and single-stage scaling, are quite similar to each other. These procedures all have a tendency to provide scaled scores that are lower at both the top and bottom when compared with scaled scores on the current scale. The results of scaling to the French Test and single-stage scaling are similar in that both the procedures provide scaled scores that are higher at the top and lower at the bottom than scaled scores on the current scale.

The effect of the various scaling procedures on the summary statistics provided for the thirteen Achievement Tests used in this study can be seen by examining the information provided in Table 5. Table 5 presents the Achievement Test scaled score summary statistics resulting from application of each of the experimental scaling procedures as well as the summary statistics and correlations of the scaling covariates, SAT-V and SAT-M, and semesters of foreign language study (for the language tests), with Achievement Test scores.

Insert Table 5 about here

It is clear from examination of the information provided in Table 5 for the English Composition Test that application of the first and second stage scaling results and the empirically based single-stage scaling procedure result in similar scaled score summary statistics. Application of the results of the single-stage scaling procedure provides scaled score summary statistics

that are quite different from those obtained by the other experimental procedures. It is clear that none of the experimental procedures result in scaled score means that are similar to those obtained using the current scale values for the test.

Information provided in Table 5 for the Literature Test shows close agreement among summary statistics obtained by application of the results of the first and second stage scalings and the empirically based single-stage scaling procedure. In addition, scaling results obtained by application of the single-stage scaling procedure agree almost perfectly with summary statistics obtained using current scale values.

Results of the experimental scalings of the American History and European History Tests, presented in Table 5, are similar to those obtained for the Literature Test. These results indicate a high level of agreement among scaled score summary statistics obtained for the first and second stage scalings and the empirically based single-stage scaling procedure applied to these two tests. As was observed for the Literature Test, results obtained for the single-stage scaling procedure used with the American History and European History Tests are very similar to summary statistics obtained using the current scale values.

Information provided in Table 5 for the Math Level I and Level II Tests are similar to that provided for the tests discussed so far in that the first and second stage scalings and the empirically based single-stage scaling all provide similar scaled score summary statistics for the respective tests. On the other hand, the results for both of these tests are similar to those obtained for the English Composition Test in that the summary statistics obtained as a result of applying the single-stage scaling procedure differ somewhat from those associated with the current scale values.

The results provided in Table 5 for the Biology, Chemistry, and Physics Tests show that the summary statistics resulting from the application of the first and second stage scalings and the empirically based single-stage scaling are similar for each of the respective tests. For the Biology Test, the summary statistics resulting from application of the single-stage scaling procedure agree quite closely with those resulting from the current scale values. For both the Chemistry and Physics Tests, summary statistics resulting from the application of the single-stage scaling procedure are somewhat different from those obtained using current scale values.

Examination of the results presented in Table 5 for the French Test indicate close agreement among the summary statistics resulting from application of the first stage scaling and the second stage scalings A and B. Of course, the summary statistics obtained from scaling to the French Test are identical to the current scale values. The summary statistics provided by the single-stage scaling results and the empirically based single-stage scaling results differ from each other and also from the current scale values.

The results presented in Table 5 for the German Test are inconsistent with those obtained for the French Test. Summary statistics obtained for the two second stage scalings agree quite closely with each other. Summary statistics obtained by scaling to the French Test and from the single-stage scaling procedure also agree closely with each other and are reasonably close to the summary statistics derived from the current scale.

The results of the Latin Test scalings presented in Table 5 indicate that the only procedures providing similar summary statistics are the two second stage scaling procedures. The summary statistics resulting from the first stage scaling and the empirically based single-stage scaling agree somewhat

for this test, as do the summary statistics obtained by application of the single-stage scaling procedure and the current scale values. Summary statistics obtained by scaling to the French Test are not in close agreement with those obtained by any of the other experimental scaling procedures.

The Spanish Test scaling results presented in Table 5 indicate reasonably close agreement among the summary statistics obtained for the first and second stage scalings and the empirically based single-stage scaling. Summary statistics resulting from scaling to the French Test and the single-stage scaling procedure are in reasonably close agreement and agree fairly well with summary statistics obtained using the current scale values.

The additional information presented in Table 5 that should be noted at this point are the correlation coefficients of Achievement Test scores with the scaling covariates. It can be seen, from examination of the information presented in Table 5, that the thirteen tests generally fall into three categories: 1) tests that correlate highly with SAT-V scores; 2) tests that correlate highly with SAT-M scores; and, 3) tests that do not correlate highly with either SAT-V or SAT-M scores.

Tests such as English Composition, American History, Literature, and European History are all tests that show a higher relationship between Achievement Test scores and SAT-V scores than between Achievement Test scores and SAT-M scores. Tests showing a higher correlation of Achievement Test scores with SAT-M scores than with SAT-V scores are the two Math tests and the Physics and Chemistry Tests. The Biology Test scores, unlike the other science test scores, show a slightly higher relationship with SAT-V scores than with SAT-M scores. Scores on the foreign language tests do not correlate particularly well with either SAT-V or SAT-M scores. The language test scores

that show the highest relationship with scores on SAT-V and SAT-M are the Latin Test scores. The language test which has scores that exhibit the lowest correlations with SAT-V and SAT-M scores is the German Test.

One way to evaluate the results of the experimental scaling procedures is to evaluate the rank ordering of the scaled score means obtained for the groups actually taking the tests in relationship to the groups' ability levels, as assessed by the scaling covariates SAT-V and SAT-M. If the underlying abilities measured by the various Achievement Tests were equally and perfectly correlated with abilities measured by the covariates, and the scales of the tests were aligned, one would expect the rank ordering of the group means obtained on the Achievement Tests to match those obtained by the groups on the covariate measures. As just noted, the tests differ in their relationship to the covariates, so an examination of the ranking of the Achievement Test scaled score means in relationship to SAT-V and SAT-M scaled score means can provide only a rough evaluation of the scaling results.

The results of the rank ordering of the scaled score means are presented in Table 6. A pragmatic criterion based on a combination of SAT-V and SAT-M was formed, and scaled score means obtained on SAT-V and SAT-M were simply summed for each group taking a particular Achievement Test. This scaled score sum was then used to rank order the thirteen tests from high to low. Scaled score means obtained using current scale values and the results of first stage scaling, second stage scaling A, single-stage scaling, empirically based single-stage scaling, and scaling to the French Test were used to rank order the tests and these orders were then compared to the rank ordering obtained using the summed SAT-V and SAT-M scaled scores. It can be seen, from examination of the information presented in Table 6, that none of the rank

orderings associated with the current Achievement Test scale or the experimental scaling procedures evaluated exactly reproduce the rank ordering that occurs using the summed SAT-V and SAT-M means.

Insert Table 6 about here

Some consistency clearly does exist in the rank orderings provided in Table 6. For example, Math Level II and Physics are the two top ranked tests regardless of scaling procedure and regardless of whether Achievement Test score or SAT sum is used. Scaling to the French Test scale definitely provides a higher rank ordering for the language tests than the other scaling procedures under consideration. This higher ranking for the language tests is consistent with the high ranking these tests receive on the SAT sum. It should also be noted that both the current scale and the second stage scaling results preserve the rank ordering of the four language tests obtained using the sum of SAT means. On the other hand, the remaining scaling procedures place the French Test scaled score mean above the German Test mean, which is inconsistent with the rank ordering obtained using the sum of SAT means.

As a means of assessing the degree of consistency between the rank orderings of the scaled score means obtained from the five scaling procedures, the rank ordering obtained from the current scale, and the rank ordering based on the summed SAT-V and SAT-M means, Spearman rank order correlation coefficients were calculated and are reported in Table 6. (Ties were resolved by reranking using data to more decimal places than shown in Table 6.) The rank order correlation between rank orderings based on summed SAT-V and -M means and the current scale is .687. The rank order correlations between the single-stage scaling and the empirically based single-stage scaling and the

rank ordering based on the summed SAT-V and SAT-M means are .538 and .896, respectively. The rank order correlations between rank orderings based on summed SAT-V and -M means and the second stage scaling and scaling to the French Test are .929 and .926, respectively. Finally, the rank order correlation between the first stage scaling results and the SAT-V and SAT-M sum is .846. The experimental scaling procedures that use empirically derived data for the reference populations appear to provide more consistent orderings of means with the ordering provided by the summed SAT-V and -M means than do the orderings based on means from the current scale or the single-stage scaling procedure.

It is also interesting to note that the current scale and the single-stage scaling procedures have a tendency to rank tests that have scores that show a strong relationship to SAT-M scores higher than tests providing scores that have a strong relationship with SAT-V scores. This is not surprising given that the current SAT-V and SAT-M scales are not well aligned and the SAT-M scale is higher than the SAT-V scale. There are, however, some exceptions to this rule, particularly the Latin Test. Another point worth noting is that all rank orderings, with the exception of that provided by the current scale, rank Math Level I as the lowest ranked test. Finally, if the language tests are ignored and only the rank orderings of the non-language tests are evaluated, it can be seen that the current scale, the second stage scaling A, the first stage scaling, the empirically based single-stage scaling, and the scaling to the French Test (which provides almost the same results as the second stage scaling) result in a rank ordering that is the same and consistent with the ranking obtained using the SAT sum for the Math Level II, Physics, Chemistry, and European History Tests. On the other hand,

ignoring the language tests, the experimental procedures result in a fairly different rank ordering of the Math Level I, Biology, Literature, American History and English Composition Tests. The current scale and the single-stage scaling procedures have a tendency to place Math I higher and the Literature and English tests lower than either the SAT sum or the remaining scaling results.

One way to evaluate the alignment of the Achievement Test scales is to examine plots of Achievement Test scaled score means, conditioned on the three covariates used for the experimental scalings. Recall a definition of scaling given by Keeves (1988) in an earlier section of this paper. Keeves quoted Howard (1958) as saying; "In practice these requirements [key requirements of a scaling procedure] demand that the same mark on different examinations should imply the same level of performance relative to a common population." The common, or reference population, used for the various experimental scaling procedures differed from procedure to procedure, as is illustrated by the information presented in Table 2; consequently, it was thought useful to evaluate the relationship among scaled score means obtained for the Achievement Tests at 100 point intervals along the SAT-V and SAT-M score scale. In addition, Achievement Test means were evaluated by examining foreign language test means conditioned on semesters of study of a foreign language, which ranged from three to nine semesters.

The results of these analyses are plotted in Figures 1a-1o and Figures 2a-2e. Figures 1a-1o contain, for the current scale and each of the experimental scaling methods, plots of Achievement Test scaled score means for groups of examinees with selected scores on the scaling covariates. Three plots appear for each scaling procedure. One plot shows Achievement Test

means conditioned on SAT-V scores, the second plot shows Achievement Test means conditioned on SAT-M scores, and the third plot shows language Achievement Test means conditioned on semesters of study of a foreign language.

Figures 2a-2e are simply a rearrangement of the plots shown in Figures 1a-1o; i.e., Figures 2a-2e show plots for all three covariates used for a single scaling procedure on the same page and hence are more useful for evaluating trends across the covariates. Because the plots provided in Figures 1a-1o are larger, they permit a clearer evaluation of the behavior of the individual tests represented by the symbols in the plots and, hence, are included in the paper.

Insert Figures 1a-1o and Figures 2a-2e about here

Figures 1a-1c show plots of Achievement Test means on the current scale conditioned on SAT-V, SAT-M, and semesters of study of a foreign language, respectively. Examination of the information provided in Figure 1a indicates a considerable spread among all the conditional means, although the spread appears to be less in the vicinity of an SAT-V scaled score of 500. If one looks only at the grouping of Achievement Test means at an SAT-V mean of 500, it is apparent that the Math Level I, Chemistry and Physics Tests form one cluster of scores, that the remaining tests, with the exception of Math Level II, form a second cluster of means, and that the Math Level II Test provides higher scaled scores than any of the other Achievement Tests. The plots provided in Figure 1b show somewhat similar results to those given in the first figure. It can be seen that the Achievement Test means tend to cluster at a scaled score mean of 500 on the SAT-M scale. Again, the Math Level II

mean scores appear higher than mean scores obtained on the other Achievement Tests. Figure 1c shows the relationship among language test means for the four foreign language tests conditioned on semesters of study. It is apparent, even after conditioning on amount of training, that there is still considerable variability in mean scores, with the Spanish Test consistently providing the lowest scores and the Latin Test, the highest.

The results of the single-stage scaling procedure presented in Figures 1d-1f can be contrasted to results shown in the previously discussed plots. It appears that the Achievement Test conditional means resulting from the single-stage scalings are slightly less dispersed than those observed for the current scale for all three covariates. A major difference between the information provided in these plots and those provided for the current scale is that the Math Level II means, conditioned on SAT-M scores, do not appear as high as the means shown for the same test resulting from the current scale.

Figures 1g-1i show conditional Achievement Test scaled score means resulting from the first stage scalings. It should be kept in mind that the reference populations are now centered at means that are above 500 on SAT-V and -M (see Table 2). An examination of the information provided in Figure 1g shows that Achievement Test means conditioned on SAT-V means are fairly tightly clustered for mid to upper SAT-V scaled score ranges. Math Level II means appear to be more closely related to the means obtained on the other tests for this particular scaling procedure than observed for the Level II conditional means for the scaling procedures previously evaluated. Conditional means provided in Figure 1h, i.e., Achievement Test means conditioned on SAT-M scores, appear to provide a slightly tighter clustering of Achievement Test means than the clustering observed in the previously

discussed plots that involved conditioning on SAT-M. The plots shown in Figure 1i, which display language test means resulting from the first stage scaling conditioned on semesters of study, also indicate a closer agreement among these means, particularly for seven semesters of study (the reference population mean) than the comparable plots previously evaluated.

Figures 1j -1l show Achievement Test conditional means resulting from scaling to the French Test. Achievement Test means shown in these plots are for the foreign language tests only. A comparison of the foreign language test means conditioned on SAT-V and SAT-M scaled scores (presented in Figures 1j and 1k) show the means to be reasonably clustered at different points on the V and M scaled score continuums. A comparison of the information shown in Figure 1l with that shown in Figures 1c, 1f, and 1i indicates that scaling to the French Test has a tendency to cluster foreign language test means, conditioned on semesters of study, a little more tightly.

Figures 1m-1o show Achievement Test conditional means resulting from the second stage scalings. The plots shown in Figures 1m-1o are almost identical to those shown for the first stage scaling results. The tighter clustering of conditional Achievement Test means, as compared to those obtained by the single stage scaling or those for the current scale, is evident. In addition, Math Level II scores resulting from the second stage scaling seem to provide conditional means closer to the other Achievement Test means than the Level II means that resulted from the single-stage scaling or the current scale.

As mentioned previously, Figures 2a-2e contain the same plots illustrated in Figures 1a-1o; however, the plots shown in Figures 2a-2e have been condensed so that plots of conditional means for all covariates used for a particular scaling method can be shown on a single page. Figure 2a contains

plots of conditional means obtained using current scale values. It can be seen, from an examination of the plots contained in Figure 2a, that there is considerable scatter of Achievement Test means conditioned on SAT-V and SAT-M scores, as well as for the language test means conditioned on semesters of study.

Examination of the information provided in Figure 2b, which illustrates conditional means resulting from the single stage procedure, shows a slight reduction in the scatter of the Achievement Test means compared to those obtained using current scale values; at least for those conditioned on SAT-V and SAT-M scores. Figure 2c contains plots of Achievement Test means resulting from the first stage scaling procedure. This scaling procedure appears to have resulted in a noticeable reduction in the scatter of the Achievement Test means for all three covariates.

An examination of the plots shown in Figure 2d with the bottom panel of Figure 2c permits a comparison of the results of scaling the language tests to the French Test scale with those obtained by the alternative first stage scaling procedure. It appears that the scatter of foreign language Achievement Test means, conditioned on semesters of study, is quite similar for the two scaling procedures.

Finally, the plots shown in Figure 2e illustrate conditional Achievement Test means resulting from the second stage scaling. The degree of scatter observed in Figure 2e for the Achievement Test means is very similar to that observed for the first stage scaling results.

Another way to evaluate the alignment of the Achievement Test scales resulting from application of the experimental scaling procedures is to examine the relationship between scaled score means on pairs of Achievement

Tests, where each pair was taken by the same group of students. The assumption is that, if the students were equally prepared in the subject matter tested by each test in the pair, and if the tests were measuring the same underlying ability, tests with aligned scales would show similar mean scores. Figures 3a-3d provide bivariate plots of Achievement Test scaled score means resulting from the experimental scaling procedures. The data used to provide the scaled score values plotted in Figures 3a-3d were obtained from a recent administration of the Achievement Tests and are not necessarily representative of the groups of students used for the experimental scalings.

Insert Figures 3a-3d about here

The plots shown in Figures 3a-3d demonstrate the relationship between pairs of Achievement Test scaled scores representing the current scale, first stage scaling results, single-stage scaling results and scaling to the French Test. (Scaled score means for the non-language tests in the scaling to the French Test plot were derived from the results of the first stage scaling of these tests.) Points falling closer to the diagonal line on the plots represent pairs of Achievement Test means that are in closer agreement with each other than pairs represented by points that are farther away from the diagonal line. Table 7 provides scaled score values for the points plotted in Figures 3a-3d. In addition, Table 7 provides SAT-V and SAT-M scaled score means for the respective groups as well as the correlation of Achievement Test scores with SAT-V and SAT-M scores. The particular scaling results used to obtain the mean scores for the pairs of Achievement Tests shown in Figures 3a-3d and Table 7 were chosen because they represent, to a certain extent, the extremes of scale alignment provided by the results obtained in this study.

Insert Table 7 about here

Examination of the information provided in Figure 3a, which shows a plot of pairs of Achievement Test scaled score means based on the current scale, indicates that test pairs showing the most agreement in scaled score means are Biology and Chemistry, English Composition and European History, English Composition and Literature, English Composition and Biology, Biology and American History, and English Composition and American History. Points falling the farthest from the diagonal line represent the following pairs of tests: Chemistry and Math Level I, English Composition and Physics, Latin and Math Level I, and Latin and English Composition. The remaining points on the plot are somewhat intermediate in agreement.

The single-stage scaling results are represented by the points plotted in Figure 3b. It can be seen, from examination of the points plotted in this figure, that points falling closest to the diagonal line are those associated with the following tests: English Composition and Biology, French and American History, and Biology and American History. Points representing test pairs falling the farthest from the diagonal are English Composition and French, English Composition and Latin, English Composition and Physics, and Chemistry and Math Level I.

Figure 3c contains a bivariate plot of scaled score means resulting from application of the first stage scaling procedure. It can be seen from an examination of this figure that points falling closest to the diagonal line represent the following test pairs: English Composition and Literature, English Composition and Chemistry, French and American History, Biology and Chemistry and Biology and Math Level I. Points falling farthest away from the

diagonal line in Figure 3b are those representing English Composition and French, English Composition and Latin, English Composition and Physics, and Chemistry and Math Level I.

Finally, the results of scaling to the French Test are shown by the points plotted in Figure 3d. It should again be noted that values presented in this plot and in Table 7 for the non-language tests are those obtained from the first stage scaling of these tests. Examination of the information provided in Figure 3d indicates that pairs of tests falling close to the diagonal line are English Composition and Literature, English Composition and French, English Composition and Chemistry, French and American History, Latin and Math Level I, and Biology and Math Level I. Pairs of tests falling the farthest from the diagonal are English Composition and Latin, English Composition and Physics, French and Math Level I, and Chemistry and Math Level I.

