ACT Research Report Series 93-6

A Multivariate Generalizability Analysis of the 1989 and 1990 AAP Mathematics Test Forms With Respect to the Table of Specifications

Dean A. Colton

December 1993

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

A MULTIVARIATE GENERALIZABILITY ANALYSIS OF THE 1989 AND 1990 AAP MATHEMATICS TEST FORMS WITH RESPECT TO ITS TABLE OF SPECIFICATIONS

Dean Alan Colton

Abstract

Tables of specifications are used to guide test developers in sampling items and maintaining consistency from form to form. This paper is a generalizability study of the AAP Mathematics test, with the content areas of the table of specifications representing multiple dependent variables. The results are presented with respect to variance and covariance components, and estimates of error variance. Also discussed are alternative weightings of the content categories.

The importance of tables of specifications in the development of standardized achievement tests is evident in both their widespread use and their contribution to consistency in measurement procedures. Different forms of a test must be used to ensure security, and these forms must be "similar" to ensure consistency in the measurement procedure. A table of specifications guides the test developer in at least two ways. First, it helps keep the test within prespecified limits such as the purpose of the test and the philosophy of testing. Second, it provides a guide or blueprint for sampling items on the basis of content, difficulty, item type, or other considerations.

The first section of this paper is a history of the evolution of tables of specifications in standardized achievement testing. The second section presents the application of a multivariate generalizability theory model (Jarjoura and Brennan, 1982, 1983) to the ACT Achievement Program (AAP) Mathematics test (American College Testing, 1989, 1991). The results of these analyses reveal psychometric characteristics of the mathematics measurement procedure as well as those of the particular forms that were analyzed. The central purpose of the analyses is to investigate the role of tables of specifications in the overall measurement procedure.

History of Tables of Specifications

The development of standardized tests according to a specified set of rules to measure different mental constructs was first introduced in intelligence testing. Originally, intelligence was believed to consist of many distinct and independent abilities or aptitudes. Thus, intelligence was measured by many distinct tests. According to Terman (1916), however, Alfred Binet abandoned that testing strategy. Believing that the aspects of intelligence are interrelated, Binet devised a test of general intelligence containing items sampled from numerous domains. Shortly before this development, standardized achievement tests were invented, and the view of achievement testing as the sampling of numerous domains followed.

Prior to 1845, students in the schools of Boston were tested orally by the Boston School Committee each year. In 1845, however, the oral examination was replaced with a written examination, requiring mostly short-answer responses. The results of this written examination were then tabulated, and school-level achievement was made public. In an 1845 issue of the Common School Journal (reprinted in Caldwell & Courtis, 1925) Horace Mann wrote that this method of written examination was impartial, "not in a limited, but in a very extended application of that term; for it submits the same question not only to all the scholars who are to be examined, in the same school, but to all the schools of the same class or grade. Scholars in the same school, therefore, can be equitably compared with each other; and all the different schools are subjected to a measurement by the same standard" (p 238).

According to several authors (e.g., Ruch and Stoddard, 1927; Traub, 1924) the inventor of the standardized achievement test was J. M. Rice.[1] Rice was a physician who abandoned that profession to devote his time to education (DuBois, 1970). At the end of the 19th century, a controversy over the school curriculum was dividing educators into two groups. Some schools were introducing courses in "practical subjects" such as manual arts and home economics. Opponents of this trend argued that time was being taken away from the teaching of traditional courses such as reading, spelling, and arithmetic. They believed that this reduction in teaching time would cause a lowering of achievement in the traditional subjects.

Rice decided that a comparison of achievement in schools using new curricula with achievement in schools using the traditional curriculum would reveal the efficacy of each. Thus, he developed a spelling test that was administered to classes in many different schools in 1894-1895. Rice (1897) reported that students taking the new curricula were as high in spelling achievement as those taking the old curriculum.

[1] Traub (1924) also mentions earlier tests, and cites an earlier test administered to different groups. In 1864 E. B. Chadwick reported a testing procedure used by the Reverend George Fisher in the Greenwich Hospital School in England. However, this did not receive much attention, and so was not an impetus for the development of other testing programs.

Then, Rice reported the results obtained with his arithmetic tests (1902). These tests were written for grades four through eight. Each test contained eight word problems, with some items common among tests. According to the author, "...for the purpose of studying the growth of mental power from year to year, some of the problems were carried through several grades. Thus of the eight questions for the fourth grade, five were repeated in the fifth, and three in the sixth, etc." (p. 283). These tests were administered to 5,963 students in 18 schools in seven cities.

Rice did not indicate how items were either generated or selected. However, his efforts inspired the work of Cliff Stone who was concerned with the content and construction of tests. According to Courtis (1913), "The evident defects of Rice's tests and methods, however, led Dr. C. W. Stone to attempt the standardization of a measure for sixth-grade work. Carefully prepared tests were given under uniform conditions. ... A full account of tests, methods, and tabulations was published making it possible for the teacher of any sixth-grade class to compare his work with that of the schools tested by giving the same tests under the same conditions, and following the same plan of scoring" (p. 397).

Stone reported the administration of two arithmetic tests to sixth grade students (Stone, 1908). One was a test of "Fundamentals" that included addition, subtraction, multiplication, and division. The other was a test of "Reasoning." Reasoning was not assembled with specific content guidelines, but items were selected that met certain criteria such as concreteness of the situation, and that did not contain any subject matter other than "whole number, fractions, and United States money." These items were all word problems. Both tests were speeded: No student finished the 14-item test of Fundamentals, and only a few students finished the 12-item test of Reasoning.

The conditions under which the tests were administered were well controlled. For example, Stone administered the tests himself. No one was allowed in the classrooms during testing except the teachers and students, and no warning about the tests was made to teachers or pupils before administration. The time limits were strictly observed: Twelve minutes were allowed for the test of Fundamentals, and 15 minutes were allowed for the test of Reasoning.

Clearly, Stone's tests were built with rather concrete guidelines. The test of Fundamentals was a stratified test, since it was specifically designed to contain items covering the four basic operations of arithmetic. The test of Reasoning followed strict rules for the inclusion of items, as previously discussed. Thus, although Stone did not describe a procedure as sophisticated as the use of a table of specifications, some characteristics of such a procedure were clearly present.

Stone reported data on approximately 6,000 students in 26 school systems in seven states. The tests were scored with more difficult items weighted heavier than easier items.