While it is difficult to draw conclusions regarding the different scaling procedures from an examination of the data provided in the four figures, a few generalizations can be made. For one, the scaling procedure resulting in the largest number of points (test pairs) falling close to the diagonal line was the procedure that involved scaling to the French Test scale. The procedure that resulted in the fewest number of pairs falling close to the diagonal line was the single-stage scaling procedure. Secondly, it seems reasonable to expect tests such as, for example, English Composition and Literature to provide scaled score means somewhat similar for a group of examinees taking both tests if the test scales were aligned. For this test pair, this is clearly the case for the means resulting from the current scale and from all experimental scaling procedures with the exception of the single-stage scaling procedure.

Three Achievement Test pairs provided results that were consistently discrepant regardless of whether the current scale or an experimental scaling procedure was used. These test pairs were English Composition and Latin, English Composition and Physics, and Chemistry and Math Level I. One would hardly expect the English Composition and Physics Tests to provide scores measuring a single underlying skill or ability and hence, similar scaled score means for the same group of students. However, it might be expected that tests such as Chemistry and Math Level I share a sufficiently common base of knowledge or skills that the test scores obtained on these tests by the same group of students should be somewhat similar. Examination of the correlation coefficients supplied in Table 7 for this pair of tests indicates that, although the tests are reasonably correlated with each other, they show a differential correlation with the scaling covariates. Math Level I is much more highly correlated with SAT-M scores than scores obtained on the Chemistry Test. This differential correlation clearly appears to be reflected in the pair of Achievement Test scaled score means and is an indication that even when Achievement Tests share some commonality in the skills they measure, they may not share a similar relationship with the scaling covariates, thus complicating even further the scaling process.

DISCUSSION AND CONCLUSIONS

The purpose of this study was to evaluate an experimental scaling procedure suggested in an earlier study by Cook (1988). The proposed experimental scaling method involved the use of a two stage scaling procedure. The first stage attempted to scale tests by cluster (language versus non-language tests) and to maximize the alignment of scales of tests within a cluster. The second stage of the procedure was carried out in an attempt to

align scales for tests across clusters. The first stage scalings of the tests in the language test cluster were based on two different procedures. One procedure used, as reference group values for the scaling covariates, data aggregated across all the language tests in the cluster. The alternative first stage scaling procedure consisted of scaling the German, Latin, and Spanish Tests to the French Test scale. In addition to the two stage scaling procedure, which was the focus of this study, a single-stage procedure and an empirically based single-stage procedure were also used for comparative purposes.

Finally, the effect of changing the procedure used to select scaling samples for the Biology and Chemistry Tests was evaluated. The first and second stage experimental scalings included high school sophomores in the scaling samples for these two tests, while the two single-stage scaling procedures did not.

The results of the two stage procedure were somewhat disappointing in that, in almost all cases, the results of the first stage scaling (scaling within cluster) and the results of the second stage scaling (scaling across clusters) were very similar. The reason for this is fairly clear. If one examines the information provided in Table 2, which shows values of scaling covariates used for the reference groups for the first and second stage scalings, it is apparent that these values change very little from first to second stage and hence have little effect on the scalings. The only exception to this statement is the reference group mean of SAT-V scores obtained for the language test cluster.

As mentioned previously, the reason the reference group values change so little from the first to the second stage scaling is due to the fact that the

groups that take the tests (language versus non-language) are similar in ability as assessed by SAT-V and SAT-M scores and also, for the second stage scalings, the non-language group, because it contains more tests and the tests are given to more students, has a greater influence on the reference group values determined by aggregating across clusters.

Although the first stage results derived from the separate scalings of the language and non-language tests did not differ much from the second stage scaling results, the results of scaling the language tests to the French Test scale did provide results differing from second stage scaling B results. The interesting thing about these results is that the second stage scaling B results, the second stage scaling A results, and the first stage scaling results all agree fairly well with each other and disagree with the results obtained by scaling the language tests to the French Test scale.

A second aspect of the study, varying the manner in which the scaling samples are selected for the Biology and Chemistry Tests by including high school sophomores in the scaling samples, can be evaluated by examining the information provided in Table 3. Perusal of this information indicates that the addition of sophomores to the samples had very little effect on the summary statistics for the covariates (SAT-V and SAT-M scores) used in the scaling procedures. The addition of sophomores to the scaling sample did, however, affect the summary statistics obtained for the respective Achievement Tests. Also, it should be noted that addition of sophomores to the scaling sample for the Biology Test had a tendency to lower the correlations between the scores on the two covariates and Achievement Test score, while, addition of sophomores to the scaling sample for the Chemistry Test had no effect on these correlations.

A useful framework for evaluating the results of the current study is provided by statements made by both Angoff (1968) and Keeves (1988) about the desired outcomes of the scaling process. Recall, Angoff stated the desired outcome of a scaling procedure is to ensure that candidates choosing to compete with more able candidates will not be put at a disadvantage. According to Angoff, "...a candidate who is average in a highly selected group of candidates should earn a higher scaled score than a candidate who is average in a less able group." According to Keeves (1988) one of the key requirements of moderation [scaling] procedures is to ensure that, "...the same mark on different examinations should imply the same level of performance relative to a common population." With Angoff's and Keeves' statements of desired outcomes of the study in mind, it is useful to focus on the information provided in Tables 6 and 7 and Figures 1a-1o, 2a-2e, and 3a-3d.

The information provided in Table 6 is an attempt to use Angoff's definition of a desired outcome of a scaling method to evaluate the experimental scaling procedures. Using Angoff's definition, if the covariates used for the scalings are measures of the abilities underlying scores on the thirteen tests and the covariates are related to the tests in a similar manner, one would expect the rank ordering of groups taking the tests to be similar to a ranking obtained using scores on these covariates. An examination of the information provided in Table 6 shows that the scaling procedures do effect the rank orderings of the tests and that the procedures employing reference group values that are empirically based provide rank orderings that are more similar, when compared to rankings obtained using the sum of SAT-V and SAT-M means, than are rank orderings produced by the current scale and by the single-stage scaling procedure which employs a reference

group with a mean and standard deviation of SAT-V and SAT-M scores of 500 and 100, respectively. The information provided in Table 6 indicates that the rank order correlation coefficient is somewhat affected by whether or not results from the first stage scaling procedure are used or results provided by the additional scaling across the language and non-language clusters represented by the second stage scaling procedure are used. It appears as though closer agreement with the ranking obtained on the sum of SAT-V and SAT-M means is realized by Achievement Test means resulting from the second stage scaling.

Although the effect on rank order correlation coefficients of the various scaling procedures is apparent, interpretation of these effects is not so clear. Recall, the assumption is that the scaling covariates, SAT-V and SAT-M, measure the same underlying abilities as measured by the thirteen Achievement Tests used in this study and that the relationship of the covariates with the Achievement tests is similar across tests. Dramatic evidence that this assumption is not true is presented by the correlation coefficients given in Table 5. As noted previously, the thirteen tests show very different patterns of correlations with the covariates. In addition, the results of the rankings of the Achievement Tests must be interpreted with caution since in a number of instances very small differences between Achievement Test means result in different rank orderings of the tests. Keeping the above mentioned caveats in mind, if one were to use agreement in rank ordering of means between a particular scaling procedure and the covariates as a criterion in the choice of a scaling procedure, it appears that either the second stage scaling procedure used with all tests or the empirical single stage scaling used for the non-language tests coupled with

the second stage scaling used with the language tests would be the procedures of choice.

A second way to evaluate the experimental scaling procedures used for this study is to employ Keeves' key requirement for scaling; i.e., that scores on the Achievement Tests should imply the same level of performance relative to a common population. The best information to use to evaluate the tests from this point of view is the information provided in the plots shown in Figures 2a-2e. An examination of the plots shown in these figures indicates that both the first stage scaling and second stage scaling procedures reduce the scatter of Achievement Test conditional means when compared to that observed for the single-stage procedure and the current scale. For all of the procedures represented in the plots shown in Figures 2a-2e, the scatter of Achievement Test conditional means is less as one approaches the value of the covariate mean of the reference population.

The fact that scatter of the conditional Achievement Test means plotted for extreme values of the covariates is much greater than that observed for values surrounding the reference population mean is understandable given that the experimental scaling procedures are all linear scaling procedures. Some type of non-linear scaling procedure would necessarily need to be used to maintain a similar level of agreement among Achievement Test conditional means throughout the entire range of covariate scores.

It is important to note again, that one would not expect exact agreement among Achievement Test means conditioned on a particular covariate value unless the covariate is measuring the same underlying ability for all the tests and is related to the tests in a perfect manner. As mentioned previously, there is clear evidence given in Table 5 that this situation is

not met. However, keeping the data presented in Table 5 in mind, one could interpret the results provided in Figures 2a-2e as providing some indication that the first and second stage scaling procedures are somewhat more successful in aligning Achievement Test means than is a procedure that is based on a reference population with scaled score means and standard deviations of 500 and 100 on SAT-V and SAT-M.

The information regarding the relationship of Achievement Test means for groups taking pairs of the tests, presented in Figures 3a-3d and Table 7, is very difficult to interpret. This may be because the covariate means of the particular group taking a pair of Achievement Tests are very different from those specified for the reference groups used for the scaling procedures as well as being very different from pair to pair. As mentioned previously, it appears as though the procedures based on empirically derived scores on the covariates for the reference populations result in pairs that are in slightly closer agreement than a procedure that involves using values of 500 and 100 for the reference population scaled score means and st ıard deviations.

The question must be asked, even if scales for, say, the Biology Test and Math Level I Test were perfectly aligned, given that the tests measure different skills and are selected by examinees for different reasons, should one expect scores on these tests for the same group of examinees to be the same? The answer to this question is probably no. Thus, examining the relationship among means obtained on pairs of Achievement Test scores is probably not the most effective way of choosing one scaling procedure from a group of potential scaling procedures.

An important point to note is the dramatic effect that the use of empirical values for reference group scores on the scaling covariates has on

the placement of the Achievement Tests on the 200 to 800 scale. Achievement Test scaled score means obtained for the experimental procedures employing empirical values for the reference group and Achievement Test means obtained employing the current scale are very different (see Table 5). Also, the relationship between scaled score means obtained on SAT-V and SAT-M and the Achievement Tests, for the respective groups taking the tests, is quite different depending upon whether or not empirical values are used for reference group covariate scores.

Examination of the information provided in Table 5 indicates that use of empirical values for the reference group results in a much lower placement on the 200 to 800 scale for Achievement Test means relative to the current scale. This is due to the fact that the empirically defined reference population is a very able group, as assessed by SAT-V and SAT-M scores. Estimating scores for this group on the respective Achievement Tests and subsequently placing these scores on the current scale results in the low scale placement of Achievement Test means observed in Table 5 for the empirically based scaling procedures. It should be recalled, however, that the primary purpose of the scaling procedures evaluated in this study is to promote comparability of scales for the thirteen Achievement Tests, not necessarily to promote comparability of Achievement Test scales with the current scale. Thus, lower scale placement of scores due to the use of observed reference group values is not necessarily a disadvantage of the empirically based procedures.

To summarize, the purpose of this study was to evaluate whether an experimental scaling procedure that employed observed reference group values for the covariates (SAT-V and SAT-M) would provide better results than a procedure that used as reference group values scaled score means of 500 and

standard deviations of 100. In addition, the feasibility of scaling in two stages was evaluated. The purpose of the first stage scaling was to align scales within clusters of tests; i.e., language versus non-language Achievement Test clusters, and the purpose of the second stage scaling was to align score scales across the two clusters. Finally, the effect of augmenting the scaling samples used for the Biology and Chemistry Tests by adding high school sophomores to these samples was evaluated.

The results of the study related to the sampling question indicate that addition of high school sophomores to the samples does not improve the relationship between Achievement Test scores and the scaling covariates (at least as evaluated by the correlations between Achievement Test score and the covariates) and thus is probably not an appropriate change to consider. The results of the study related to the investigation of the two stage scaling procedure indicate there is some reasonable evidence to suggest that the use of empirical values for the reference group and use of the procedure that involves scaling the tests in two stages may improve the alignment of the Achievement Test scales. A viable alternative involves application of the empirical single-stage scaling to the non-language tests and the two-stage scaling to the language tests. This combination of procedures appears to provide a comparable degree of alignment of the scales as that provided by applying the two stage procedure to all tests, and should be somewhat easier to implement.

The results of the study should be interpreted with caution because of the circular nature of the criterion. In other words, the requirements of the scaling were specified, a scaling procedure was developed based on these requirements, and the criterion used to evaluate the results of the scalings

was based on the same requirements. As mentioned several times, the scaling procedures, as well as the criterion used to evaluate the procedures, were based on an untenable assumption; i.e., that the covariates measured the same underlying abilities for all the tests and that the relationship between the covariates and the Achievement Test was similar for all tests. This assumption is clearly impossible to meet. Given this situation, the best that can be expected is a rough alignment of the test scales. This rough alignment does appear to be provided by the two stage procedure evaluated in this study.

It is important that the results of this study be considered tentative and that further research be carried out. Procedures that involve modeling and correcting for the selection bias present in the Achievement Test scores might prove fruitful. In addition, use of non-linear scaling procedures that should provide improved alignment for scores throughout the entire scaled score range for the tests might also be desirable and should be further investigated.

REFERENCES

Angoff, W. H. (1968). How we calibrate College Board scores. The College Board Review, 68, 11-14.

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

Cook, L. L. (1988). Achievement Test scaling (RR-88-34). Princeton, NJ: Educational Testing Service.

Cooney, G. H. (1975). Standardization procedures involving moderator variables--some theoretical considerations. Australian Journal of Education, 19, 50-63.

Cooney, G. H. (1976). Scaling procedures: A review of Australian practices. Australian Mathematics Teacher, 32, 57-62.

Dorans, N. J. (1985). Achievement Test rescaling: Help or hindrance? (SR-85-92). Princeton, NJ: Educational Testing Service.

Howard, M. (1958). The conversion of scores to a uniform scale. British Journal of Statistical Psychology, 11, 199-207.

Keeves, J. P. (1988). Scaling achievement test scores. In J. P. Keeves (Ed.), Educational research, methodology, and measurement. An international handbook. Elmsford, NY: Pergamon Press, Inc.

Wilks, S. S. (1961). Scaling and equating College Board tests. Princeton, NJ: Educational Testing Service.

/gt
dre\aligning.rr2

Table 1

Achievement Test Background Questionnaire Used to Collect Covariate Information

French Test[1]

In the group of nine spaces labeled Q, you are to blacken ONE and ONLY ONE space, as described below, to indicate how you obtained your knowledge of French. The information that you provide is for statistical purposes only and will not influence your score on the test.

Question 1

If your knowledge of French does not come primarily from courses taken in grades 9 through 12, blacken space 9 and leave the remaining spaces blank, regardless of how long you studied the subject in school. For example, you are to blacken space 9 if your knowledge of French comes primarily from any of the following sources: study prior to the ninth grade, courses taken at a college, or special study, residence abroad, or living in a home in which French is spoken.

If your knowledge of French does come primarily from courses taken in grades 9 through 12, blacken the space that indicates the level of the French course in which you are currently enrolled. If you are not now enrolled in a French course, blacken the space that indicates the level of the most advanced course in French that you have completed.

Level I:	first or second half	- blacken space 1
Level II:	first half	- blacken space 2
	second half	- blacken space 3
Level III:	first half	- blacken space 4
	second half	- blacken space 5
Level IV:	first half	- blacken space 6
	second half	- blacken space 7
Advanced Placement or course that represents a level of study higher than Level IV: second half		- blacken space 8

If you are in doubt about whether to mark space 9 rather than one of the spaces 1-8, mark space 9.

[1]The same questionnaire (with the appropriate test name) appears in the French, German, Latin and Spanish Tests. The Latin questionnaire differs slightly in that the phrase, "...or living in a home in which [language] is spoken" is eliminated.

Table 2

Scaled Score Summary Statistics for Reference Populations Used in Achievement Test Scaling Study

	First Stage Scaling Lang.	First Stage Scaling Non-lang.	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling	Scaling to the French Test		
						French/German	French/Spanish	French/Latin
SAT-V								
Mean	527	514	515	500	515	541	524	542
Var.	9,434	10,465	10,397	10,000	10,397	8,990	9,383	9,008
SAT-M								
Mean	575	579	579	500	579	583	572	583
Var.	9,265	10,394	10,308	10,000	10,308	8,649	9,263	8,657
SAT-V/ SAT-M								
Cov.	5,456	5,733	5,708	6,000	5,708	4,938	5,436	4,975
Sem. of Study								
Mean	7.1365	---	---	7.1365	7.1365	7.2685	7.1793	7.2310
Var.	2.0482	---	---	2.0482	2.0482	1.8913	1.9888	2.0087
SAT-V/ Sem. of Study								
Cov.	26.4297	---	---	26.4297	26.4297	23.7577	28.4348	25.1457
SAT-M/ Sem. of Study								
Cov.	21.8334	---	---	21.8334	21.8334	19.9454	23.3862	20.4793

Table 3

Scaled Score Summary Statistics for Achievement Test Samples
Used in Achievement Test Scaling Study

Test[1]	Achievement Test Mean	Achievement Test Var	SAT-V Mean	SAT-V Var	SAT-M Mean	SAT-M Var	Semesters of Study Mean	Semesters of Study Var	Covariance of Achievement Test Score and Scaling Covariates $Cov_{v,m}$	$Cov_{v,sem}$	$Cov_{m,sem}$	Correlation of Achievement Test Score and Scaling Covariates $r_{x,v}$	$r_{x,m}$	$r_{x,sem}$
Eng. Comp.	518	9,872	514	10,196	576	10,392	---	---	5763	---	---	.779	.548	---
Literature	528	10,632	527	10,562	545	10,400	---	---	6176	---	---	.831	.526	---
Amer. Hist.	528	9,406	515	9,862	557	10,328	---	---	5755	---	---	.752	.549	---
Europ. Hist.	547	9,070	554	10,323	562	10,700	---	---	5623	---	---	.672	.441	---
Math I	543	8,134	496	9,572	557	8,940	---	---	4984	---	---	.479	.820	---
Math II	660	7,184	545	11,111	646	6,736	---	---	4349	---	---	.435	.783	---
Biology[2]	540 (532)	11,343 (11,449)	514 (513)	10,766 (11,025)	573 (571)	10,305 (10,816)	---	---	6006 (6396)	---	---	.698 (.717)	.616 (.624)	---
Chemistry[2]	572 (567)	10,373 (10,404)	525 (525)	11,575 (11,664)	624 (621)	8,716 (9,025)	---	---	5343 (5824)	---	---	.575 (.575)	.649 (.649)	---
Physics	594	9,451	536	11,708	653	6,995	---	---	4330	---	---	.510	.642	---
French	530	10,333	540	8,976	581	8,622	7.3085	1.8629	4939	25.7969	21.3919	.497	.413	.417
German	533	8,750	551	8,993	599	8,572	6.9217	2.0027	4743	10.0177	11.8836	.355	.304	.336
Latin	548	10,970	559	8,928	600	8,581	6.6340	2.7285	4922	31.5855	25.1534	.533	.485	.362
Spanish	513	9,153	509	9,928	564	9,705	7.0621	2.0740	5634	27.2123	23.1856	.381	.310	.412

[1]Sample sizes are given in Table 5 of the Results section.
[2]Values in parentheses are for samples that do not contain sophomores. These samples were used for the two single-stage scalings.

Table 4

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

ENGLISH COMPOSITION

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	781	782	825	781
750	731	732	773	731
700	681	681	720	681
650	631	631	668	631
600	581	581	615	581
550	531	531	563	531
500	481	481	511	481
450	431	431	458	430
400	381	380	406	380
350	331	330	354	330
300	281	280	301	280
250	231	230	249	230
200	181	180	196	180

LITERATURE

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	773	773	801	773
750	725	724	751	724
700	676	675	701	675
650	627	627	651	627
600	578	578	601	578
550	530	529	551	529
500	481	480	501	480
450	432	431	451	431
400	383	382	401	382
350	335	334	351	334
300	286	285	301	285
250	237	236	251	236
200	189	187	201	187

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

AMERICAN HISTORY

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	774	774	811	774
750	723	723	759	723
700	672	672	707	672
650	621	621	655	621
600	571	570	604	570
550	520	519	552	519
500	469	468	500	468
450	418	417	448	417
400	367	366	396	366
350	316	315	345	315
300	265	264	293	264
250	214	213	241	213
200	163	162	189	162

EUROPEAN HISTORY

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	787	787	818	786
750	734	734	765	734
700	682	682	711	682
650	630	629	658	629
600	577	577	604	577
550	525	524	551	524
500	473	472	497	472
450	420	420	444	420
400	368	367	390	367
350	316	315	337	315
300	263	262	283	262
250	211	210	230	210
200	159	157	176	157

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

MATHEMATICS LEVEL I

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	752	753	835	753
750	699	700	780	700
700	647	647	725	647
650	594	595	671	595
600	541	542	616	542
550	489	489	561	489
500	436	436	506	436
450	383	383	452	383
400	331	331	397	331
350	278	278	342	278
300	226	225	288	225
250	173	172	233	172
200	120	119	178	119

MATHEMATICS LEVEL II

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	699	700	773	700
750	648	648	722	648
700	596	597	670	597
650	545	545	619	546
600	494	494	568	494
550	443	443	516	443
500	391	391	465	391
450	340	340	413	340
400	289	288	362	288
350	237	237	310	237
300	186	185	259	185
250	135	134	207	134
200	83	82	156	83

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

BIOLOGY

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	744	744	792	744
750	697	697	744	697
700	650	649	696	650
650	602	602	648	602
600	555	555	599	555
550	508	507	551	507
500	460	460	503	460
450	413	413	454	413
400	366	365	406	365
350	319	318	358	318
300	271	270	309	270
250	224	223	261	223
200	177	176	213	176

CHEMISTRY

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	744	745	801	745
750	696	697	753	697
700	648	648	704	649
650	600	600	656	600
600	552	552	608	552
550	504	504	560	504
500	456	456	511	456
450	408	408	463	408
400	360	360	415	360
350	312	312	366	312
300	264	264	318	264
250	216	216	270	216
200	168	168	221	168

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

PHYSICS

Current Scale	First Stage Scaling	Second Stage Scaling	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	743	744	806	744
750	696	696	757	696
700	648	648	708	648
650	600	601	659	601
600	553	553	610	553
550	505	505	561	505
500	457	457	511	457
450	410	410	462	410
400	362	362	413	362
350	314	314	364	314
300	267	266	315	266
250	219	219	266	219
200	172	171	216	171

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

FRENCH

Current Scale	First Stage Scaling	Scaling to French Test	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	772	800	770	770	802	773
750	723	750	722	722	754	725
700	675	700	674	674	706	677
650	626	650	626	626	658	629
600	578	600	578	578	610	581
550	529	550	530	530	562	532
500	481	500	482	482	514	484
450	432	450	434	434	466	436
400	384	400	385	385	418	388
350	335	350	337	337	370	340
300	286	300	289	289	322	292
250	238	250	241	241	274	244
200	189	200	193	193	226	196

GERMAN

Current Scale	First Stage Scaling	Scaling to French Test	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	785	814	795	795	805	786
750	733	760	742	742	753	734
700	680	706	689	689	701	682
650	628	653	637	637	649	630
600	575	599	584	584	597	577
550	523	545	531	531	545	525
500	470	491	478	478	493	473
450	418	438	426	425	441	421
400	365	384	373	373	389	369
350	313	330	320	320	337	317
300	261	276	267	267	285	264
250	208	222	215	214	232	212
200	156	169	162	161	180	160

Table 4 (cont.)

Results of Application of Experimental Scaling Procedures to Selected Scale Score Points

LATIN

Current Scale	First Stage Scaling	Scaling to French Test	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	752	779	757	757	787	752
750	704	730	711	711	740	704
700	656	681	664	664	693	657
650	608	631	618	618	646	610
600	560	582	572	572	599	563
550	513	533	525	525	552	516
500	465	483	479	479	505	468
450	417	434	432	432	458	421
400	369	385	386	386	411	374
350	321	336	339	339	364	327
300	274	286	293	293	317	280
250	226	237	246	247	270	232
200	178	188	200	200	223	185

SPANISH

Current Scale	First Stage Scaling	Scaling to French Test	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling
800	793	821	794	794	812	794
750	740	767	742	742	760	742
700	688	713	690	690	707	690
650	636	659	638	638	655	638
600	583	605	586	586	603	585
550	531	551	534	534	551	533
500	478	497	482	482	499	481
450	426	443	430	430	447	429
400	374	389	378	378	395	377
350	321	335	327	327	343	325
300	269	281	275	275	291	273
250	217	227	223	223	239	221
200	164	173	171	171	187	168

Table 5

Summary Statistics Resulting from
Application of Experimental Scaling Parameters

ENGLISH COMPOSITION (n=216,735)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	518	499	499	530	499	514	576
s.d.	99	99	100	104	100	101	102
r(ACH,SATV)	.78						
r(ACH,SATM)	.55						

LITERATURE (n=25,006)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	528	508	507	529	507	527	545
s.d.	103	101	101	103	101	103	102
r(ACH,SATV)	.83						
r(ACH,SATM)	.53						

Table 5 (cont.)

Summary Statistics Resulting from
Application of Experimental Scaling Parameters

AMERICAN HISTORY (n=47,639)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	528	497	496	529	496	515	557
s.d.	97	99	99	100	99	99	102
r(ACH,SATV)	.75						
r(ACH,SATM)	.55						

EUROPEAN HISTORY (n=3,785)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	547	522	521	547	521	554	562
s.d.	95	100	100	102	100	102	103
r(ACH,SATV)	.67						
r(ACH,SATM)	.44						

Table 5 (cont.)

Summary Statistics Resulting from
Application of Experimental Scaling Parameters

MATH I (n=155,671)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	543	481	481	553	481	496	557
s.d.	90	95	95	99	95	98	95
r(ACH,SATV)	.48						
r(ACH,SATM)	.82						

MATH II (n=54,787)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	660	555	556	629	556	545	646
s.d.	85	87	87	87	87	105	82
r(ACH,SATV)	.43						
r(ACH,SATM)	.78						

Table 5 (cont.)

Summary Statistics Resulting form
Application of Experimental Scaling Paramenters

BIOLOGY (n=23,634)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	540	498	497	541	497	514	573
s.d.	107	101	101	103	101	104	102
r(ACH,SATV)	.70						
r(ACH,SATM)	.62						

CHEMISTRY (n=29,238)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	572	525	525	581	525	525	624
s.d.	102	98	98	98	98	108	93
r(ACH,SATV)	.58						
r(ACH,SATM)	.65						

Table 5 (cont.)

Summary Statistics Resulting from
Application of Experimental Scaling Parameters

PHYSICS (n=18,415)

	Achievement Test Information						
	Current Scale	First Stg Scaling	Second Stg Scaling	Single-Stg Scaling	Emp. Based Single-Stg Scaling	SAT-V	SAT-M
Mean	594	547	547	604	547	536	653
s.d.	97	93	93	96	93	108	84
r(ACH,SATV)	.51						
r(ACH,SATM)	.64						

Table 5 (cont.)

Summary Statistics Resulting from
Application of Experimental Scaling Parameters

FRENCH (n=20,660)

	Achievement Test Information									
	Current Scale	First Stage Scaling	French Test Scaling	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling	SAT-V	SAT-M	Semester of study
Mean	530	510	530	510	510	542	513	540	581	7.3085
s.d.	102	99	102	98	98	98	98	95	93	1.3649
r(ACH,SATV)	.50									
r(ACH,SATM)	.41									
r(ACH,sem study)	.42									

GERMAN (n=2,387)

	Achievement Test Information									
	Current Scale	First Stage Scaling	French Test Scaling	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling	SAT-V	SAT-M	Semester of study
Mean	533	505	527	514	514	528	508	552	599	6.9217
s.d.	94	98	101	99	99	97	98	95	93	1.4152
r(ACH,SATV)	.36									
r(ACH,SATM)	.30									
r(ACH,sem study)	.34									

Table 5 (cont.)

Summary Statistics Resulting form
Application of Experimental Scaling Parameters

LATIN (n=2,683)

	Achievement Test Information									
	Current Scale	First Stage Scaling	French Test Scaling	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling	SAT-V	SAT-M	Semester of study
Mean	548	511	531	523	523	551	514	559	600	6.6340
s.d.	105	100	103	97	97	98	99	94	93	1.6518
r(ACH,SATV)	.53									
r(ACH,SATM)	.48									
r(ACH,sem study)	.36									

SPANISH (n=22,772)

	Achievement Test Information									
	Current Scale	First Stage Scaling	French Test Scaling	Second Stage Scaling A	Second Stage Scaling B	Single-Stage Scaling	Emp. Based Single-Stage Scaling	SAT-V	SAT-M	Semester of study
Mean	513	492	511	496	496	513	495	509	564	7.0621
s.d.	96	100	103	99	99	100	100	96	99	1.4402
r(ACH,SATV)	.38									
r(ACH,SATM)	.31									
r(ACH,sem study)	.41									

Table 6

Rank Ordering of SAT-V and SAT-M Composite Scaled Score Means and Achievement Test Scaled Score Means for Experimental Scalings

	Scaled Score Means							Rank Ordering of Scaled Score Means[1]						
Test (Code)	SAT-V and SAT-M	Current Scale	Single Stage Scaling	Second Stage Scaling A[2]	Emp. Based Single Stage Scaling	Scaling to French Test[3]	First Stage Scaling	SAT-V and SAT-M	Current Scale	Single Stage Scaling	Second Stage Scaling A[2]	Emp. Based Single Stage Scaling	Scaling to French Test	First Stage Scaling
English Comp. (EN)	1090	518	530	499	499	499	499	M2	M2	M2	M2	M2	M2	M2
Literature (LR)	1072	528	529	507	507	508	508	PH	PH	PH	PH	PH	PH	PH
Amer. Hist. (AH)	1072	528	529	496	496	497	497	LT	CH	CH	CH	CH	LT	CH
Eur. Hist (EH)	1116	547	547	521	521	522	522	GM	LT	M1	LT	EH	FR	EH
Math I (M1)	1053	543	553	481	481	481	481	CH	EH	LT	EH	LT	GM	LT
Math II (M2)	1191	660	629	556	556	555	555	FR	M1	EH	GM	FR	CH	FR
Biology (BY)	1087	540	541	497	497	498	498	EH	BY	FR	FR	GM	EH	LR
Chemistry (CH)	1149	572	581	525	525	525	525	EN	GM	BY	LR	LR	SP	GM
Physics (PH)	1189	594	604	547	547	547	547	BY	FR	EN	EN	EN	LR	EN
French (FR)	1121	530	542	510	513	530	510	SP	LR, AH	LR, AH	BY	BY	EN	BY
German (GM)	1151	533	528	514	508	527	505	LR, AH			SP, AH	SP, AH	BY	AH
Latin (LT)	1159	548	551	523	514	531	511		EN	GM			AH	SP
Spanish (SP)	1073	513	513	496	495	511	492	M1	SP	SP	M1	M1	M1	M1
Spearman Rank Order Correlation Coefficient[4]									.687	.538	.929	.886	.926	.846

[1]Orderings are high to low.

[2]Results of the second stage scaling A and B procedures were quite similar, thus only second stage scaling A results are presented here.

Table 7

Summary Statistics for Achievement Test Pairs Resulting from Application of Selected Experimental Scaling Procedures

ACH Test Pair	N	Current Scale		First Stage Scaling		Single-Stage Scaling		French Test Scaling		SAT-V		SAT-M		Correlation of Achievement Test Scores with SAT-V and SAT-M Scores				
		M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	r_{xy}	r_{xv}	r_{xm}	r_{yv}	r_{ym}
EN (X)	5,180	539	102	520	102	552	107	520	102	531	105	512	99	.74	.78	.56	.83	.53
LR (Y)		532	105	512	102	533	105	512	102									
EN (X)	7,551	557	95	538	95	570	100	538	95	534	100	552	94	.50	.75	.49	.47	.35
FR (Y)		528	100	508	97	540	96	528	100									
EN (X)	959	564	100	545	100	578	105	545	100	552	105	572	98	.57	.77	.50	.54	.49
LT (Y)		526	98	490	94	530	92	509	97									
EN (X)	7,551	509	99	490	99	520	104	490	99	505	101	539	102	.61	.76	.56	.73	.63
BY (Y)		509	100	469	95	511	97	469	95									
EN (X)	7,488	519	100	500	100	531	105	500	100	522	101	610	95	.51	.76	.54	.58	.66
CH (Y)		547	100	502	96	557	97	502	96									
EN (X)	3,487	521	103	502	103	533	108	502	103	528	105	630	90	.50	.77	.54	.57	.63
PH (Y)		578	95	532	91	588	93	532	91									
EN (X)	16,070	501	99	482	99	512	104	482	99	502	101	534	100	.57	.76	.47	.73	.44
AH (Y)		497	94	466	96	497	97	466	96									
EN (X)	1,832	550	101	531	101	563	106	531	101	567	102	551	99	.56	.77	.50	.70	.45
EH (Y)		544	101	519	106	544	108	519	106									
LR (X)	1,034	521	108	501	105	522	108	501	105	525	112	484	103	.68	.84	.52	.76	.50
AH (Y)		503	97	472	99	503	100	472	99									
FR (X)	416	510	96	490	93	523	92	510	96	547	98	507	98	.45	.57	.45	.68	.48
AH (Y)		531	86	500	88	532	89	500	88									
FR (X)	4,742	514	99	494	96	527	95	514	99	518	99	548	86	.37	.43	.36	.40	.77
M1 (Y)		538	88	476	93	548	96	476	93									
LT (X)	553	508	96	472	92	513	90	491	95	534	100	559	89	.49	.51	.48	.45	.79
M1 (Y)		554	90	493	95	566	99	493	95									
BY (X)	202	562	111	519	105	563	107	519	105	512	118	586	113	.84	.71	.74	.55	.73
CH (Y)		572	111	525	107	581	107	525	107									
BY (X)	449	498	106	459	100	501	102	459	100	514	109	494	113	.73	.76	.67	.77	.57
AH (Y)		506	97	475	99	506	100	475	99									
BY (X)	5,700	497	99	458	94	500	96	458	94	491	99	534	94	.60	.71	.63	.52	.81
M1 (Y)		525	92	462	97	534	101	462	97									
CH (X)	5,245	523	95	478	91	533	92	478	91	498	100	584	92	.64	.55	.62	.46	.81
M1 (Y)		582	93	522	98	596	102	522	98									