Analysis of the data was approached at two levels: scores associated with school systems and scores associated with individual students. One question addressed by Stone (1908) was, "How far does the possession of one ability imply the possession of others?" (p. 36). To answer this question, school system scores were submitted to correlation analysis. Pearson coefficients among the fundamental operations ranged from 0.805 to 0.933. Pearson coefficients between Reasoning and each of the four Fundamental operations ranged from 0.062 to 0.338. From these results it was concluded that, "Ability in any fundamental except addition implies nearly the same ability in other fundamentals in both systems and individuals; but ability in any fundamental implies ability in reasoning in individuals to a lesser degree than ability in such a subject as geography... Of fundamentals, division seems to be most like reasoning, perhaps subtraction next, then multiplication, a close third, and addition least of all" (p. 43).

In summarizing the work thus far, Courtis (1913) stated that, "To Rice is due the credit for the fundamental idea of comparison of schools by the results of tests given to all under uniform conditions; Stone emphasized standard achievement and scientific care in the preparation of the tests and in control of conditions; Courtis extended the idea of standards and adapted both tests and testing to the measurement and improvement of the efficacy of classroom work" (p. 398).

Courtis did not classify items within a test according to content. Instead, he developed eight tests, each covering a different content area (i.e., addition, subtraction, multiplication, division, copying figures, reasoning in one-step problems, abstract examples in the fundamentals, and reasoning in two-step problems). In the first four tests, however, items were classified according to judged difficulty. The items in the five difficulty classes were then dispersed throughout the test. Some classification of items was done in the test for abstract examples in the fundamentals. There, some items did not involve borrowing or carrying. Others involved borrowing or carrying of "small numbers," and the remaining items involved borrowing or carrying of larger numbers.

Although Courtis did raise and address many issues still of interest today (e.g., coaching, standard setting, and measurement of growth), he probably did not contribute positively to the development of specifications for constructing tests. Indeed, he may have slowed such progress in the area of mathematics testing, since his tests became widely used for both practical and research purposes (e.g., Ashbaugh, 1914; Haggerty, 1915; Monroe, 1918). This was unfortunate since mathematics testing was probably the most heavily researched area of achievement testing at the time.

One other early attempt at selection of items according to content was also in mathematics achievement testing. A report by Rugg and Clark (1918) suggests that their ninth grade algebra tests were constructed using a set of content guidelines. Although the authors did not describe their approach to the generation of items, they did say that, prior to construction of the tests, they reviewed textbooks to discover the content being taught. Further, they stated that the construction of a test required "Classifying clearly the subject matter of the course for which the test is being designed" (p. 52). Unfortunately, no results related to classification by content were reported, and their tests do not appear to have been adopted by others. No work on classification of items or the development of guidelines for test construction was evident for approximately ten years after their paper.

During this period, most researchers were concentrating on item statistics for the evaluation and inclusion of items. Not only was attention focused on item statistics, but some writers were openly critical of developing tests to measure more than one skill or knowledge in a broad content area. For example, Osburn (1933) cited Thorndike's extension of the Hillegas Scale of English Composition. In this instrument, ten compositions dealt with supplying details, three with causal relationships, one with inference, etc. Thus, the test sampled a variety of skills, and could be called a stratified test. For sampling from this broad domain, Thorndike was criticized by Ballou (1914) for trying to measure too much with a single test. Ballou stated that, "A scale should not try to measure too complex a product. To attempt to measure the several forms or types of English compositions by one and the same scale is like trying to measure heat, light and color by the same instrument ..." (p. 93). Ballou seems to have been arguing for the use of separate tests, much as Courtis (1913) separated Stone's (1908) Fundamental operations into four tests.

In distinguishing between "traditional" (essay) tests and "new-type" (objective) tests, Talbott and Ruch (1929) described essay tests as "intensive," and objective tests as "extensive." Essay tests were preferred by many educators because they require the student to write all he knows about the topic being tested. Thus, they allow for intensive sampling of knowledge on a topic. Objective tests, on the other hand, were described as extensive because they allowed sampling on many different topics in the same amount of testing time.

Table 1 contains reproductions of two tables Talbott and Ruch used to illustrate this distinction. In these tables, uppercase letters represent different topics, and lower case letters in the columns represent subtopics. The table labeled Scheme I represents sampling characteristics of essay tests. Only a subset of the major topics are sampled, but within each topic every subtopic may be sampled. In contrast, Scheme II illustrates that, with objective tests, every major topic is sampled, but only a subset of subtopics are sampled.

Insert Table 1 about here

The authors argued that the objective tests should be more "reliable" than essay tests since, "Knowledge of any one subtopic in a given column (such as b under A) is more likely to guarantee knowledge of other subtopics in different columns (such as columns D or N). It follows then, that the omissions in any examination, since most of the subject matter taught must be omitted, should not be entire columns (as in Scheme I) but rather subtopics in all columns with few or no columns omitted (as in Scheme II)" (p. 201).

Talbott and Ruch administered both an essay test and an objective test to a class of students. Then, the authors compared the two tests and methods of sampling knowledge. From their results, it was concluded that, "Since the essay examination requires twice as much time and evokes less than half as much knowledge, the objective test is from four to five times as efficient as a device for sampling. This will probably hold only for factual tests like the present ones, and even then very roughly" (p. 205).

Refinements of and Research on Tables of Specifications

Apparently, G. M. Ruch (1929) coined the term "table of specifications." According to Ruch, "The term table of specifications was adopted for the sake of emphasizing the need for a general guide as a skeleton in building a test. Such a table guards against the omission of essential items, the over-emphasis of minor topics, and improper balance of the sampling. The drawing-up of a working plan before drafting specific items goes a considerable distance in establishing the validity of the final test when completed" (p. 150).

Ruch listed ten steps in test construction:

"I. Drawing up a Table of Specifications

II. Drafting the items in preliminary form

III. Deciding upon the scope (length)

IV. Editing and selecting the final items

V. Rating the items for difficulty

VI. Breaking the items into alternative forms

VII. Rearranging the items in order of difficulty

VIII. Preparing the instructions for the tests

IX. Making the answer keys or stencils" (p. 149)

Ruch noted that in his plan it may seem illogical to decide on the length of the test after drafting the items, rather than as part of the table of specifications. He argued that one must know the number of good items available before deciding how many items from each topic will be included. On the other hand, in a sample table of specifications, he included the percentage of the total number of items on a test to be allotted to each major topic. The percentage of items allotted to each subtopic was not specified.

Working at about the same time as Ruch (1929), Ralph W. Tyler set forth a general procedure for the construction of achievement tests based on his work with classroom zoology tests. Although Tyler wrote several earlier papers on this (e.g., Tyler, 1930; Tyler, 1931), his most comprehensive statement was published in 1934 (Tyler, 1934). There he wrote that "A fundamental task in constructing achievement tests which will be used by college instructors is to make certain that the important objectives of the subject and course are adequately measured. This is so obvious a requirement for a valid examination that there is nothing new in the suggestion that it is the essential criterion for validity. However, techniques of test construction in which test items are consciously derived from the specific objectives of the course are much more rare" (p. 4).