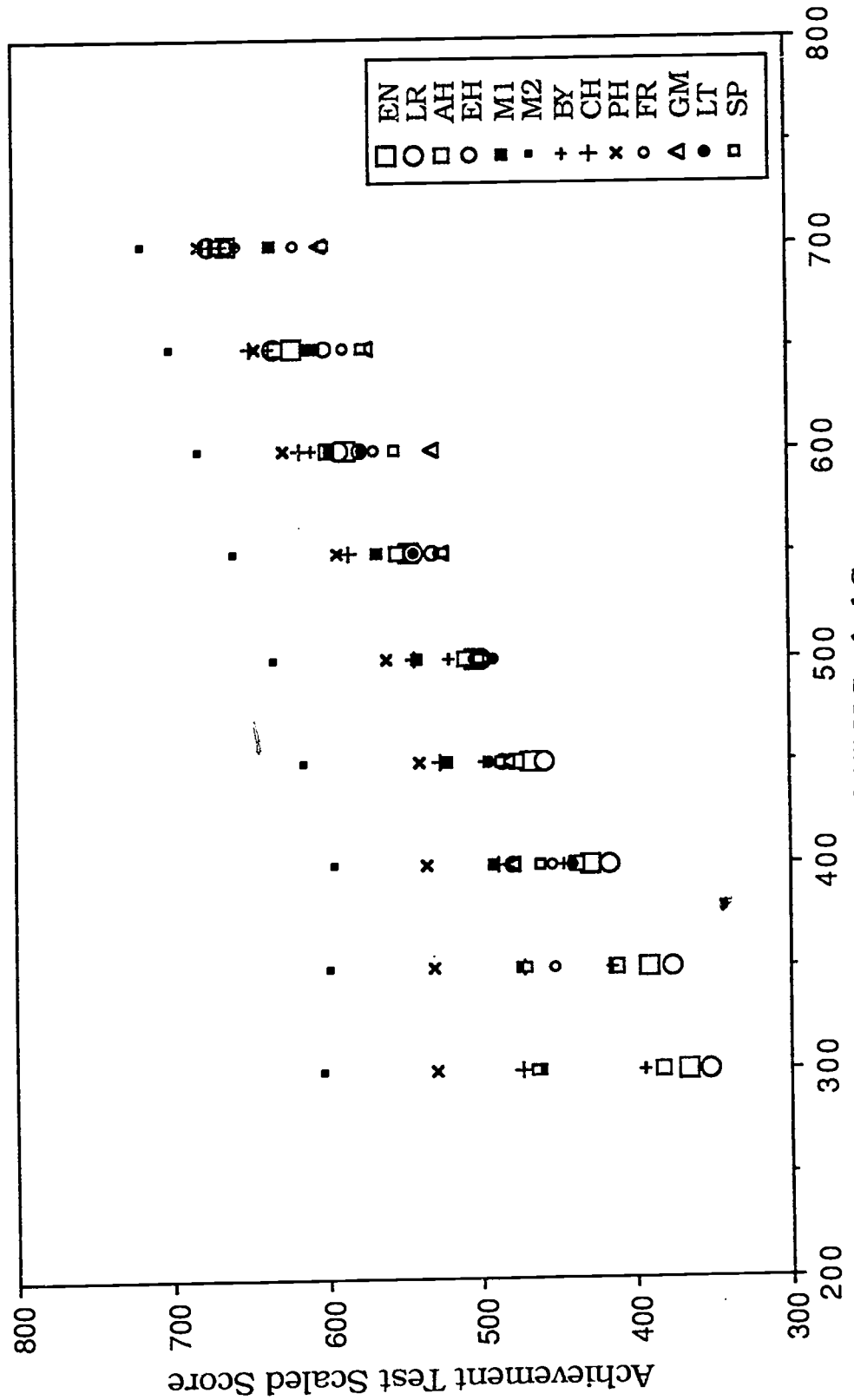


Figure 1a: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

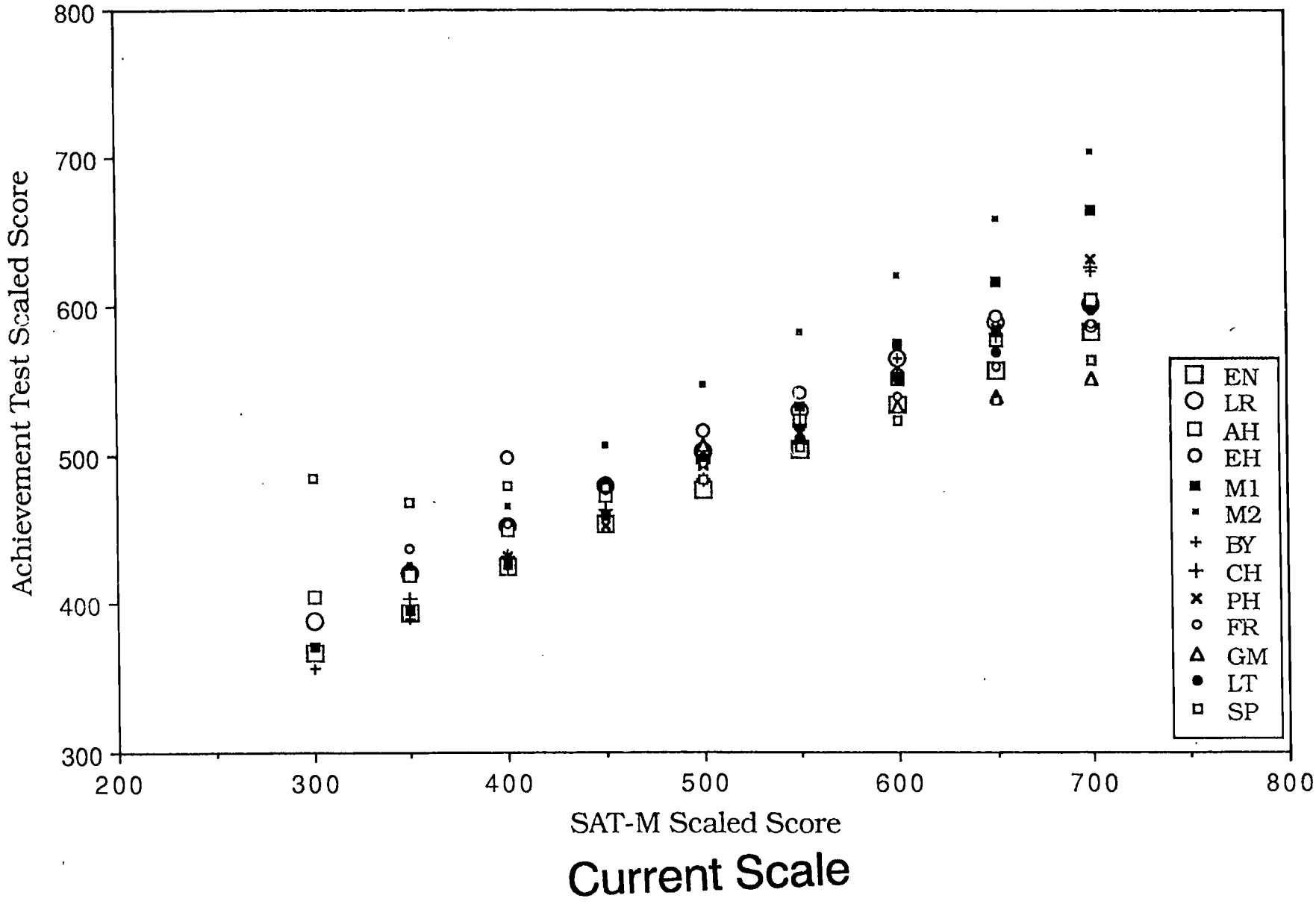


Figure 1b: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

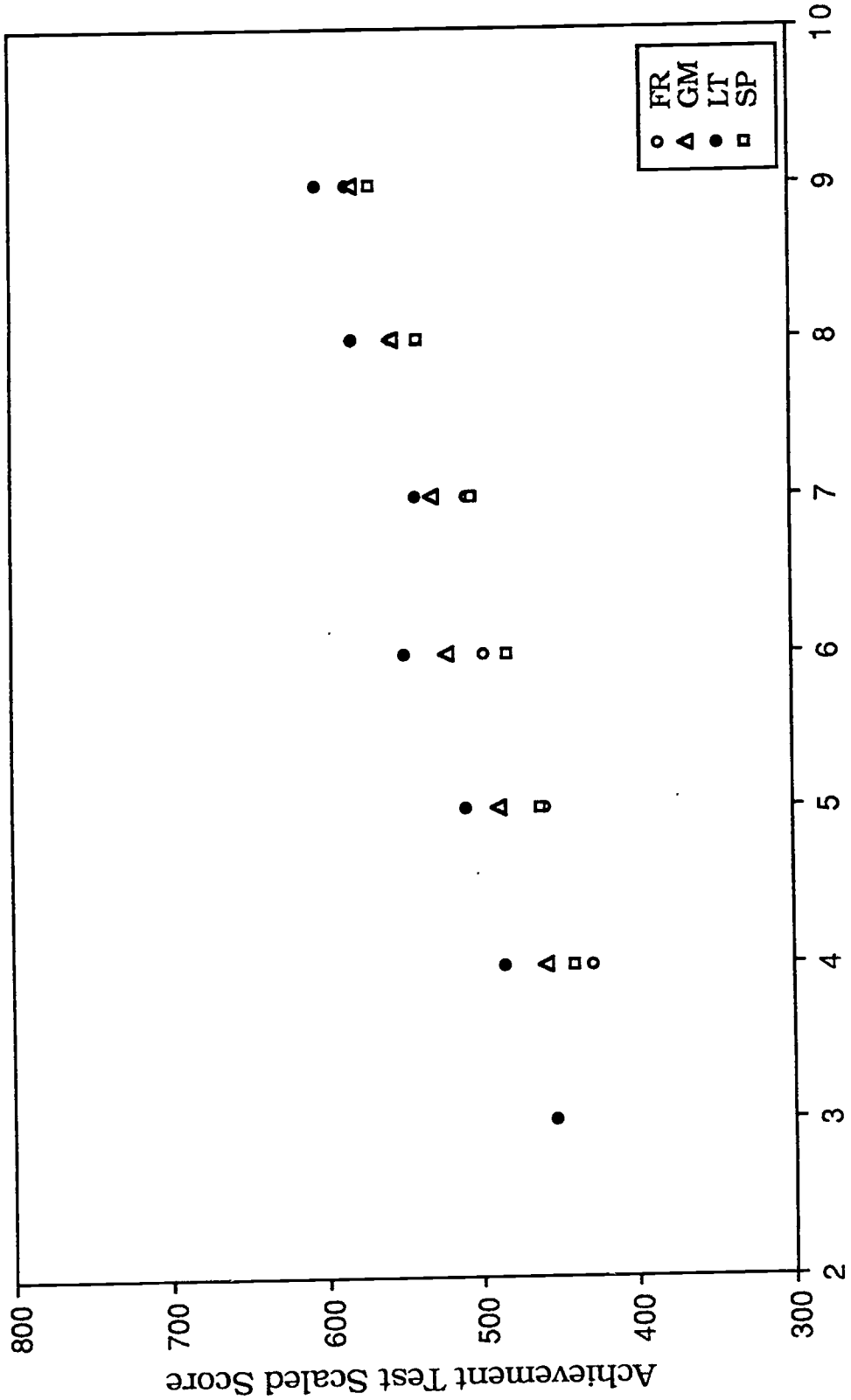


Figure 1c: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

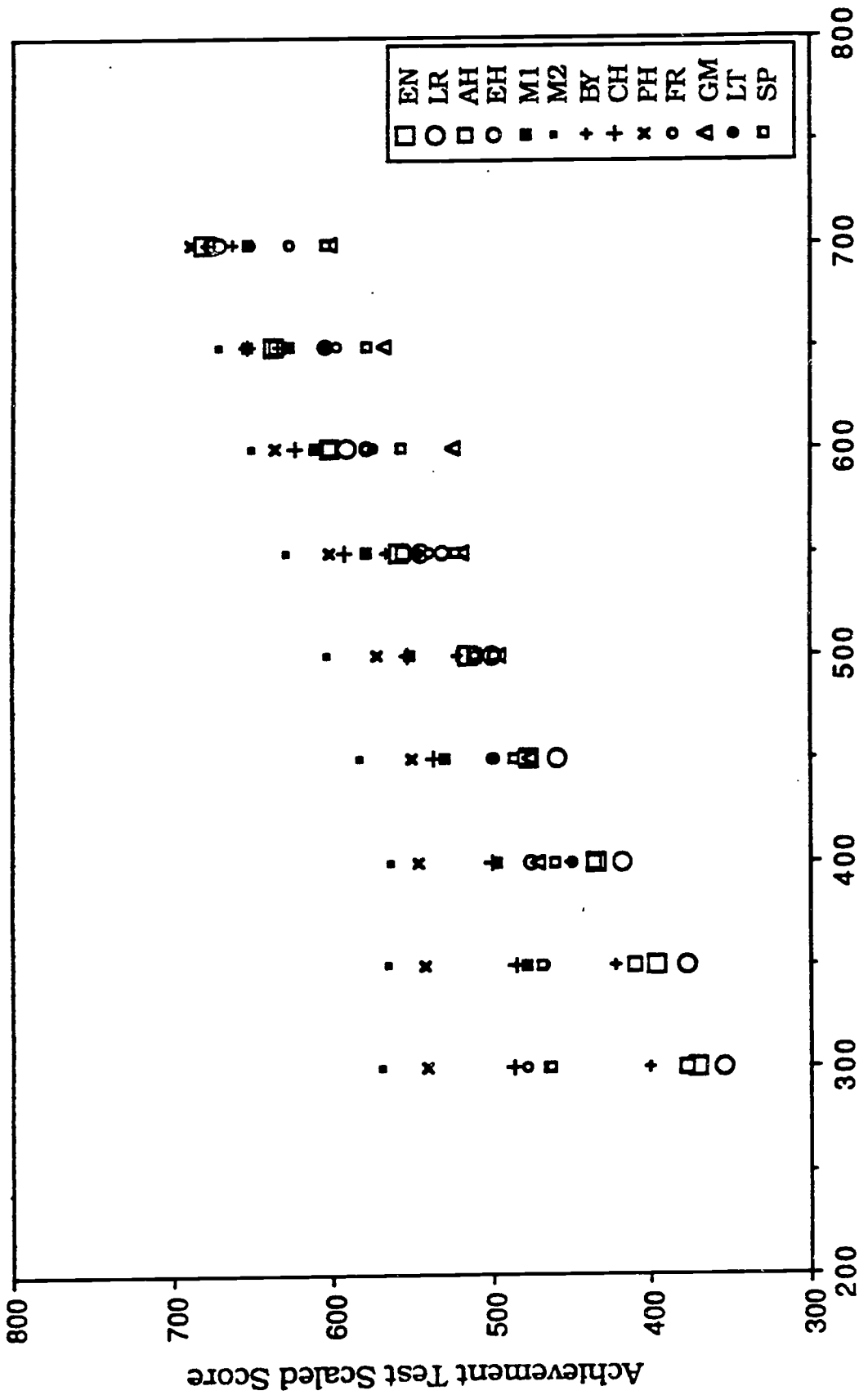


Figure 1d: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

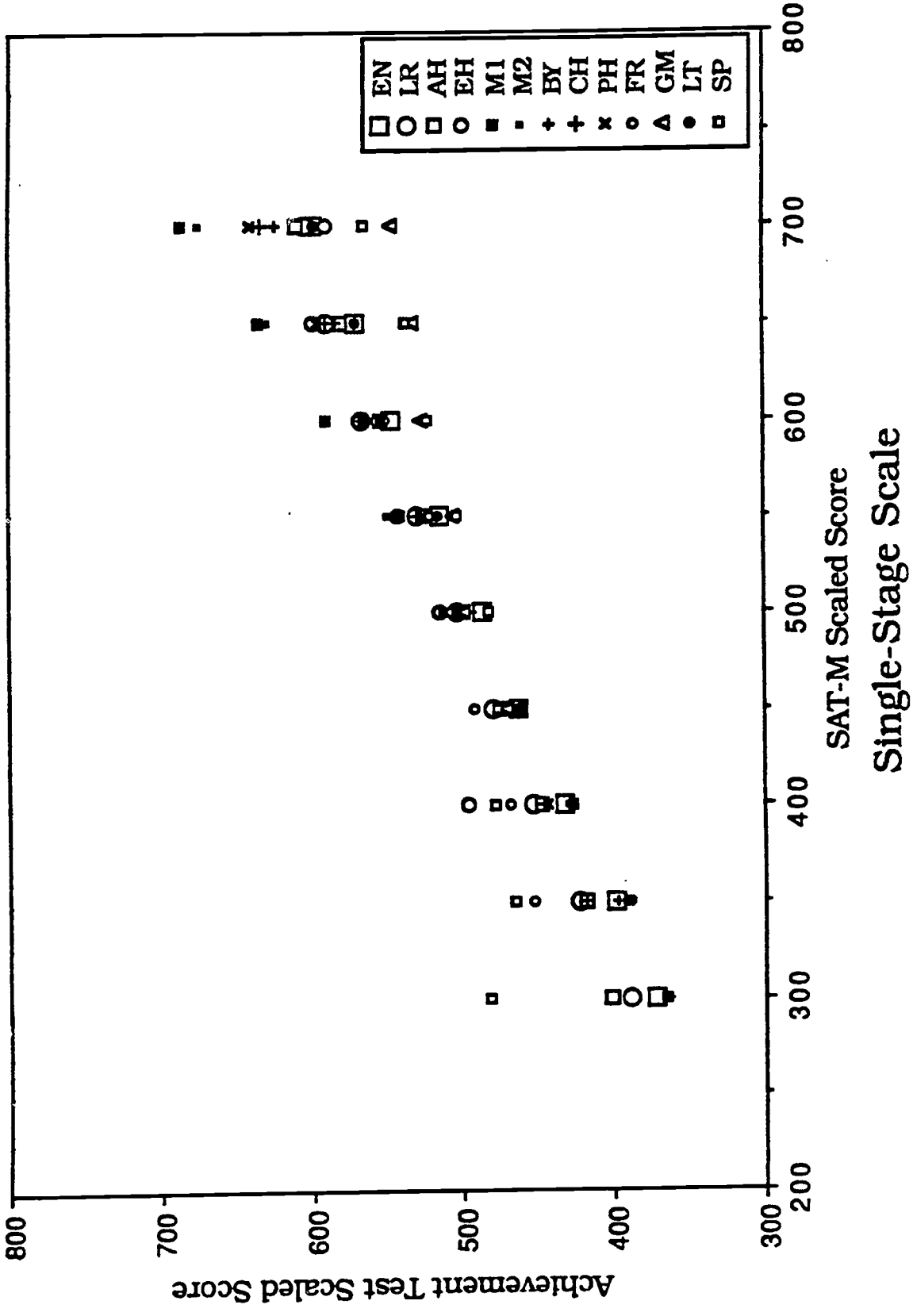


Figure 1e: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

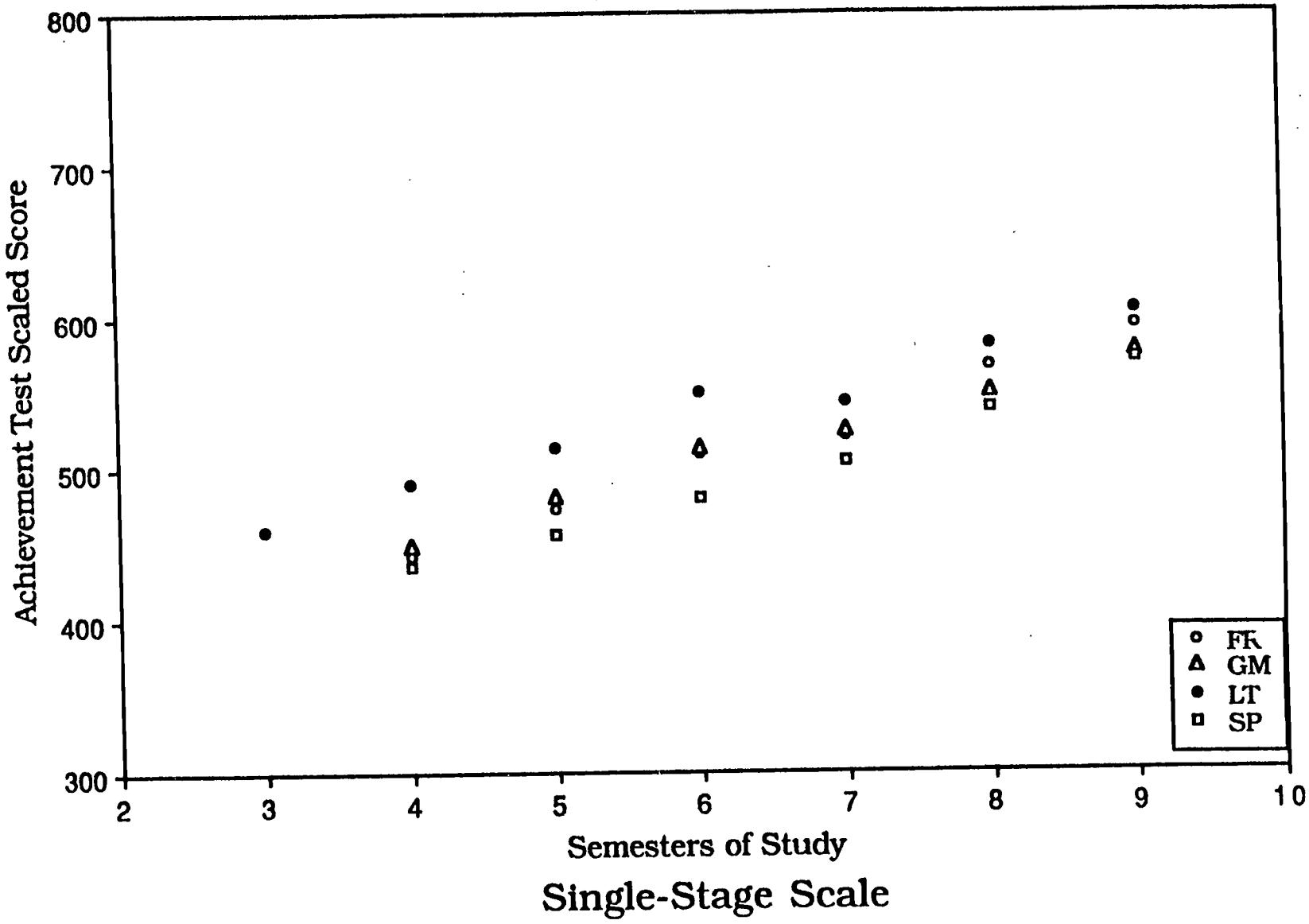


Figure 1f: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

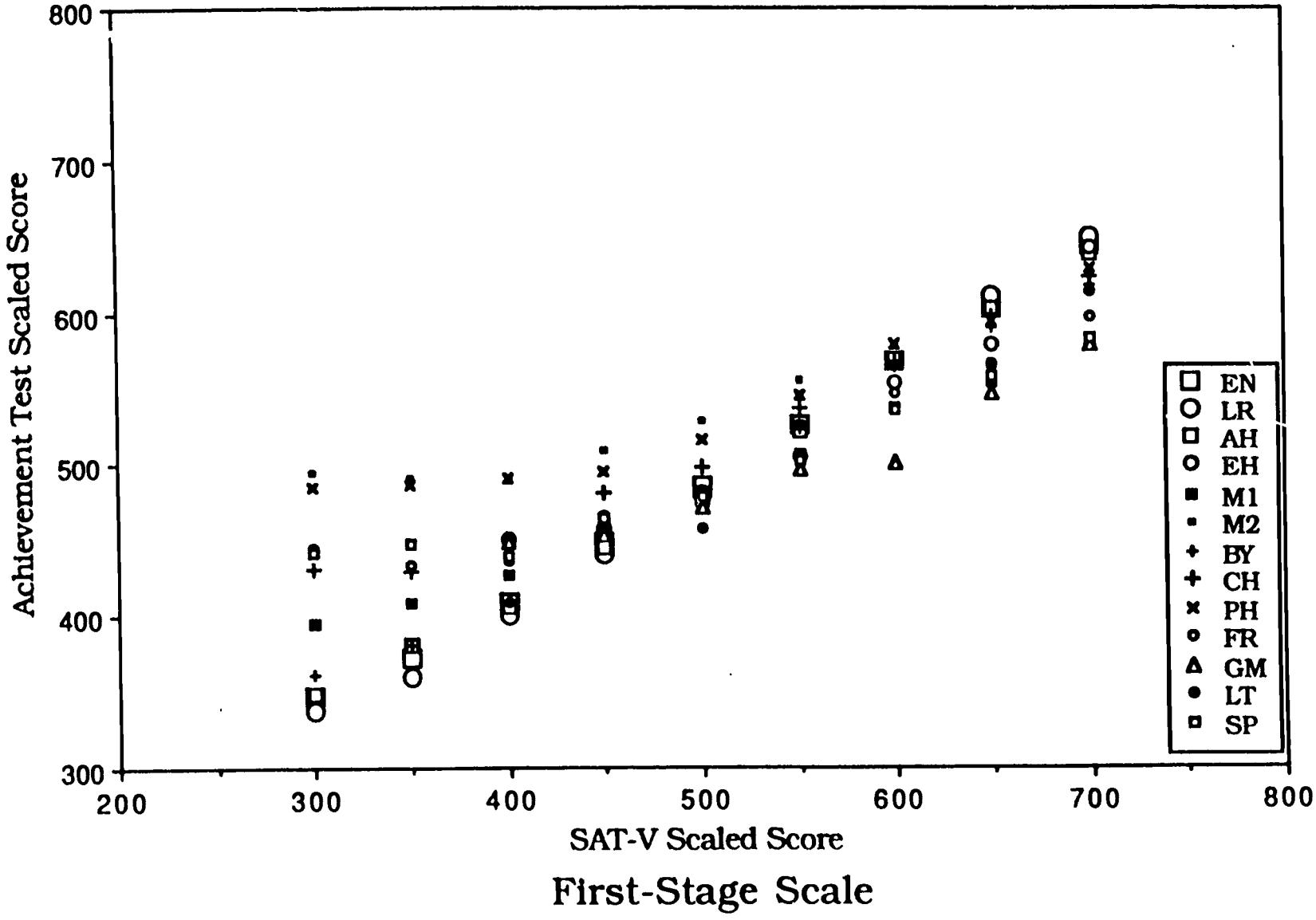
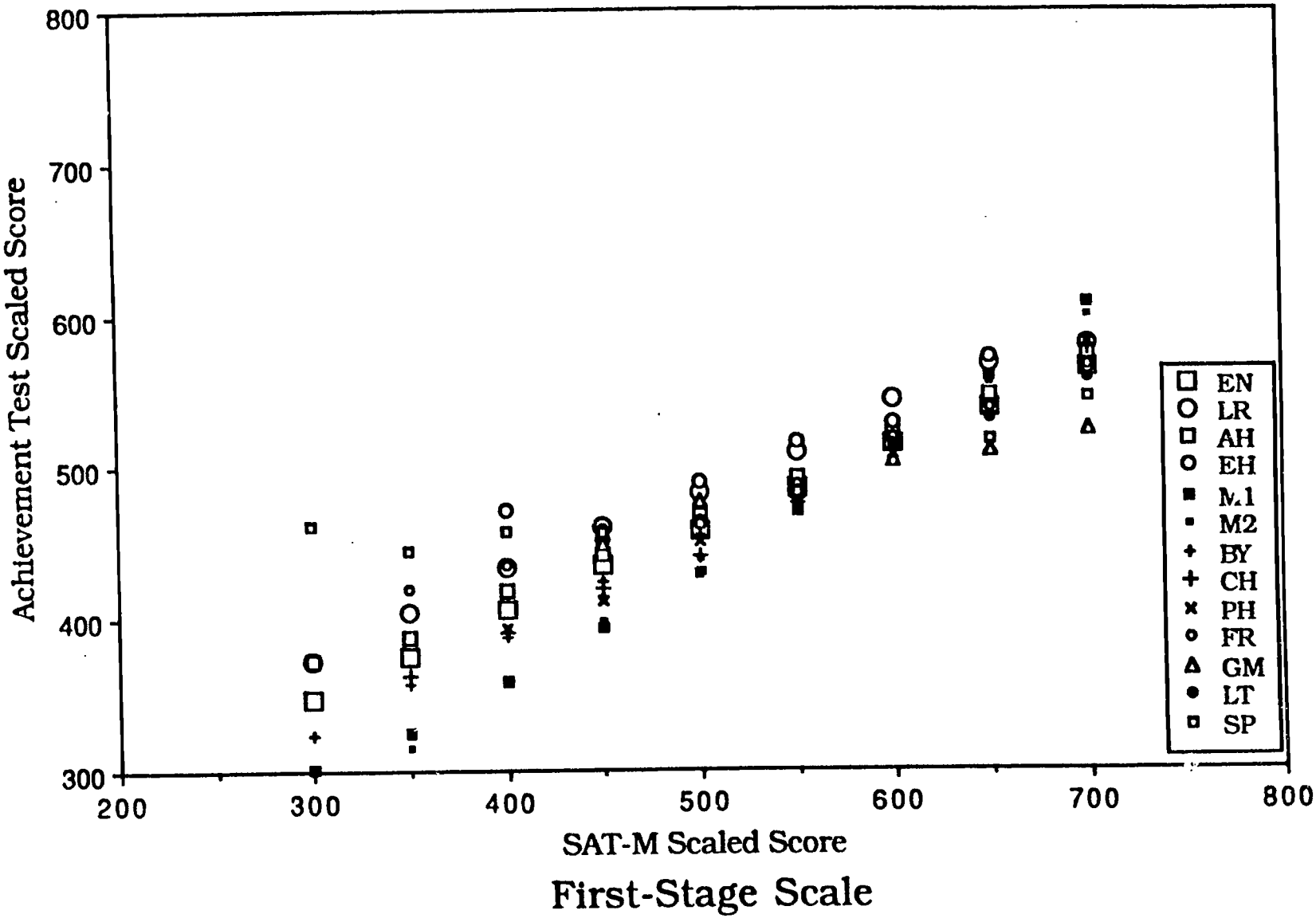


Figure 1g: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.