Tyler then summarized the steps of objective achievement test construction as:

"1. Formulation of course objectives

2. Definition of each objective in terms of student behavior

3. Collection of situations in which students will reveal presence or absence of each objective

4. Presentation of situations to students

5. Evaluation of student reactions in light of each objective

6. Determination of objectivity of evaluation

7. Improvement of objectivity, when necessary

8. Determination of reliability

9. Improvement of reliability, when necessary

10. Development of more practicable methods of measurement, when necessary" (p. 5)

Next, Tyler specified eight types of objectives used in his work at Ohio State University. "These are:

Type A, Information, which included terminology, specific facts, and general principals

Type B, Reasoning, or scientific method, which includes induction, testing hypotheses, and deduction

Type C, Location of Relevant Data, which involves a knowledge of sources of usable data and skill in getting information from appropriate sources

Type D, Skills Characteristic of Particular Subjects, which include laboratory skills in the sciences, language skills, and the like

Type E, Standards of Technical Performance, which includes the knowledge of appropriate standards, ability to evaluate the relative importance of several standards which apply, and skill or habits in applying these standards" (p. 475)

In addition to calling for the formulation of objectives to use in the construction of tests, and regardless of whether objectives were stated prior to construction, Tyler suggested that guidelines for item content be used in test construction. These objectives would serve then as a table of content categories to be sampled by the test. Consideration of the type of objective gives this approach the appearance of using a table of specifications in the sense that term is commonly used today. That is, objectives crossed with type of objective would yield specifications containing classes of items defined from more than one perspective. Further, consideration of the type of objective is clearly similar to the later, more widely cited work of Bloom (1956).

At about the same time, Brueckner and Elwell (1932) reported a study of differences among item classes in errors in multiplication of fractions. Using the Brueckner Diagnostic Test of Fraction Multiplication, the authors selected six item-types that differed in the number systems of the factors and products (e.g., a mixed number multiplied by a whole number with a mixed number product). Using four examples of each item type, the 24 items were randomly ordered for administration. Students (N=327) from five cities worked the items in an unspeeded test.

The authors first examined the number of errors made. They concluded that "An error on a single example of a given type is not at all a reliable index of what a pupil is likely to do on another example of that type since in 59.8 percent of the cases pupils who solved one example of the four of a given type correctly, missed from one to three of the remaining three. In only one case in five did a pupil who worked one example incorrectly also have errors on the other three examples of the same type included in the test" (p. 177).

Next, the authors examined the types of errors made, and how these were distributed over the six item types. In reviewing the subjects' work, the authors found errors associated with (1) not knowing how to proceed, (2) computation, (3) cancellation, (4) working with improper fractions, and (5) legibility; among others. Along this line of investigation, the authors examined consistency among types of errors within item types that a subject made. They reported "A relatively high degree of consistency of the type of error or specific fault found in a pupil's work when as many as three or four examples of a single type were solved incorrectly" (p. 185).

Brueckner and Elwell (1932) did not investigate the table of specifications for a full test. They did, however, examine some strata of the test to investigate questions related to item type and minimum length of strata. Such investigations have been rare.

E. F. Lindquist (1936) made explicit the concept of a multidimensional table of specifications. He stated that, the "table of specifications should in most cases be multiple in nature; i.e., it should provide for several independent classifications of the content to be tested, each with reference to a different point of view. In building an American history test, for example, the elements of content might be classified chronologically, or topically, or according to the type of history involved, such as social, economic, political and cultural history, or according to types of associations, such as between men and events, ..., historical terms and meanings, etc." (p. 108). He also advised that the number of items to be sampled from each category be estimated.

This concept of a table of specifications is very similar to that described in current measurement textbooks (e.g., Ebel and Frisbie, 1991). Lindquist's description does not include classification according to student behaviors, which was later advocated by Benjamin Bloom.

At the 1949 convention of the American Psychological Association, a meeting led to the agreement among attending college examiners that a system of classifying educational objectives would aid in "the exchange of test materials and ideas about testing... in stimulating research on examining and on the relations between examining and education" (Bloom, 1956, p. 4). Meetings in subsequent years led to the development of what is now popularly called Bloom's Taxonomy.

According to Bloom (1956), the "taxonomy is designed to be a classification of student behaviors which represent the intended outcomes of the educational process. It is assumed that essentially the same classes of behavior may be observed in the usual range of subject matter content, at different levels of education (elementary, high school, college) and in different schools" (p. 12). Further, the taxonomy is believed to have a logical order. That is, the order of levels in the taxonomic structure correspond to the order of levels in student behavior according to a theoretical structure of cognition, from simple to complex. These levels of the taxonomy are:

(1) Knowledge

(2) Comprehension

(3) Application

(4) Analysis

(5) Synthesis

(6) Evaluation

Not only are these levels believed to be ordered from simple to complex, but also "The objectives in one class are likely to make use of and be built on the behaviors found in the preceding classes in the list" (Bloom, 1956, p. 18).

This hierarchical structure was investigated by Madaus, Woods, and Nuttall (1973). Two social science and two natural science achievement tests, developed according to Bloom's Taxonomy, were administered to 1,128 grade 9-12 students. Using a path analysis approach, the authors investigated the variance in each level explained by the variance in the next lower level and in nonadjacent levels. In addition, for some analyses a general factor for ability, g, was introduced into the model.

It was reported that both the adjacent and nonadjacent links were highly dependent upon the g factor. This was especially true of Synthesis and Evaluation for the lower grades. Although this relationship was less pronounced at the higher grades, with familiar content the link between Analysis and Evaluation was retained in the model.

Also, at the higher grades and with familiar content, the indirect links between the three lower levels and the higher levels were retained in the model. From these results, the authors concluded that the hierarchical structure of the taxonomy is highly dependent upon general ability, but that Synthesis and Evaluation may be hierarchically related to the lower levels.

Another approach to examining stratified tests developed according to a table of specifications was proposed by Bock and Bargman (1966). From their perspective, there is apriori information about the items which is used to create subclasses of items in a hierarchical design. The authors developed a method for estimating variance components when the design contains one random factor, and one fixed and possibly unbalanced factor. The subclasses of items constitute the fixed classification.

The method employs covariance structure models, a linear least-squares estimator of "latent scores," and maximum-likelihood estimators of variance components. As well as the technical development, the authors provide examples that illustrate use of the method under various conditions. That is, homogeneous error variance, nonhomogeneous error variance, and an incomplete factorial design (some subclasses of items are not sampled). The authors argue that their "structural analysis" of a sample covariance matrix "is more general than the conventional mixed model analysis in that the design for the fixed classifications may be nonorthogonal, the replication error variance for different subclasses of the fixed classifications may be nonhomogeneous, and the measurements for these subclasses may be in different metrics" (p. 533).

Finally, Jarjoura and Brennan (1981) developed three variance components models of tests developed from a table of specifications. Using a multivariate generalizability perspective (see Cronbach, Gleser, Nanda, and Rajaratnam, 1972, and Brennan, 1992a), the authors model tests in which universes of items are nested within levels of test strata. In the first model, the strata (e.g., content categories, type of objective) are assumed to be random variables. In the other two models, strata are assumed to be fixed. These last two models differ in the restrictiveness of their assumptions, and in their generality. The “restrictive model” is similar to ANOVA models, and the “non-restrictive model” is a multivariate approach.

Jarjoura (1981) extended this approach by developing multivariate models for English and natural sciences subtests of the ACT Assessment Program. Those tests were stratified by content. The English Usage test contained reading passages and items testing punctuation, grammar, sentence structure, diction, style, and logic and organization. The natural science test contained passages and discrete items covering physics, chemistry, physical sciences, and biology.

After applying these models to the tests, it was reported that they provide more “detailed results for studying the structure of measurement procedures” (p. 35) of the stratified type, than would analyses that ignore strata. Further, it was concluded that “through such an analysis, there is a clear potential for suggesting improvements in a procedure; and perhaps more importantly, for suggesting further research on aspects of a procedure that are not readily recognizable in less detailed analyses” (p. 36).

A Model for Tests Developed According to Tables of Specifications

According to Jarjoura and Brennan (1982), the model involves "the responses of P persons to I_+ items where the items fall into C fixed categories or cells of a table of specifications with I_c items in category c, so that $I_+ = \Sigma_{c=1}^{C} I_c$." (p 162). At the most elementary sampling level, an observation is

$$Y_{pic} = \mu_c + \pi_{pc} + \beta_{i:c} + \pi\beta_{pi:c} ,$$

$$p=1, ...,P; \;\; i=1, ...,I_c; \;\; c=1, ...,C , \qquad (1)$$

where Y_{pic} is the response of a person p to an item i in category c. The i:c means that the item is nested within the category. Persons and items are random, and categories are fixed effects. Thus, the category means μ_c are fixed effects in the universe of generalization; whereas, the universe scores π_{pc} , item effects $\beta_{i:c}$, and residuals $\pi\beta_{pi:c}$ are random with expectations of zero. It is assumed that persons are randomly sampled from the population, items are randomly sampled from an infinite universe of items associated with each category, and the categories are both exhaustive of a finite set and are mutually exclusive. Since each category represents a separate universe of items, the model given in Equation 1 is multivariate. It is further assumed that all effects other than universe score effects are uncorrelated.

Under the assumptions of random sampling, the random effects can be defined as expectations over samples. Then, variance and covariance components are defined as expectations of squares and products of effects taken over the population of persons and universes of items within content categories. Finally, unbiased estimators of variance and covariance components can be found by linear functions of mean squares and products. It is only necessary for analysis purposes to assume that the model is appropriate for the data and that every category contains two or more items. Jarjoura and Brennan (1982; 1983) describe this generalizability approach in detail, and give an example of its application. Colton (1983) replicated the Jarjoura and Brennan (1982) study and reported the analyses in greater detail. Brennan (1992a) also discusses many aspects of this approach.

The purpose of the study reported here was to analyze the Mathematics Test of the ACT Assessment Program (AAP) using data obtained from the new version of the test that was introduced in 1989. This new version of the Mathematics Test (American College Testing, 1989, 1991) is constructed according to a different table of specifications than was the original. The analysis was conducted using the generalizability model discussed above (Jarjoura & Brennan, 1982).

Method

Subjects. Data on two groups of examinees were analyzed. The examinees were all persons at selected AAP test administration centers in 1989 and 1990. Data on such samples are routinely used in the annual equating of new forms. These samples are considered to be relatively stable and reasonably representative of the population of examinees.

Data. Between 3,000 and 3,500 examinees were administered each of eight forms in 1989 and each of nine forms in 1990. These forms were spiraled within test centers. The eight forms administered in 1989 will be referred to as Forms A through H, and the nine forms administered in 1990 will be referred to as Forms G through 0. Notice that Forms G and H were administered in both years. Item responses were scored zero-one, and raw test scores were calculated as the sum of the scored responses.

The AAP Mathematics Test. The Mathematics test of the AAP (American College Testing, 1989) consists of 60 five-alternative, multiple-choice items. Examinees are allowed 60 minutes to respond to all items, although the test is not considered to be speeded. In the building of test forms, items are selected with respect to the table of specifications, which represents four broad content categories. These categories are Pre-algebra and Elementary Algebra (PEA), Intermediate Algebra and Coordinate Geometry (IAG), Plane Geometry (GEO), and Trigonometry (TRG). Following the table of specifications, items are sampled from these categories in the following numbers and proportions: PEA has 24 items or .40, IAG has 18 items or .30, GEO has 14 items or .23, TRG has 4 items or .07.

Results

The multivariate generalizability p x i:c model was used to conceptualize the present study. The generalizability analyses were conducted using a computer program called GAST (Brennan, 1992b). As described above, the data included test scores on eight forms administered in 1989, and nine forms administered in 1990. The data from these two years were analyzed separately.

Means of proportion-correct scores and their standard deviations are presented in Table 2 for 1989 data, and in Table 3 for 1990 data. The means in Tables 1 and 2 are graphically presented in the box and whisker plots of Figure 1. In Tables 2 and 3 the first column identifies the form designation. Columns one through four contain the means and standard deviations for each content category, column five contains the statistics for the entire test, and the last column presents the sample sizes. For the 1989 data, the means for the whole test (the column headed ALL) have a range of .054 and a standard deviation of .020. The mean taken over forms has a standard error of only .007.[2] It appears that the forms are similar in difficulty, but the standard deviations for TRG are generally small. Form-to-form differences are also small for the content categories. Examination of the row labeled "Average" reveals that the content categories differ in difficulty. Ordered from easiest to most difficult, the content categories are arranged: PEA, GEO, IAG, and TRG. These form profiles are graphically presented in Figure 2 for the 1989 data, and in Figure 3 for the 1990 data.