First-Stage Scale

Figure 1h: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

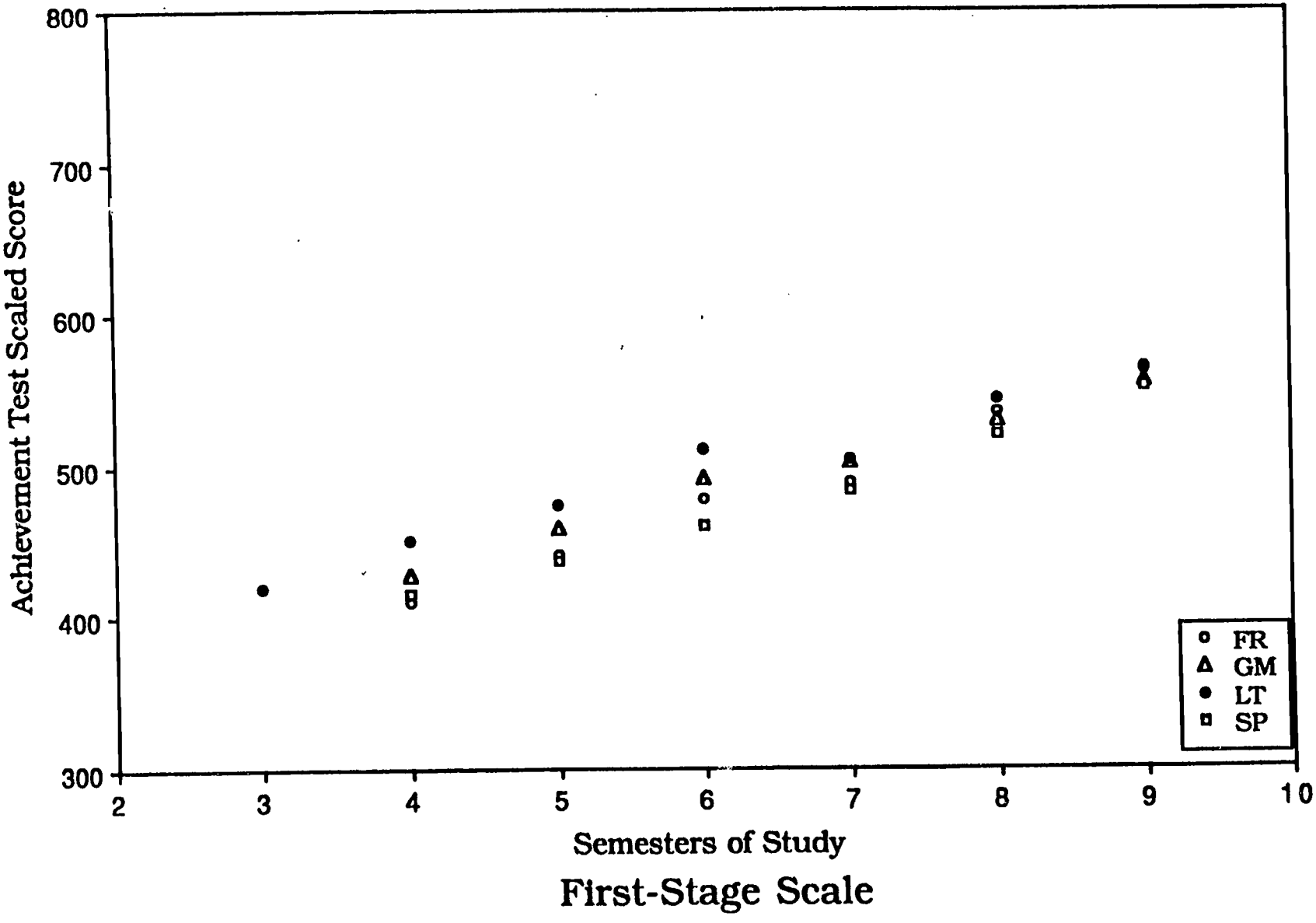


Figure 11: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

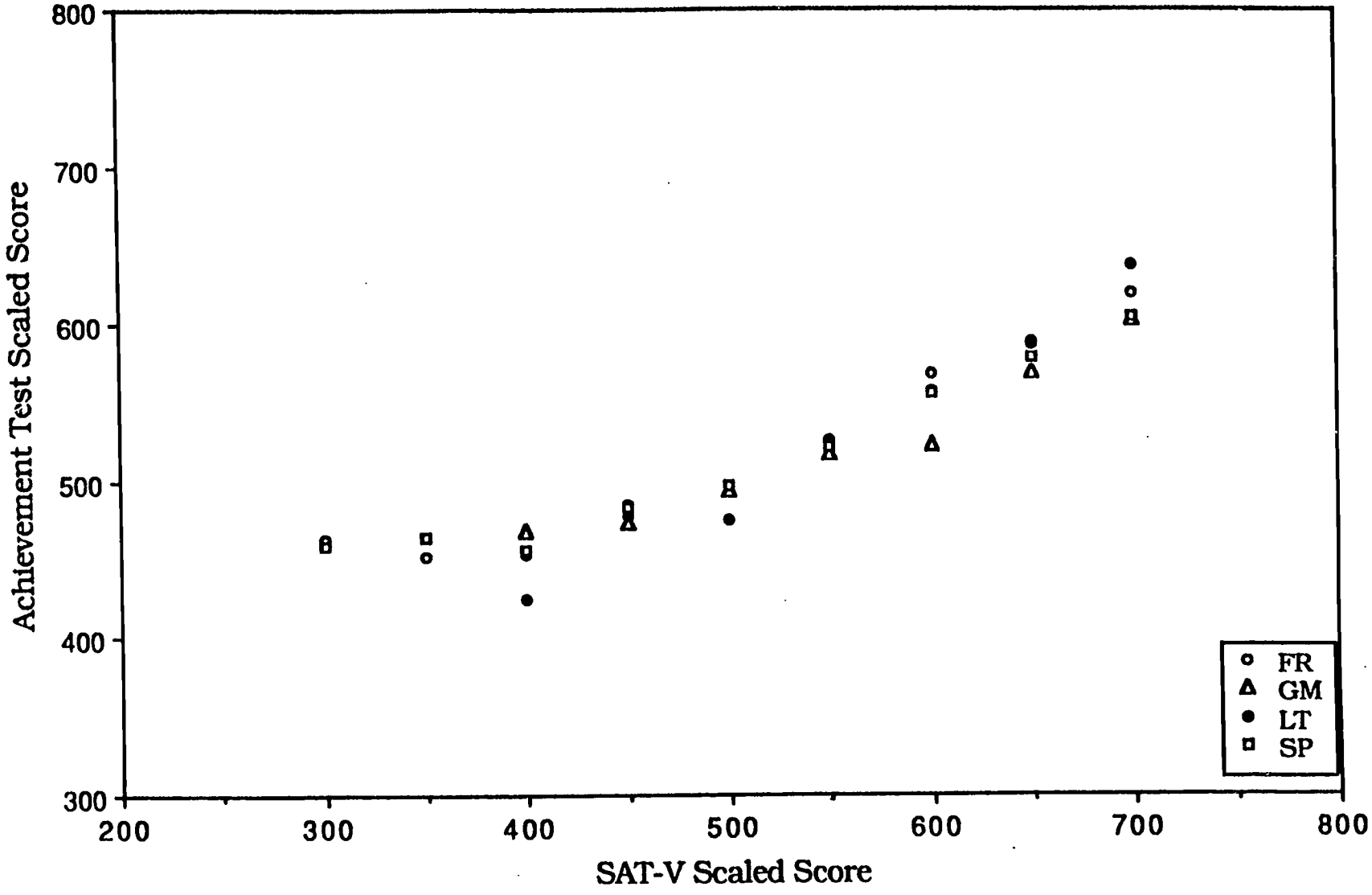


Figure 1j: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

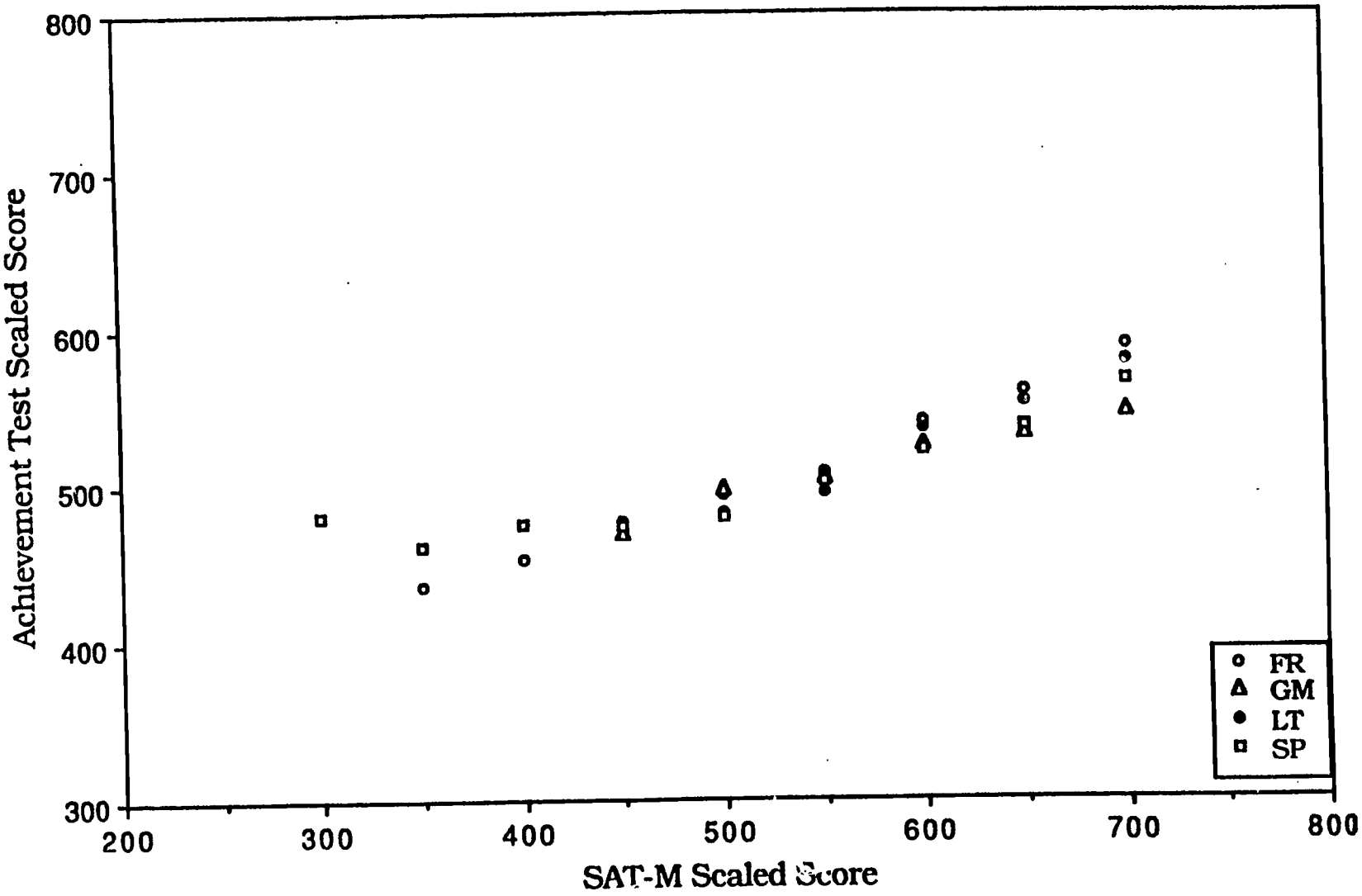


Figure 1k: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

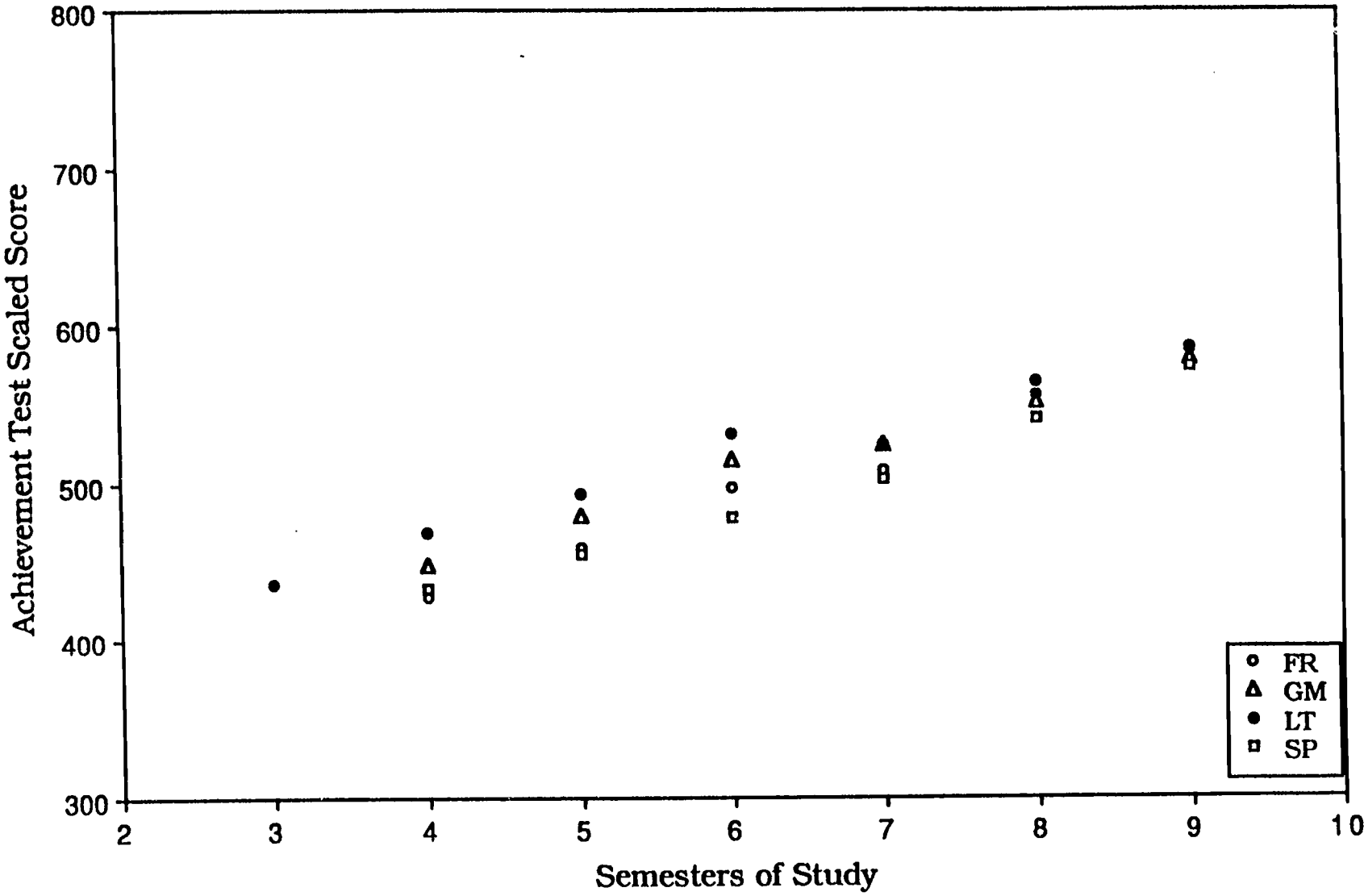
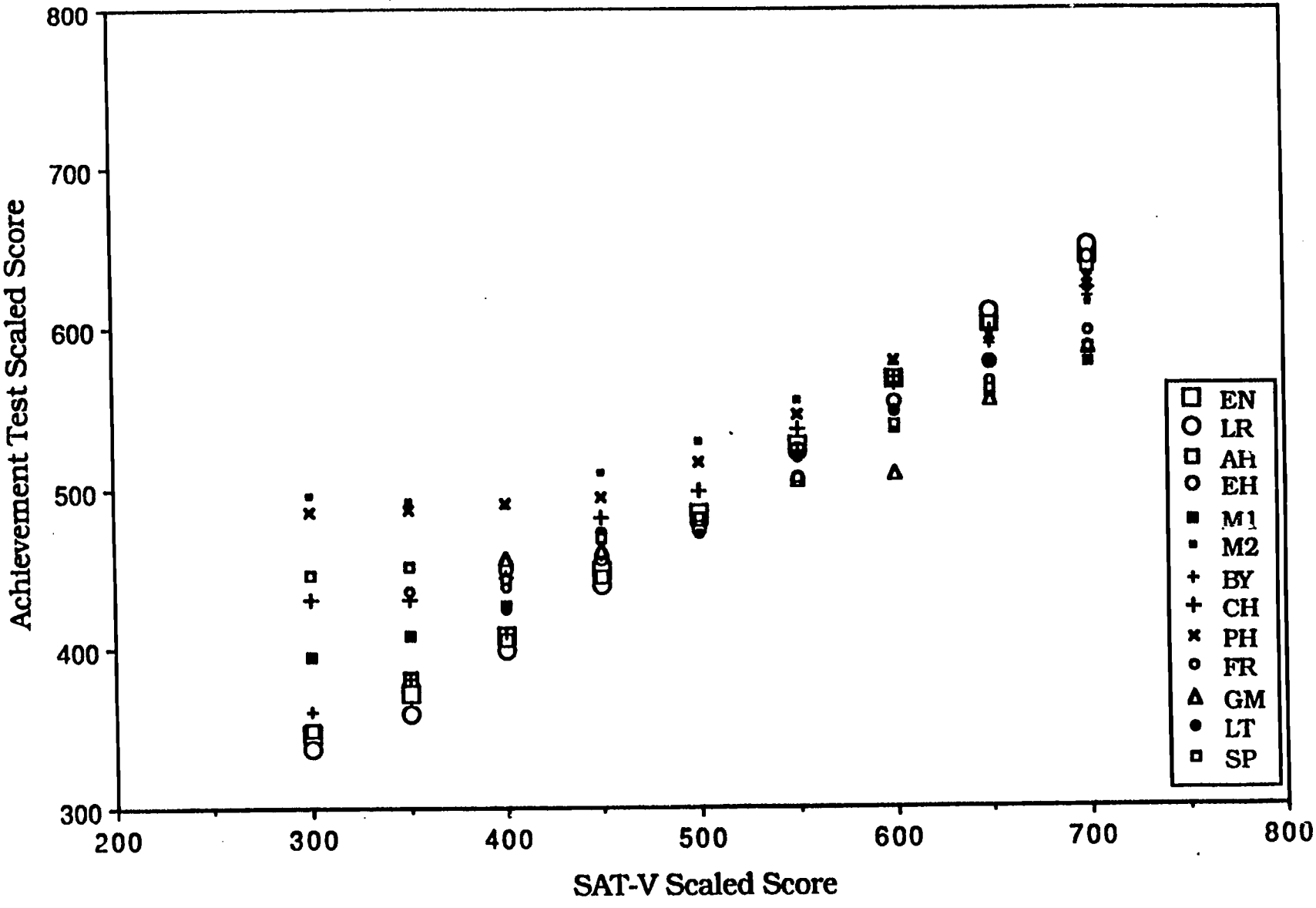
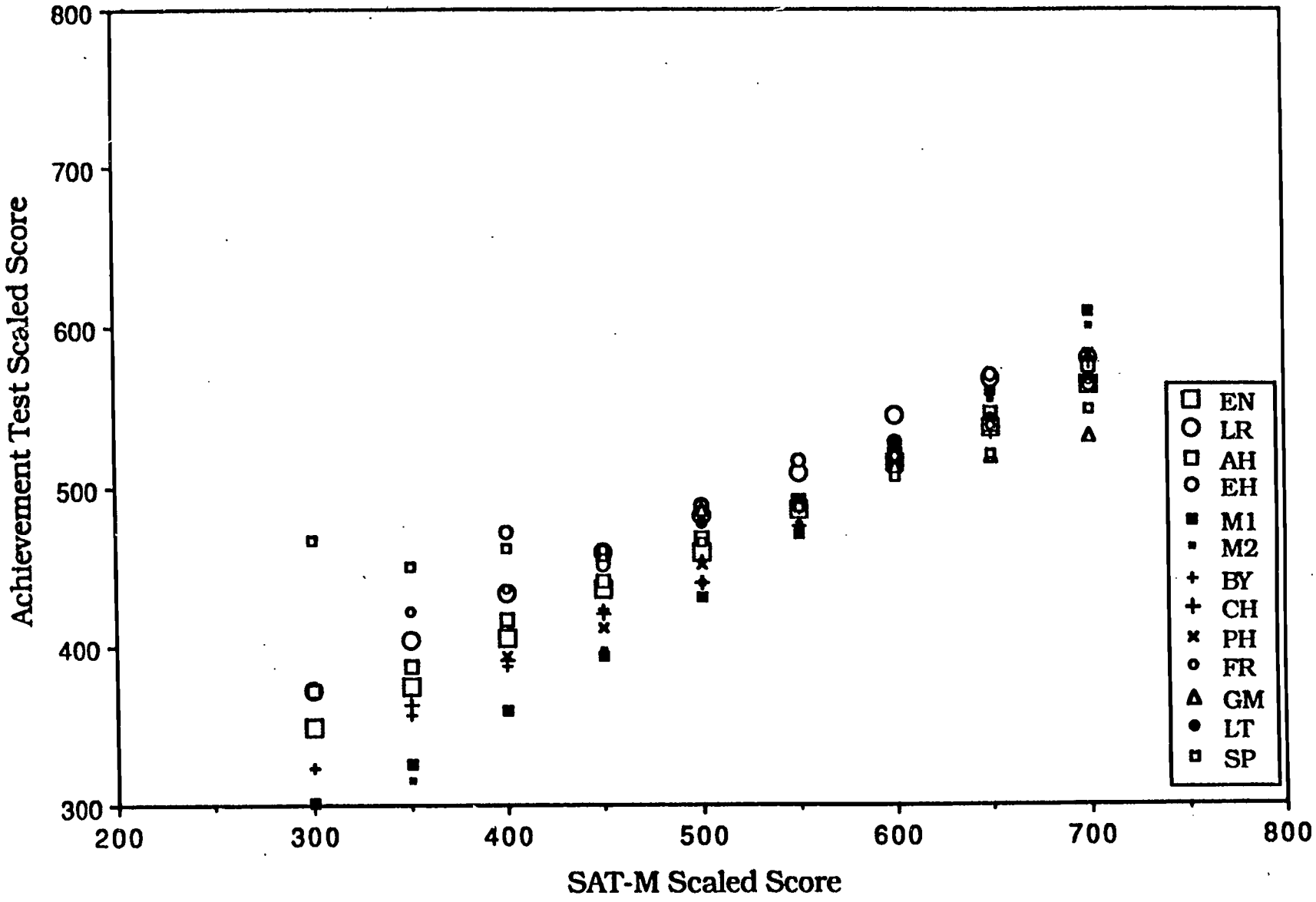