Insert Tables 2 and 3 about here

Insert Figures 1, 2, and 3 about here

For the 1990 data, the means for the whole test have a range of .042, and the mean taken over forms has a standard error of only .004. Thus, the forms administered in 1990 appear to be somewhat more consistent in difficulty than do those administered in 1989. Ordering the content categories from easiest to most difficult gives the arrangement: PEA, GEO, IAG, TRG. This is the same ordering as was found for the 1989 data. On

[2] The standard error of a mean of k elements is the standard deviation divided by the square root of k. In this report, standard errors of means over forms in a given year are routinely reported. These statistics are especially appropriate for the AAP because ACT typically reports results over forms and test dates for a given year, rather than for single forms or test dates.

the other hand, the IAG, GEO and TRG content categories were more difficult in 1990 than in 1989.

Universe score variances and covariances for each form are presented in Tables 4 and 5. The 1989 universe score variances for PEA appear relatively stable, having a range of .009 over forms. For IAG, the range is larger, .012. The ranges for GEO and TRG are notably larger, .021 and .055 respectively; probably, at least in part, due to rather small numbers of items in these categories, especially in TRG.

In Table 4 the 1990 universe score variances have ranges of .017, .016, .009, and .037, for PEA, IAG, GEO, and TRG, respectively. Here, the range for GEO is quite small, even though one might expect a less stable statistic for a category with relatively few items.

Insert Tables 4, 5, 6 and 7 about here

Estimated universe score variance and covariance components averaged over forms are presented in Tables 6 and 7. The standard errors of these averages are presented in italics. A category-to-category comparison reveals that the estimated universe score variance components are of similar magnitude, at least when examined in the aggregate over test forms. The mean covariance components also appear to be similar across categories. The values for the TRG category are the most differing, though not greatly. Notice, however, that the standard error of the mean universe score variance components for TRG is distinctly larger than those of the other categories.

That the average universe score variance and covariance components are of about the same magnitude suggests that the universe score correlations between content categories should be relatively large. This expectation is substantiated in Tables 8 and 9.

Insert Tables 8 and 9 about here

Estimated item effect variance components are presented in Tables 10 and 11. Within most categories the values for 1989, in general, vary substantially among test forms. The IAG category has a small range of values, .008, and TRG has the largest range, .050. Again, the large range for TRG is at least partially due to the small number of items. The 1990 estimated item effect variance components also vary within categories and across test forms. Again, the range for IAG, .022, is smallest, and the range for TRG, .049, is the largest.

Insert Tables 10 and 11 about here

Estimated residual effect variance components are given in Tables 12 and 13. In Table 8, the ranges of values over the 1989 forms is .014, .009, .016, and .039 for categories PEA, IAG, GEO, and TRG, respectively. In Table 9 the 1990 ranges are .025, .027, .034, and .039. Thus, except for TRG, the residual variance components were nominally more variable within categories and over forms in 1990 than in 1989.

Average estimates of item effect variance components, taken over forms, are provided in the next to the last row of Tables 10 and 11. (The standard errors of these averages are in the last row of this table.) These averaged estimates range from .022 for IAG to .031 for PEA, in the 1989 data. They range from .018 for TRG to .036 for PEA in the 1990 data.

Insert Tables 12 and 13 about here

Average estimates of residual effects variance components, taken over forms, are provided in the next to the last row of Tables 12 and 13. The average of the estimates for TRG is the lowest, .173, and the averages for IAG and GEO are the highest, .187, in the 1989 data. They range from .172 for TRG to .184 for GEO in the 1990 data.

Estimates of Error

The absolute error variance (i.e., the variance of the difference between the observed and universe scores) is .00352 for the 1989 data and .00353 for the 1990 data. (Note that these values are based on mean scores, not total scores.) These were calculated using the item and residual variance components averaged over forms. Clearly, the values are similar. The average of the 1989 composite universe score variances was found to be .032, and the 1990 value was .029. Using the error variance and composite universe score variance, we can construct a signal-noise ratio as described by Brennan & Kane (1977). For the 1989 data, the signal-noise ratio is 9.2. For the 1990 data the signal-noise ratio is 8.3. Using the error variance and composite universe score variance, we can also calculate a dependability index. For the 1989 and 1990 data, the dependability indices are .90 and .89, respectively.

Given that the several content categories differentially contribute to the error variance, it is interesting to investigate how changes in category size might affect the magnitude of the absolute error variance. Following the development of Jarjoura and Brennan (1982), the "optimal" content category lengths, in number of items, were found for minimizing the estimated error variance. For the 1989 data, the optimal lengths for PEA, IAG, GEO, and TRG are 24.1, 17.9, 14.1, and 3.9, respectively. For the 1990 data, the optimal lengths are 24.0, 18.0, 14.2, and 3.8. These values round to the operational lengths of 24, 18, 14, and 4. Furthermore, recalculation of the absolute error variances based on the optimal category lengths, produced variances of .00352 and .00353 for 1989 and 1990, respectively. Not surprisingly, these are equal to those reported earlier for the operational length category strata.

Following the work of Wang and Stanley (1970), Jarjoura and Brennan (1982) described an "effective weight" as "the covariance between a variable and the composite variable" (p. 165). For the 1989 data, the effective weights for PEA, IAG, GEO, and TRG were .013, .010, .007, and .002. The 1990 values were .013, .008, .006, and .002, respectively. These effective weights sum to the composite universe score variance (.032 for the 1989 data, and .029 for the 1990 data). Thus, taking the effective weights in ratio to the composite universe score variance gives the proportional contributions of the categories to the composite universe score variance. In the 1989 data, these proportional weights were .409, .297, .233, and .062 for PEA, IAG, GEO, and TRG, respectively. These values are close to the nominal weights of .4, .3, .2333, and .0667, (defined as the proportion of items in each category in the table of specifications). Likewise, in the 1990 data the proportional weights were found to be .429, .289, .222, and .060, for PEA, IAG, GEO, and TRG in that order.

Jarjoura and Brennan (1982) also developed an estimate of mean squared error for the hypothetical case in which items are sampled in proportions different from those laid out in the table of specifications. For example, we might ask how error variance would change if we obtained the sample of 60 test items by drawing equally (i.e., 15 items) from each category. These estimates of error variance were found to be .00361 and .00411 for 1989 and 1991, respectively. Recalling that the usual estimates were .00352 and .00353, it is clear that sampling equally from the categories yields slightly higher estimates of error variance.

In addition to the absolute error variance reported earlier, we may find a relative error variance. In the present model, relative error variance does not contain a term for items, and may be considered to be an error variance adjusted for form-to-form differences in difficulty. Relative error variance is appropriate for cases in which the composite universe score has been adjusted for form-to-form differences in difficulty -- as when scores from two forms have been equated. Thus, for the present model, relative error variance of unequated scores should be approximately equal to the absolute error variance of equated scores. Estimating the 1989 error variance for a composite universe score that is adjusted for difficulty, a value of .00306 was found, which is smaller than that found for the absolute variance (.00352). For the 1990 data, the adjusted error variance is .00299, which is smaller than the estimate of absolute variance, .00353.

Recall the signal-noise ratios reported for the 1989 and 1990 data, 9.2 and 8.3. Using the error variances adjusted for form-to-form differences in difficulty, the signal-noise ratios become 10.6 and 9.8 for the 1989 and 1990 data, respectively.

The relative error variance can be used to calculate a generalizability coefficient. For the 1989 data the generalizability coefficient was .90. For the 1990 data the generalizability coefficient was .91.

Discussion

The pattern of mean difficulty of the four content areas is the same for the 1989 and 1990 data. It appears that the specifications for the mathematics examination are resulting in fairly consistent levels of difficulty, at both the test and content strata levels. However, the form-level standard deviations for trigonometry are generally smaller than the standard deviations of the other content areas.

Examination of the universe score variance components revealed that within content category, estimated universe score variances were fairly stable over forms for some content categories. However, the smaller the number of items in the content category, the more variable were the values. Examination of the item effect variance components leads to a similar conclusion.

In general, examination of the variance components leads to the conclusion that universe scores, item effects, and residuals made fairly consistent form-to-form contributions to the variability of scores in content areas with relatively large proportions of the test's items. On the other hand, content areas with relatively small proportional representation in the test, had greater form-to-form differences in their variance components.

Averaging universe score variance components over forms, it was found that the content category values were of about the same magnitude. However, the standard error of the mean universe score variance component for trigonometry items was substantially larger than the standard errors of the other content category means. This reflects the relatively large form-to-form differences in the magnitude of these universe score variance components. The same phenomenon was found for the standard errors of the mean variance component for item effects.

Form-to-form differences in the trigonometry items are of some interest. Three observations go hand in hand. First, the standard deviation of the item difficulties were small compared with the other content categories, except on Forms A, H, and O. Second, the item effect variance components were small to moderate in magnitude compared with the other categories, except on forms A, H, and O. Third, the standard error of the mean universe score variance component for the trigonometry category was relatively large. Thus, there appears to be form-to-form inconsistencies in the performance of the trigonometry category. This is probably partly due to the small number of items in that category. It is also partly a result of an apparent form-to-form inconsistency in the variability in the difficulty of the trigonometry items.

Examination of various estimates of error and reliability-like coefficients, revealed that the average error variance was quite stable over the two years that were studied. Further, estimates of error variance were not greatly affected by the different item sampling plans that were considered.

References

American College Testing. (1989). *Preliminary technical manual for the Enhanced ACT Assessment.* Iowa City, IA: American College Testing.

Ashbaugh, E. J. The arithmetical skill of Iowa school children. *University of Iowa Extension*, LC6301, I65, No. 24.

Ayres, L. P. (1918). History and present status of educational measurements. In Guy Montrose Whipple (Ed.), *The seventeenth yearbook of the National Society for the Study of Education, part II, the measurement of educational products.* Bloomington, IL: Public School Publishing Company.

Ballou, F. W. (1914). Scales for the measurement of English composition. *Harvard Bulletins in Education* (No. 2). Cambridge, MA: Harvard University Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, the classification of educational goals, handbook I: Cognitive domain.* New York: David McKay Company, Inc.

Bock, R. D., & Bargman, R. E. (1966). Analysis of covariance structures. *Psychometrika, 31*, 507-543.

Brennan, R. L. (1992a). *Elements of generalizability theory* (revised edition). Iowa City, IA: American College Testing.

Brennan, R. L. (1992b). *Manual for GAST.* Unpublished manuscript. Iowa City, IA: American College Testing.

Brennan, R. L., & Kane, M. T. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika, 42*, 609-625.

Brennan, R. L., & Kane, M. T. (1979). Generalizability theory: A review. In R. F. Trabue (Ed.), *New Directions for Testing and Measurement.* San Francisco: Jossey-Bass, *4*.

Brueckner, L. J. *Diagnostic test in multiplication of fractions.* Minneapolis: Educational Test Bureau.

Brueckner, L. J., & Elwell, M. (1932). Reliability of diagnosis error in multiplication of fractions. *Journal of Educational Research, 26*, 175-185.

Caldwell, O. T., & Courtis, S. A. (1925). *Then & now in education: 1845-1923.* Younkers-on-Hudson, NY: World Book Co.

Colton, D. A. (1983). *A multivariate generalizability analysis of a test developed according to a table of specifications.* Unpublished master's thesis. Iowa City, IA: University of Iowa.

Courtis, S A. (1911). Standard scores in arithmetic. *The Elementary School Teacher, 12*, 127-137.

Courtis, S. A. (1911-1913). The Courtis tests in arithmetic. Section D, Subdivision I, Part II, Volume I of *Report of Committee on School Inquiry.* City of New York.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

DuBois, P. H. (1970). *A history of psycological testings.* Boston: Allyn and Bacon, Inc.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Haggerty, M. E. (1915). Arithmetic: A cooperative study in educational measurements. *Indiana University Studies* (No. 27). Bloomington: Indiana University.

Jarjoura, D. (1981). *Variance components models for content and passage effects in the English usage and natural sciences subtests of the ACT Assessment Program.* (ACT Technical Bulletin No. 39). Iowa City, IA: American College Testing.

Jarjoura, D., & Brennan, R. L. (1981). *Three variance component models for some measurement procedures in which unequal numbers of items fall into discrete categories* (ACT Technical Bulletin No. 37). Iowa City, IA: American College Testing.

Jarjoura, D., & Brenna, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6,* 161-171.

Jarjoura, D., & Brennan, R. L. (1983). Multivariate generalizability models for tests developed from tables of specifications. In L. J. Fyans (Ed.), *Generalizability theory inferences and applications.* New directions for testing and measurement, No. 18. San Francisco: Jossey-Bass.

Lindquist, E. F. (1936). The construction of tests. In Herbert E. Hawkes, E. F. Lindquist and C. R. Mann (Eds.), *The construction and use of achievement examinations.* Boston: Houghton Mifflin Company.

Madaus, G. F., Woods, E. M., & Nuttall, R. L. (1973). A causal analysis of Bloom's taxonomy. *American Educational Research Journal, 10,* 253-262.