Figure 11: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.



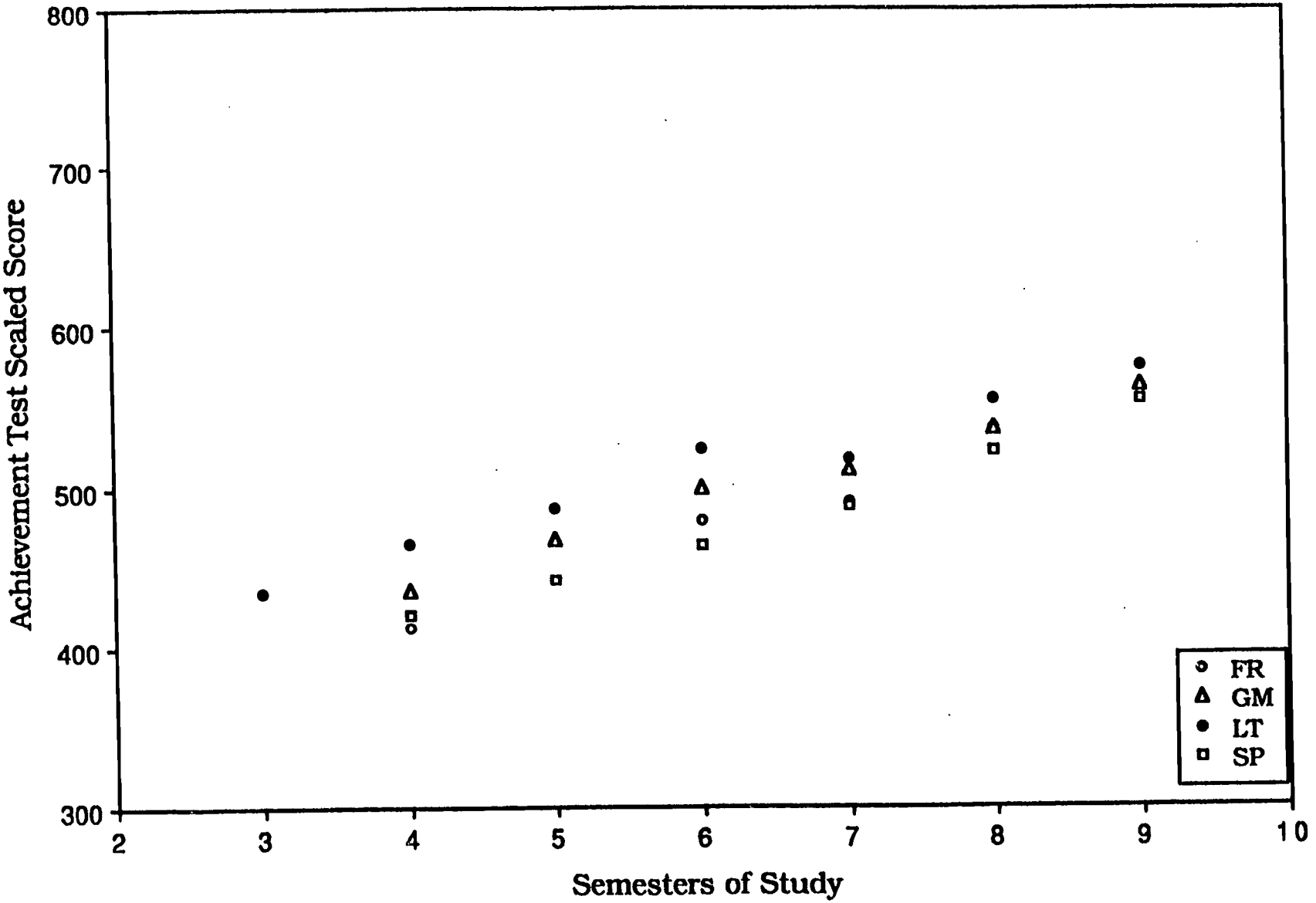
Second-Stage Scale

Figure 1m: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.



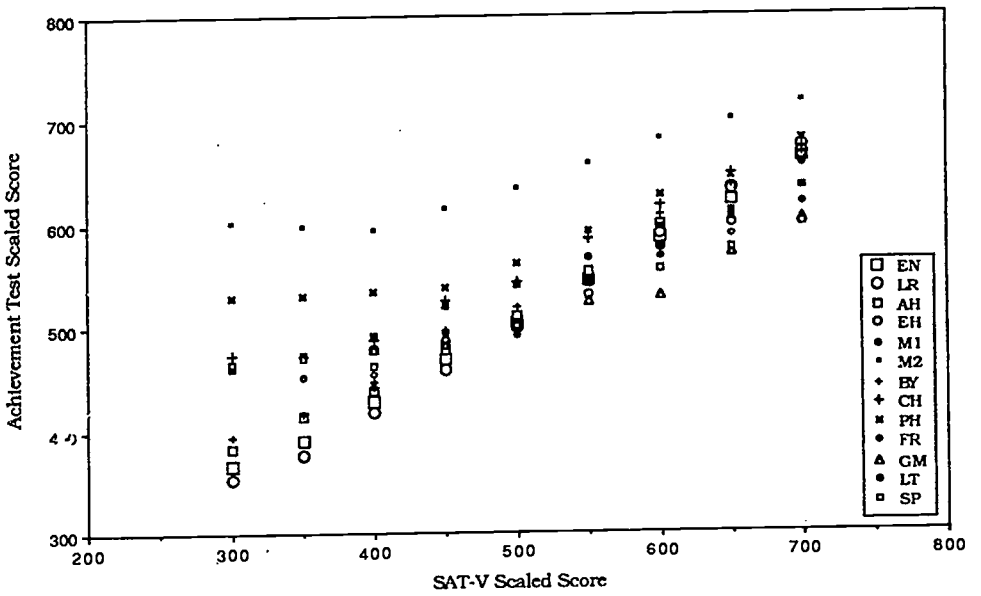
Second-Stage Scale

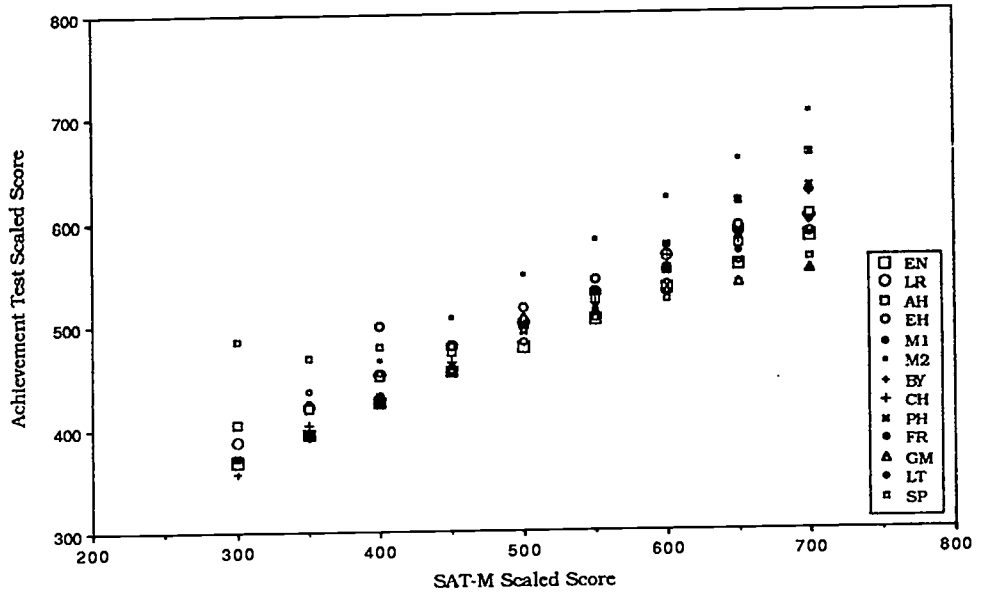
Figure 1n: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

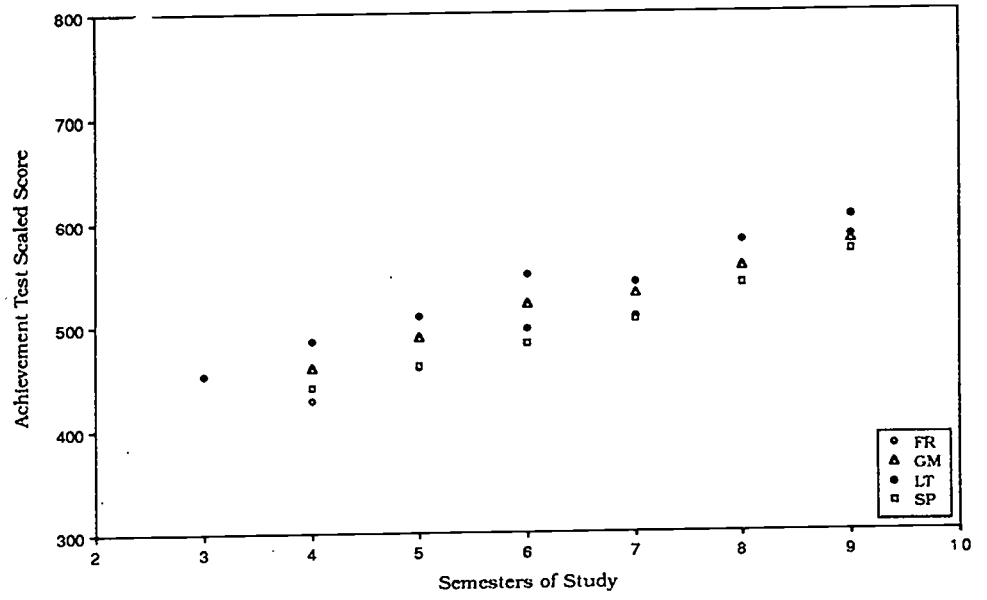


Second-Stage Scale

Figure 1o: Achievement Test scaled score means resulting from experimental scalings, conditional on SAT-V and SAT-M scaled scores and semesters of foreign language study.

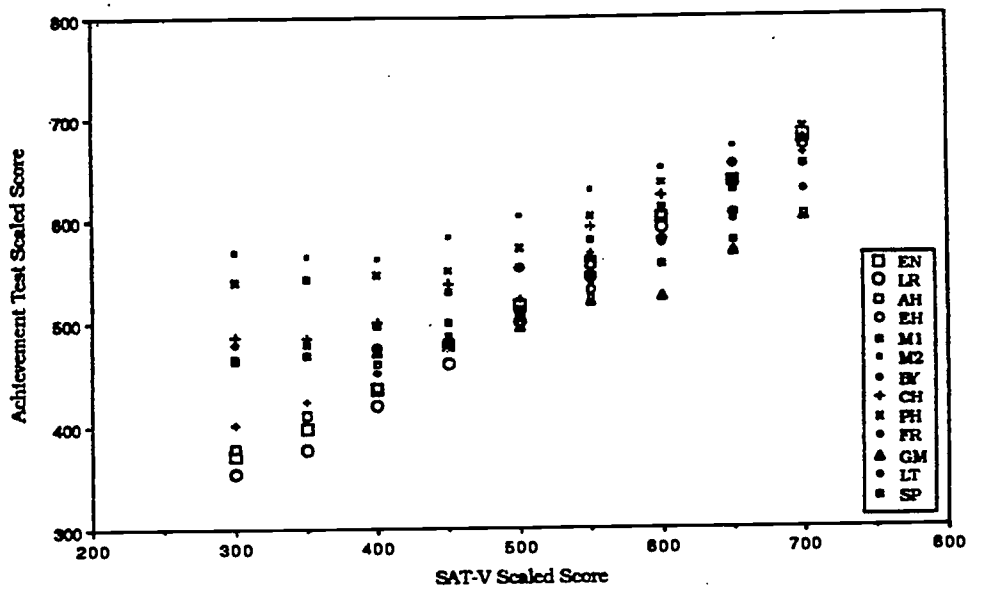


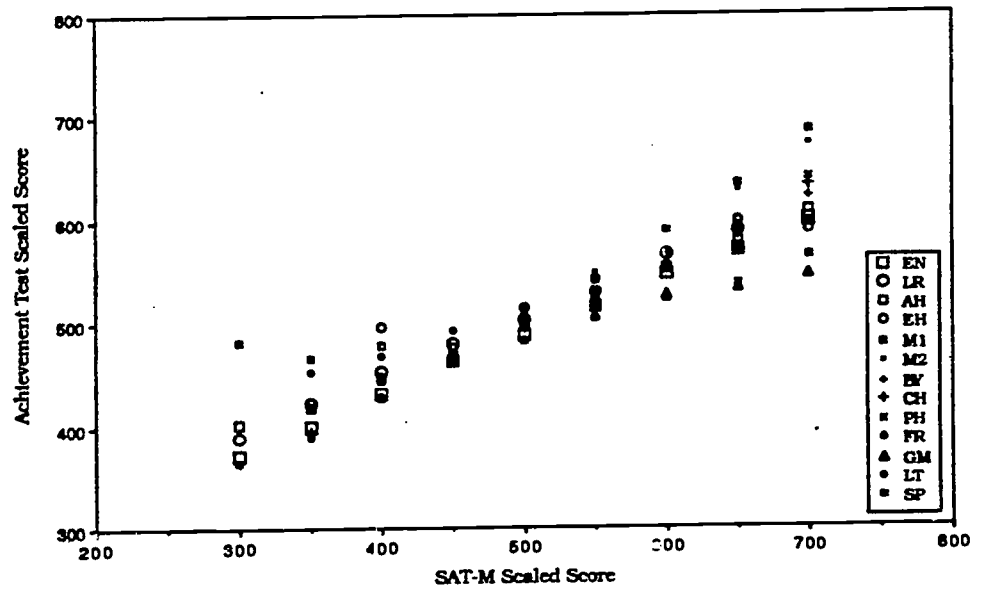


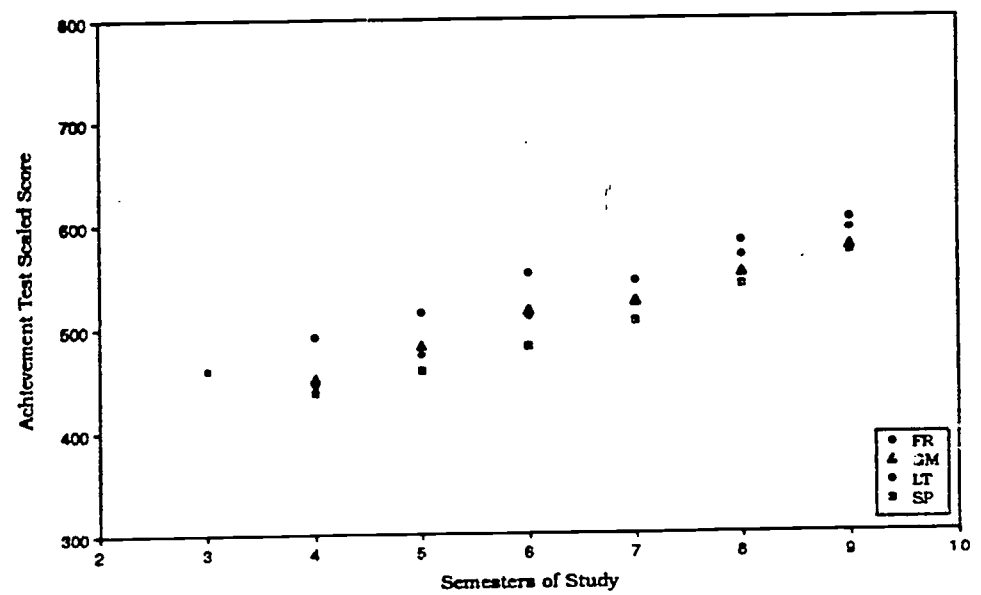


Current Scale

Figure 2a: Comparison of Achievement Test conditional scaled score means for three scaling covariates.

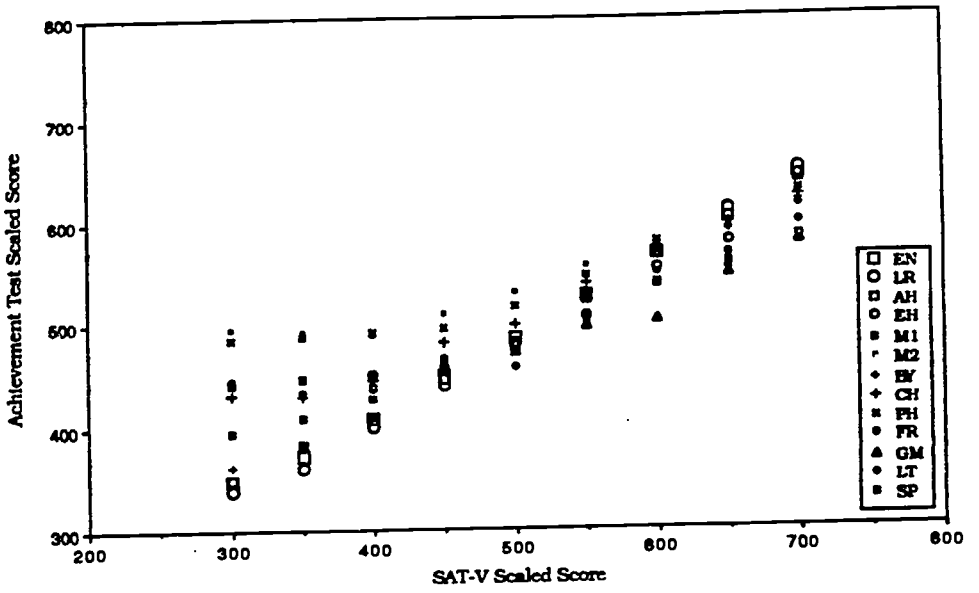


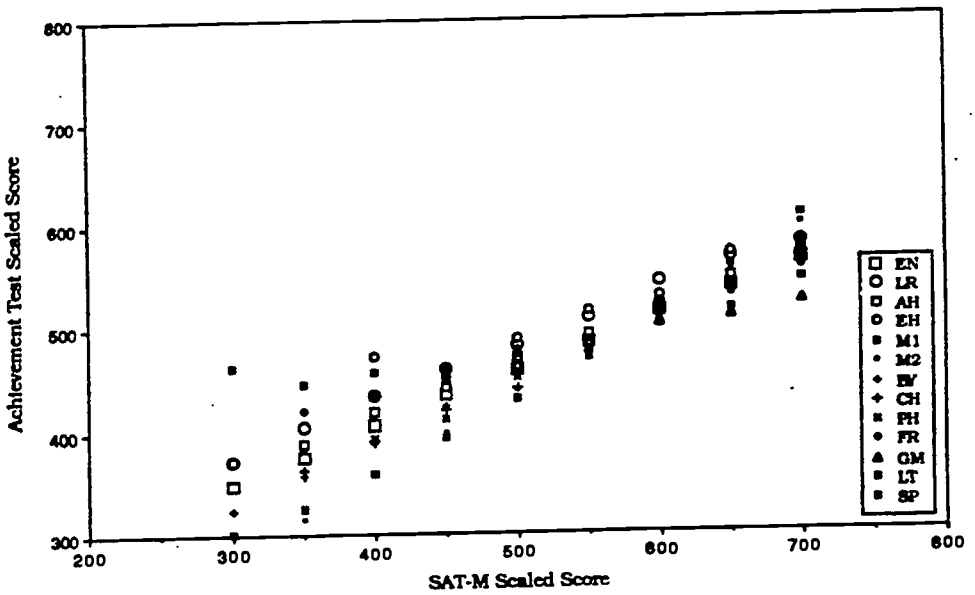


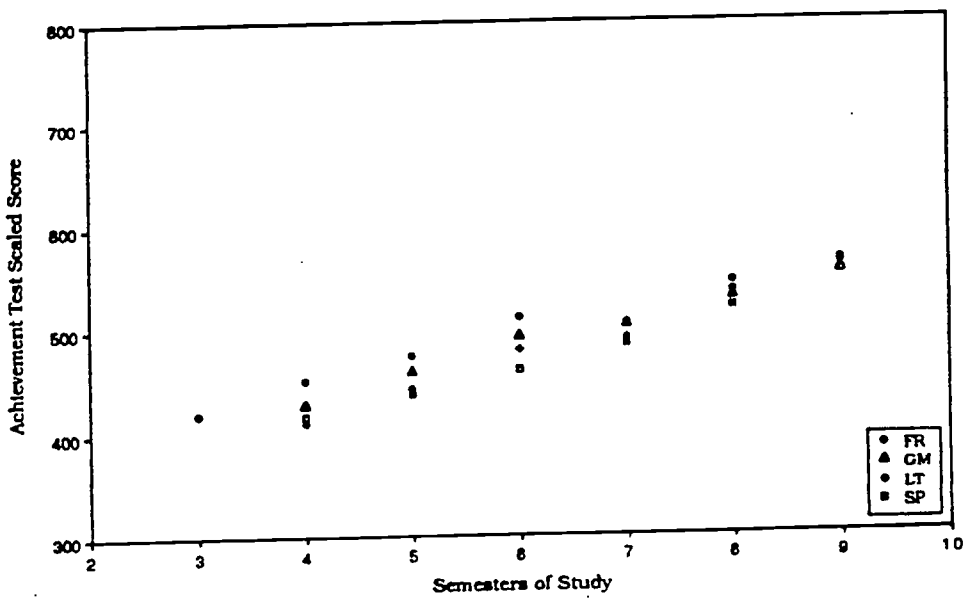


Single-Stage Scale

Figure 2b: Comparison of Achievement Test conditional scaled score means for three scaling covariates.

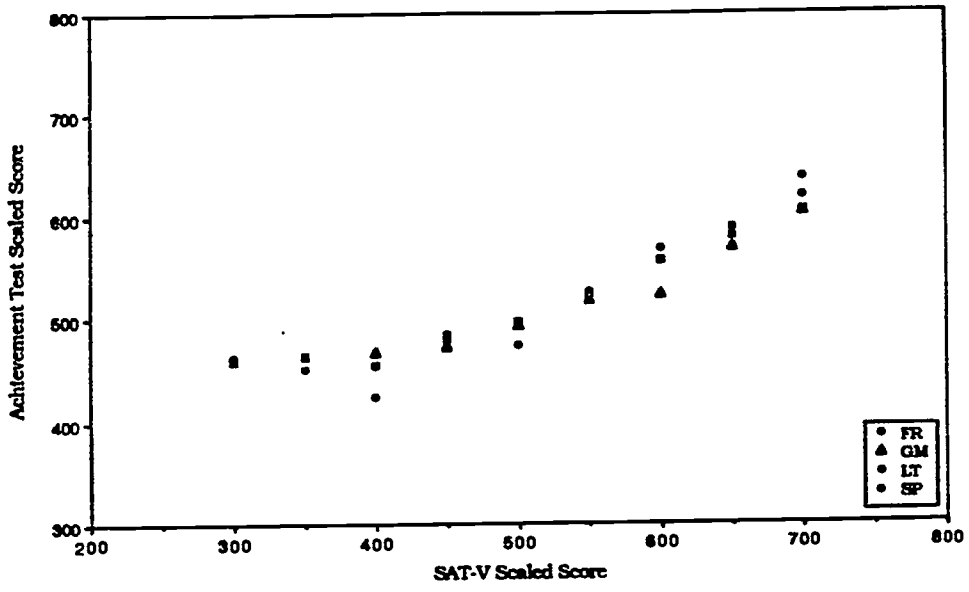


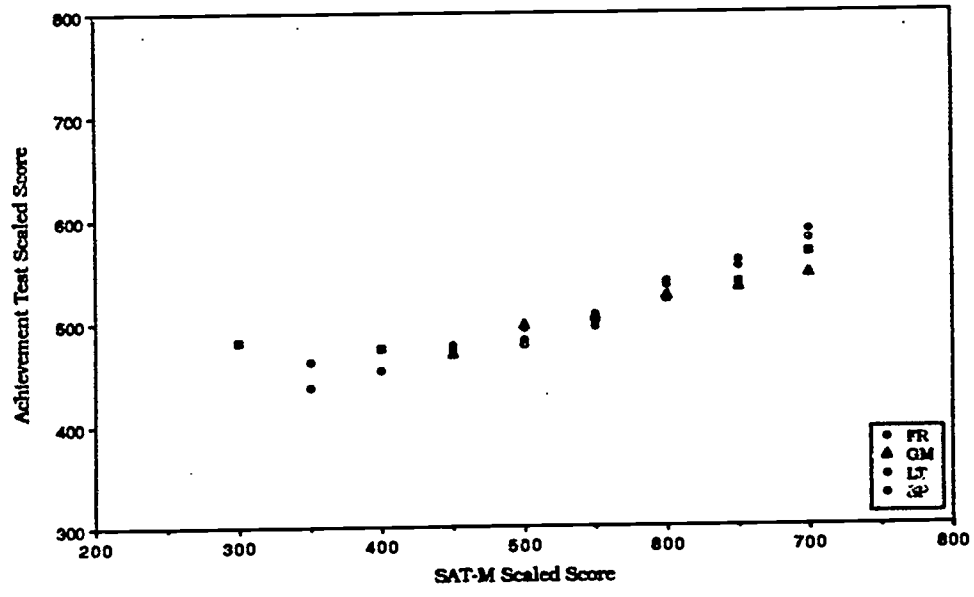


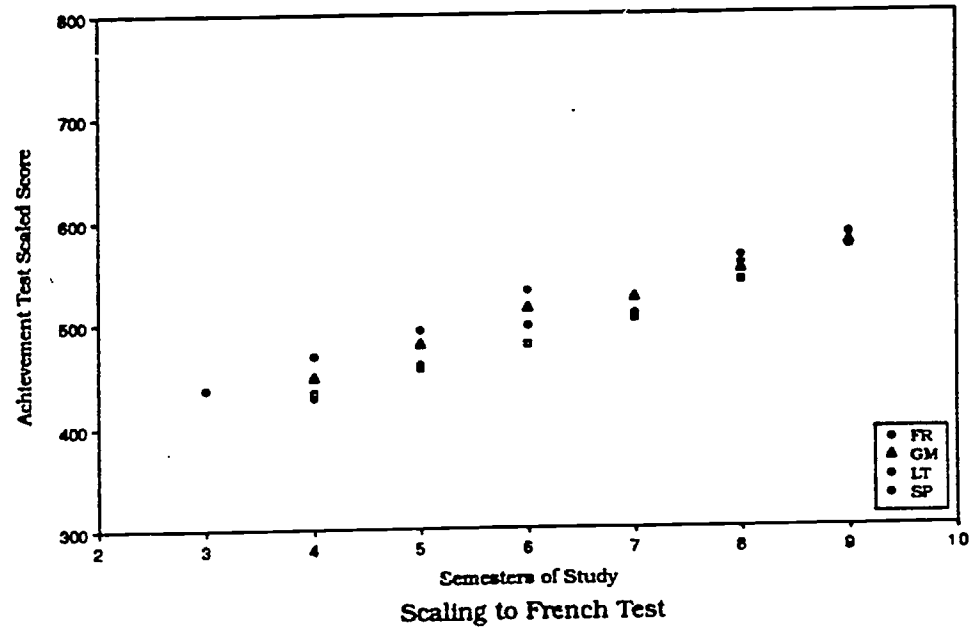


First-Stage Scale

Figure 2c: Comparison of Achievement Test conditional scaled score means for three scaling covariates.