Mann, H. (1845). Boston grammar and writing schools. *The Common School Journal, 7*(19).

Monroe, W. L. (1918). A report of the use of the Courtis standard research tests in arithmetic in twenty-four cities. *Indiana University Studies* (No. 38). Bloomington: Indiana University.

Osburn, W. J. (1933). The selection of test items. *Review of Educational Research, 3,* 21-32.

Rice, J. M. (1897). The futility of the spelling grind. *The Forum, 23,* 163-172, 409-414.

Rice, J. M. (1902-1903). Educational research: A test in arithmetic. *The Forum, 34,* 281-297.

Ruch, G. M., & Stoddard, G. D. (1927). *Tests and measurements in high school instruction.* Younkers-on-Hudsn, England: World Book Company.

Ruch, G. M. (1929). *The objective or new-type examination.* Chicago: Scott Foresman and Company.

Rugg, H. O., & Clark, J. R. (1913). *Scientific method in the reconstruction of ninth-grade mathematics.* Supplementary Educational Monographs (Vol. II, No. 1). Chicago: The University of Chicago Press.

Stone, C. W. (1908). Arithmetical abilities and some factors determining them. *Contributions to Education* (No. 19). New York: Teachers College, Columbia University.

Talbott, E. O., & Ruch, G. M. (1929). Minor studies on objective examination methods, II--Theory of sampling as applied to examinations. *Journal of Educational Research, 20,* 199-206.

Thorndike, E. L. (date unknown). *Preliminary extension of the Hillegas scale for the measurement of quality in English composition by young people.* New York: Teachers College, Columbia University.

Trabue, M. R. (1924). *Measuring results in education.* New York: American Book.

Tyler, R. W. (1930). Measuring the ability to infer. *Educational Research Bulletin, 9,* 475.

Tyler, R. W. (1931). A generalized technique for constructing achievement tests. *Educational Research Bulletin, 10,* 199-208.

Tyler, R. W. (1933). Formulating objectives for tests. *Educational Research Bulletin, 21,* 197-206.

Tyler, R. W. (1934). *Constructing achievement tests.* Columbus: The Bureau of Educational Research, Ohio State University.

TABLE 1
Talbott and Ruch's (1929) Two Sampling Schemes

SCHEME I*
The Character of the Sampling Afforded by the Traditional or Essay Examination

A	B	C	*D*	E	*F*	G	H	I	*J* ...	*N*
a	a	a	*a*	a	*a*	a	a	a	*a* ...	*a*
b	b	b	*b*	b	*b*	b	b	b	*b* ...	*b*
c	c	c	*c*	c	*c*	c	c	c	*c* ...	*c*
d	d	d	*d*	c	*d*	d	d	d	*d* ...	*d*
e	c	c	*e*	e	*e*	c	c	c	*e* ...	*e*
f	f	f	*f*	f	*f*	f	f	f	*f* ...	*f*
g	g	g	*g*	g	*g*	g	g	g	*g* ...	*g*
h	h	h	*h*	h	*h*	h	h	h	*h* ...	*h*
i	i	i	*i*	i	*i*	i	i	i	*i* ...	*i*
j	j	j	*j*	j	*j*	j	j	j	*j* ...	*j*
.	.	.	.	.	.	.	.	.	. . . .	.
.	.	.	.	.	.	.	.	.	. . . .	.
.	.	.	.	.	.	.	.	.	. . . .	.
n	n	n	*n*	n	*n*	n	n	n	*n* ...	*n*

SCHEME II*
The Character of the Sampling Afforded by the New-Type or Objective Examination

A	B	C	D	E	F	G	H	I	J ...	N
a	a	a	*a*	a	a	*a*	*a*	*a*	a ...	a
b	b	*b*	b	*b*	b	b	*b*	b	*b* ...	b
c	c	c	c	c	*c*	c	c	c	c ...	c
d	*d*	d	*d*	d	d	d	d	*d*	d ...	d
c	e	e	c	*e*	e	c	*e*	e	e ...	e
f	f	*f*	f	f	*f*	*f*	f	f	f ...	*f*
g	g	g	*g*	g	g	g	g	g	*g* ...	*g*
h	h	h	h	h	h	h .	h	h	h ...	*h*
i	i	i	*i*	i	i	*i*	i	i	i ...	i
j	j	j	j	*j*	j	j	j	j	j ...	j
.	.	.	.	.	.	.	.	.	. . . .	.
.	.	.	.	.	.	.	.	.	. . . .	.
.	.	.	.	.	.	.	.	.	. . . .	.
n	*n*	n	n	n	*n*	n	n	n	n n n n	n

*Italicized letters represent sampled subtopics.

TABLE 2

Means and Standard Deviations of Proportion Correct Scores by Content Category[a]: 1989

FORM	PEA	IAG	GEO	TRG	ALL	n
A	.580	.411	.462	.293	.483	3293
	.167	*.142*	*.128*	*.190*	*.177*	
B	.515	.399	.412	.376	.447	3258
	.168	*.137*	*.165*	*.128*	*.166*	
C	.566	.411	.481	.360	.495	3230
	.200	*.151*	*.147*	*.127*	*.182*	
D	.523	.403	.476	.364	.465	3202
	.174	*.133*	*.120*	*.058*	*.156*	
E	.572	.436	.508	.342	.501	3158
	.174	*.148*	*.209*	*.090*	*.185*	
F	.529	.388	.449	.336	.455	3153
	.165	*.161*	*.168*	*.128*	*.175*	
G	.583	.405	.460	.372	.487	3239
	.188	*.141*	*.151*	*.111*	*.182*	
H	.522	.413	.439	.384	.461	3091
	.150	*.140*	*.169*	*.201*	*.164*	
Average	.549	.408	.461	.353	.474	
SE[b]	.010	.005	.010	.010	.007	

[a] Standard deviations are in italics.

[b] Standard error of the average.

TABLE 3

Means and Standard Deviations of Proportion Correct Scores by Content Category[a]: 1990

FORM	PEA	IAG	GEO	TRG	ALL	n
G	.577	.401	.460	.377	.484	2695
	.189	*.142*	*.158*	*.113*	*.182*	
H	.517	.410	.432	.379	.455	2898
	.149	*.135*	*.168*	*.199*	*.162*	
I	.534	.390	.396	.293	.442	2648
	.168	*.172*	*.191*	*.115*	*.189*	
J	.561	.382	.423	.317	.459	2848
	.180	*.181*	*.154*	*.061*	*.190*	
K	.565	.394	.434	.352	.469	2838
	.195	*.183*	*.156*	*.124*	*.197*	
L	.563	.379	.395	.339	.454	2831
	.195	*.159*	*.168*	*.077*	*.195*	
M	.589	.382	.442	.310	.474	2791
	.150	*.172*	*.168*	*.084*	*.186*	
N	.572	.423	.438	.271	.476	2770
	.198	*.196*	*.194*	*.052*	*.209*	
O	.568	.381	.435	.285	.462	2728
	.241	*.132*	*.238*	*.145*	*.227*	
Average	.561	.394	.428	.325	.464	
SE[b]	.007	.005	.007	.013	.004	

[a] Standard deviations are in italics.
[b] Standard error of the average.