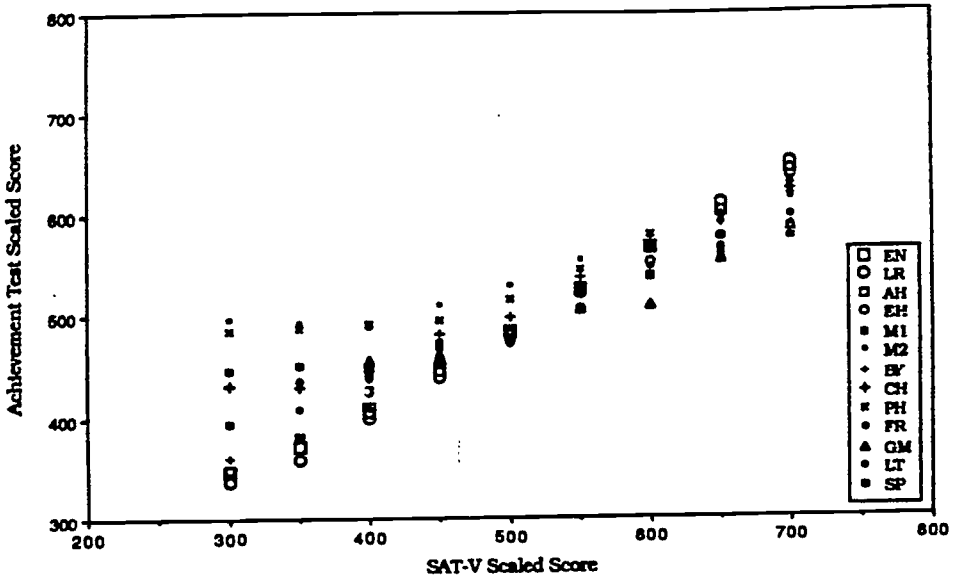


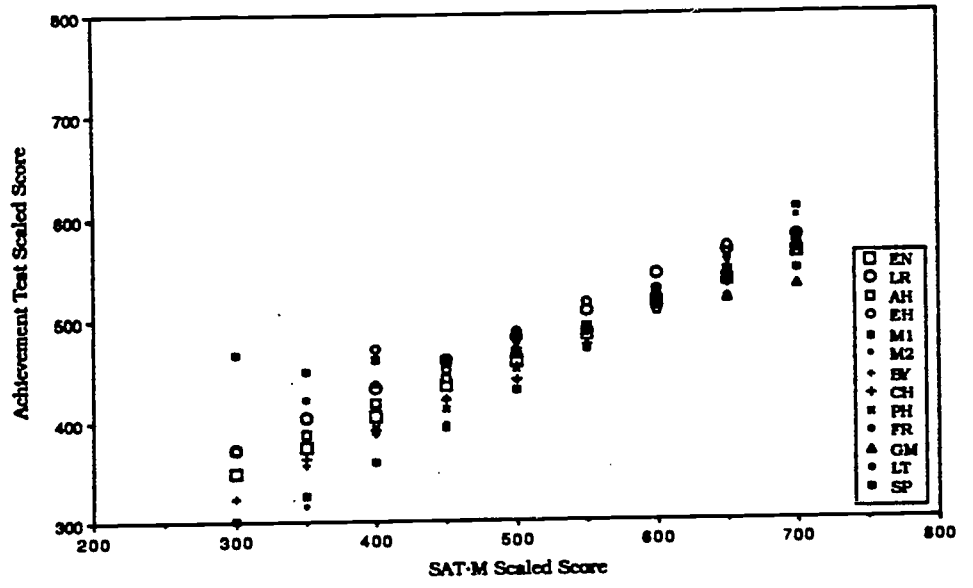




Scaling to French Test

Figure 2d: Comparison of Achievement Test conditional scaled score means for three scaling covariates.





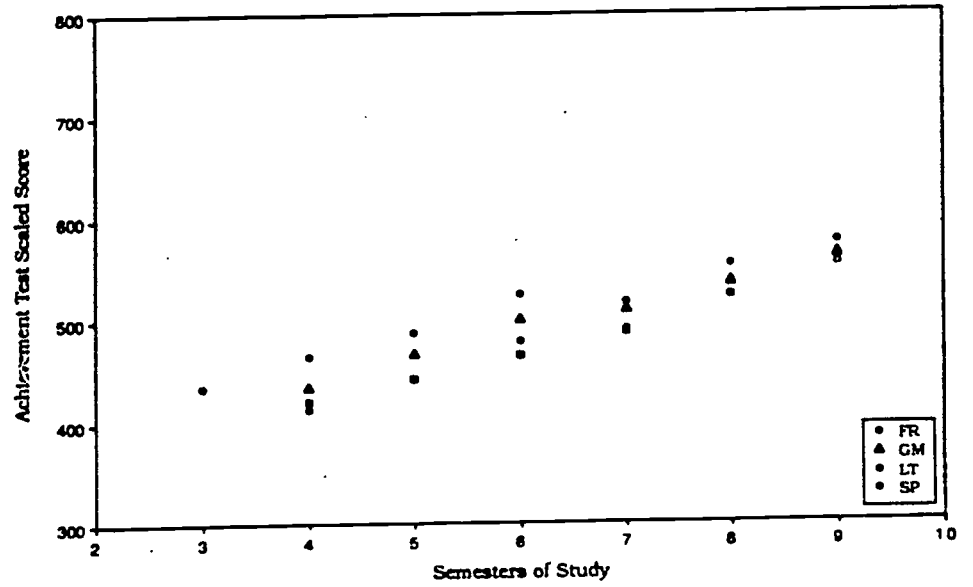


Figure 2e: Comparison of Achievement Test conditional scaled score means for three scaling covariates.

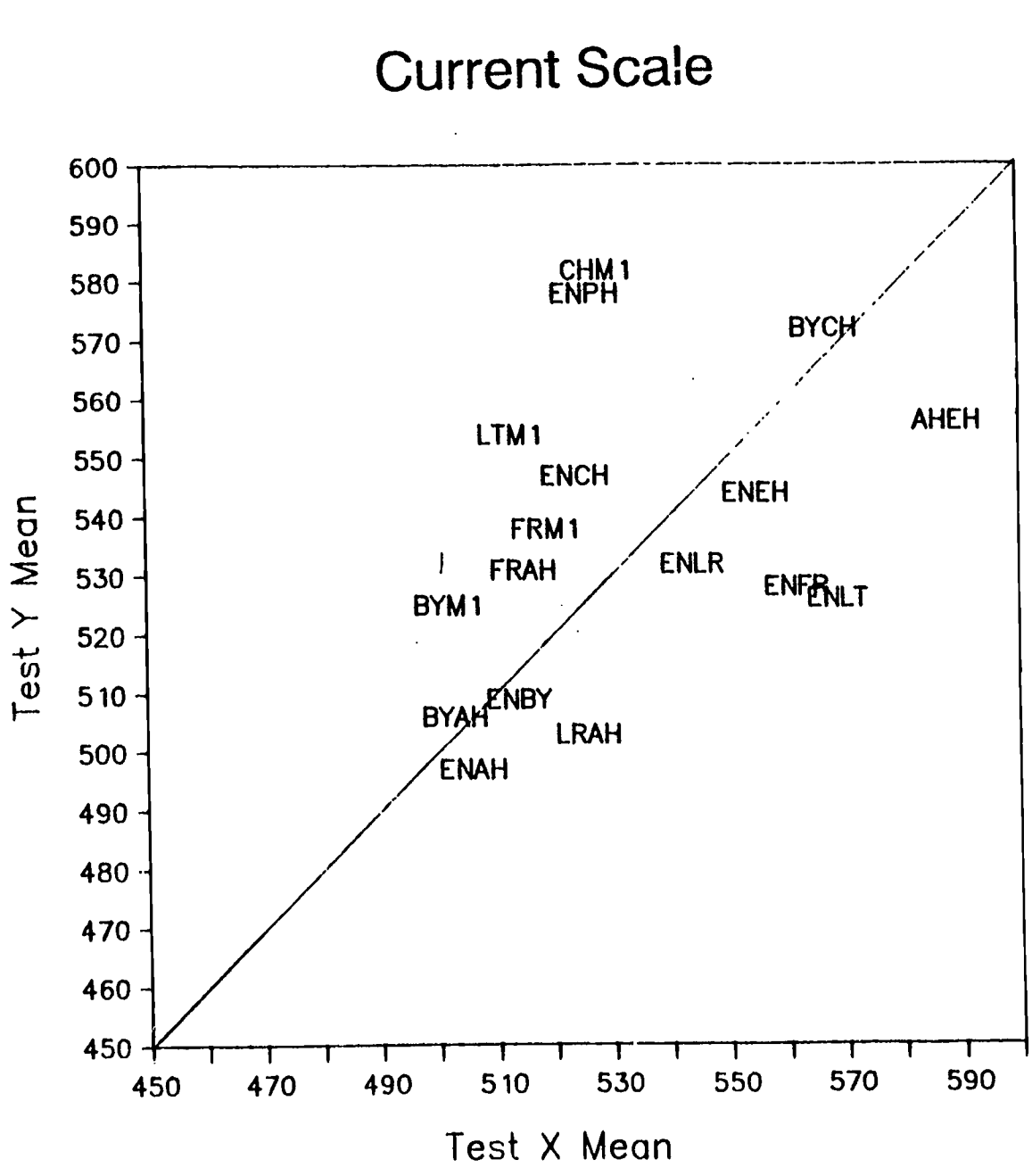


Figure 3a: Bivariate plot of Achievement Test scaled score means resulting from application of experimental scaling procedures.

Single-stage Scale

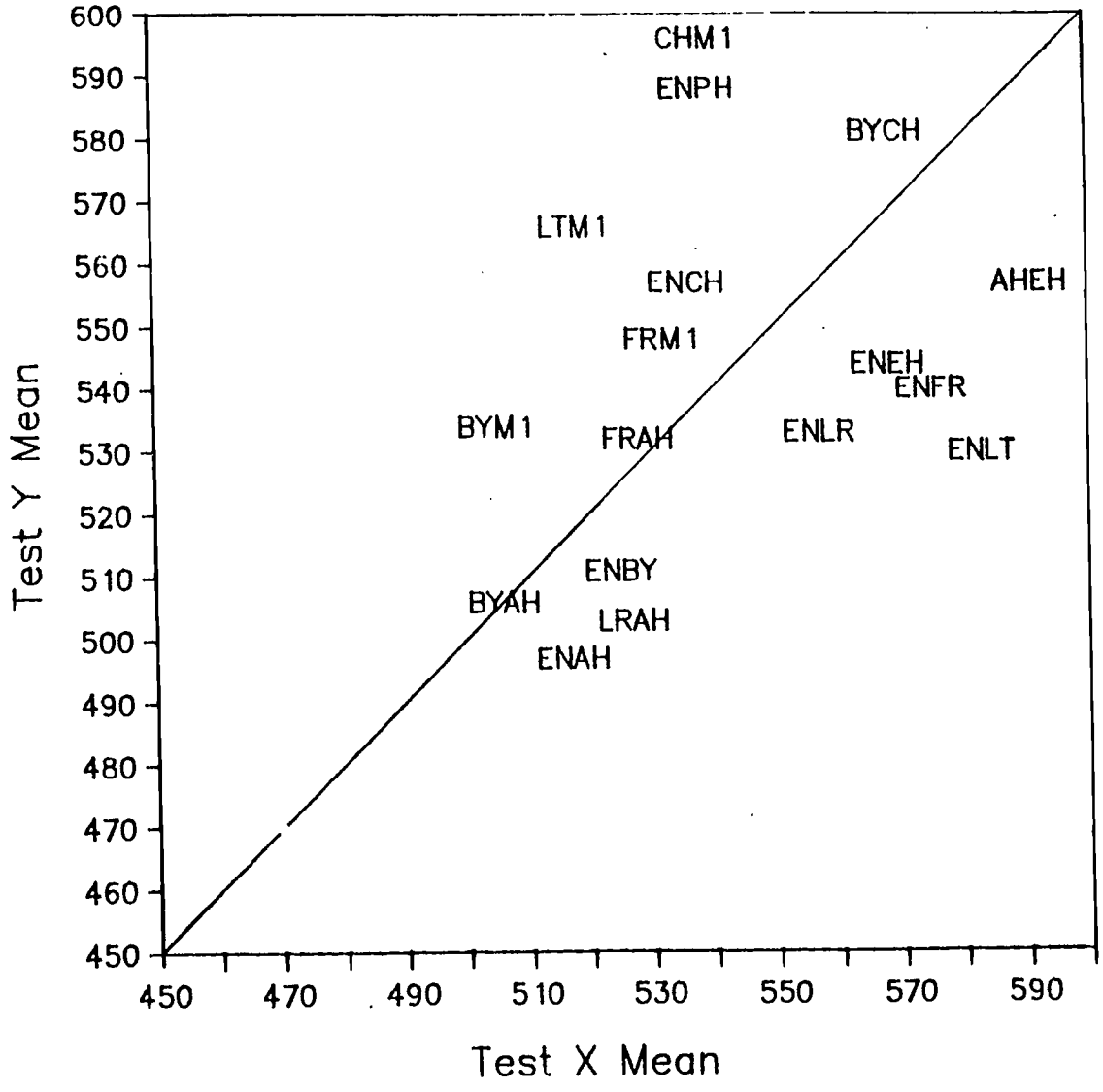


Figure 3b: Bivariate plot of Achievement Test scaled score means resulting from application of experimental scaling procedures.

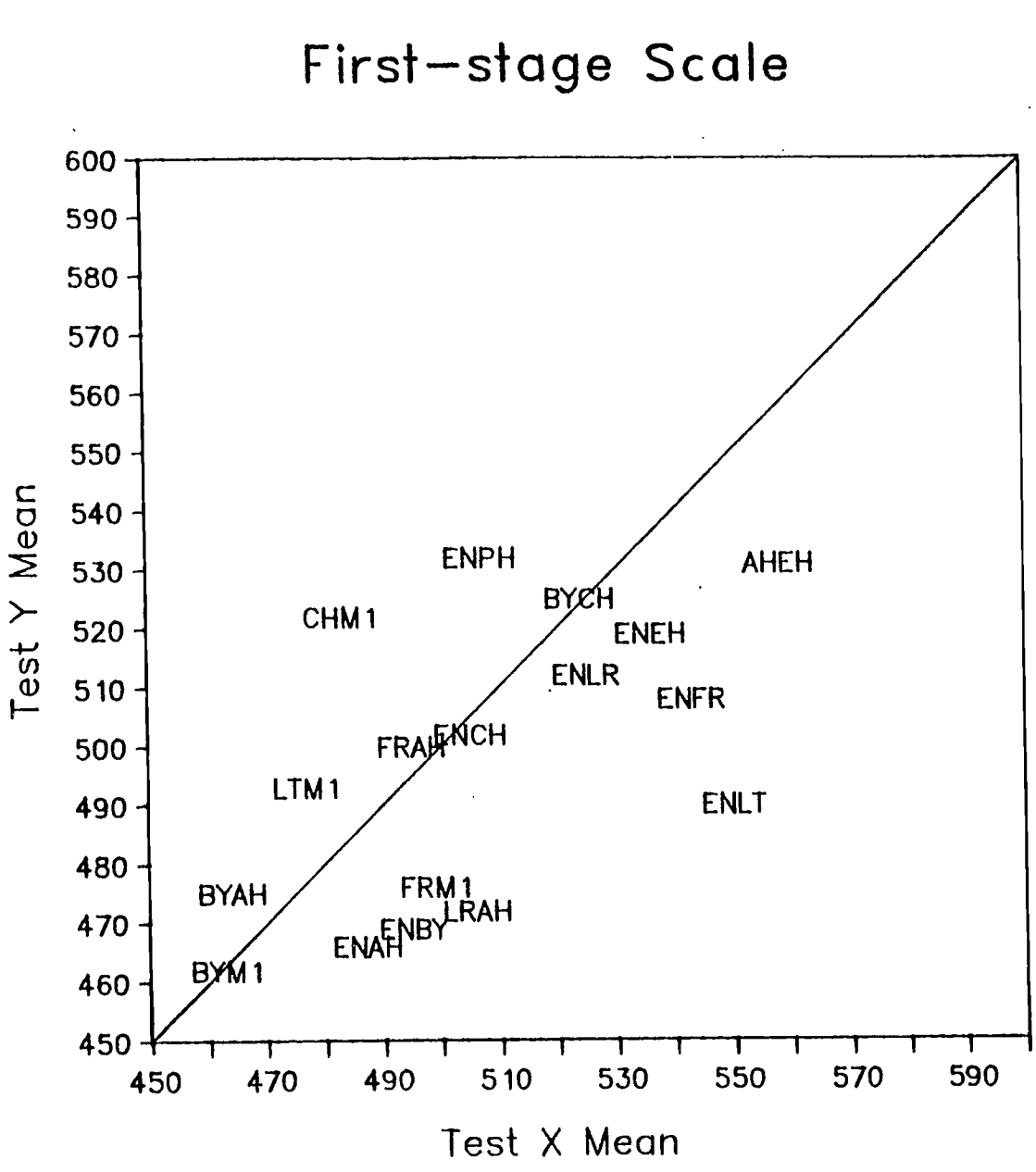


Figure 3c: Bivariate plot of Achievement Test scaled score means resulting from application of experimental scaling procedures.

Scaling to French Test

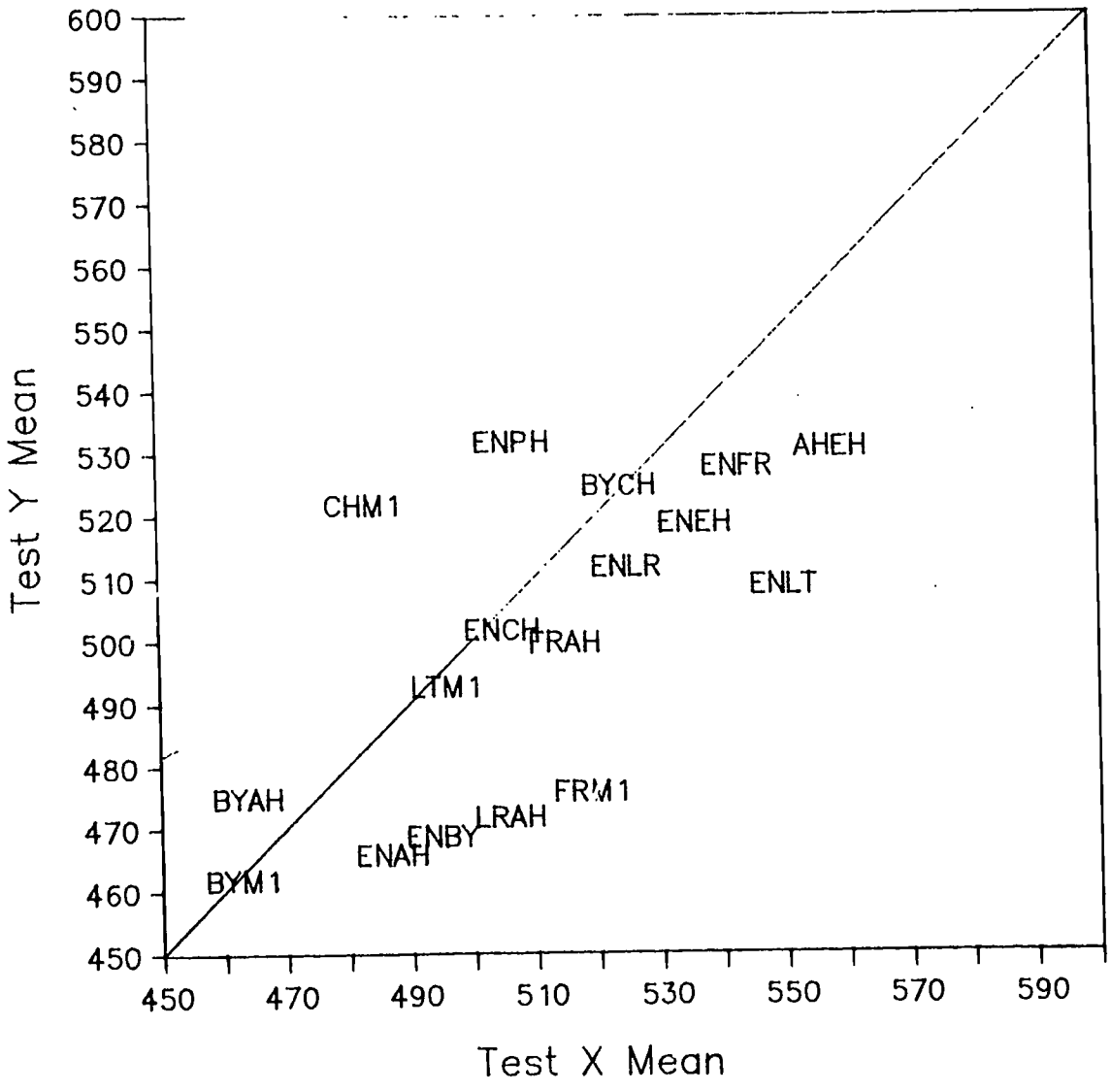


Figure 3d: Bivariate plot of Achievement Test scaled score means resulting from application of experimental scaling procedures.