TABLE 4
Estimates of Universe Score Variance and Covariance Components: 1989

FORM A

PEA	IAG	GEO	TRG
.033			
.030	.033		
.034	.034	.044	
.019	.021	.023	.017

FORM B

PEA	IAG	GEO	TRG
.036			
.035	.039		
.030	.031	.030	
.029	.032	.028	.041

FORM C

PEA	IAG	GEO	TRG
.030			
.032	.037		
.035	.040	.049	
.027	.032	.038	.031

FORM D

PEA	IAG	GEO	TRG
.034			
.033	.039		
.035	.036	.042	
.027	.032	.032	.036

FORM E

PEA	IAG	GEO	TRG
.031			
.030	.037		
.027	.029	.028	
.035	.041	.037	.069

FORM F

PEA	IAG	GEO	TRG
.039			
.033	.033		
.031	.028	.031	
.015	.015	.016	.014

FORM G

PEA	IAG	GEO	TRG
.036			
.027	.027		
.030	.026	.031	
.033	.030	.033	.051

FORM H

PEA	IAG	GEO	TRG
.042			
.036	.032		
.033	.028	.025	
.035	.030	.029	.031

TABLE 5
Estimates of Universe Score Variance and Covariance Components: 1990

FORM G

PEA	IAG	GEO	TRG
.034			
.025	.025		
.029	.024	.030	
.031	.029	.032	.045

FORM H

PEA	IAG	GEO	TRG
.041			
.035	.031		
.031	.027	.023	
.034	.030	.028	.028

FORM I

PEA	IAG	GEO	TRG
.037			
.032	.030		
.029	.028	.028	
.016	.016	.015	.009

FORM J

PEA	IAG	GEO	TRG
.033			
.026	.022		
.031	.025	.032	
.020	.028	.023	.019

FORM K

PEA	IAG	GEO	TRG
.035			
.027	.024		
.030	.026	.032	
.024	.023	.026	.034

FORM L

PEA	IAG	GEO	TRG
.032			
.030	.031		
.026	.025	.025	
.030	.031	.029	.045

FORM M

PEA	IAG	GEO	TRG
.043			
.032	.027		
.034	.026	.031	
.031	.027	.029	.046

FORM N

PEA	IAG	GEO	TRG
.034			
.032	.038		
.026	.027	.025	
.025	.032	.025	.038

Table 5 (cont)

FORM O

PEA	IAG	GEO	TRG
.026			
.025	.030		
.025	.028	.030	
.021	.025	.027	.028

TABLE 6

Mean Estimates of Universe Score Variance and Covariance Components with their Standard Errors[a]: 1989

	PEA	IAG	GEO	TRG
PEA	.035 *.001*			
IAG	.032 *.001*	.034 *.002*		
GEO	.032 *.001*	.031 *.002*	.035 *.003*	
TRG	.028 *.002*	.029 *.002*	.030 *.002*	.038 *.006*

[a]Standard errors of the means.

TABLE 7

Mean Estimates of Universe Score Variance and Covariance Components with Their Standard Errors[a]: 1990

	PEA	IAG	GEO	TRG
PEA	.035 *.002*			
IAG	.029 *.001*	.029 *.002*		
GEO	.029 *.001*	.026 *.000*	.029 *.001*	
TRG	.026 *.002*	.026 *.002*	.026 *.002*	.032 *.004*

[a]Standard errors of the means.

TABLE 8

Correlations Among Universe Scores: 1989

	PEA	IAG	GEO	TRG
PEA	1.0			
IAG	.914	1.0		
GEO	.910	.901	1.0	
TRG	.768	.811	.822	1.0

TABLE 9

Correlations Among Universe Scores: 1990

	PEA	IAG	GEO	TRG
PEA	1.0			
IAG	.929	1.0		
GEO	.916	.916	1.0	
TRG	.762	.843	.850	1.0

TABLE 10

Estimates of Variance Components for Item Effects: 1989

FORM	PEA	IAG	GEO	TRG
A	.029	.021	.018	.048
B	.029	.020	.029	.022
C	.042	.024	.023	.022
D	.032	.019	.015	.004
E	.032	.023	.047	.011
F	.028	.027	.030	.022
G	.037	.021	.025	.016
H	.023	.021	.031	.054
Average	.031	.022	.027	.025
SE[a]	.002	.001	.003	.006

[a] Standard error of the average.

TABLE 11

Estimates of Variance Components for Item Effects: 1990

FORM	PEA	IAG	GEO	TRG
G	.037	.021	.027	.017
H	.023	.019	.030	.053
I	.029	.031	.039	.018
J	.034	.035	.026	.005
K	.040	.035	.026	.021
L	.040	.027	.030	.008
M	.023	.031	.030	.009
N	.041	.041	.040	.004
O	.061	.018	.061	.028
Average	.036	.029	.034	.018
SE[a]	.004	.003	.004	.005

[a] Standard error of the average.

TABLE 12

Estimates of Variance Components for Residual Effects: 1989

FORM	PEA	IAG	GEO	TRG
A	.183	.189	.188	.154
B	.186	.182	.186	.177
C	.175	.187	.179	.183
D	.185	.184	.194	.193
E	.184	.188	.178	.148
F	.183	.179	.188	.193
G	.172	.194	.194	.170
H	.185	.191	.193	.163
Average	.182	.187	.187	.173
SE[a]	.002	.002	.002	.006

[a] Standard error of the average.

TABLE 13

Estimates of Variance Components for Residual Effects: 1990

FORM	PEA	IAG	GEO	TRG
G	.174	.195	.193	.177
H	.187	.193	.194	.168
I	.184	.179	.174	.185
J	.180	.181	.188	.194
K	.173	.182	.190	.178
L	.176	.180	.186	.174
M	.177	.180	.187	.161
N	.171	.168	.184	.157
O	.162	.189	.160	.155
Average	.176	.183	.184	.172
SE[a]	.002	.003	.004	.004

[a] Standard error of the average.

Figure 1

Content Area Means from Tables 1 and 2

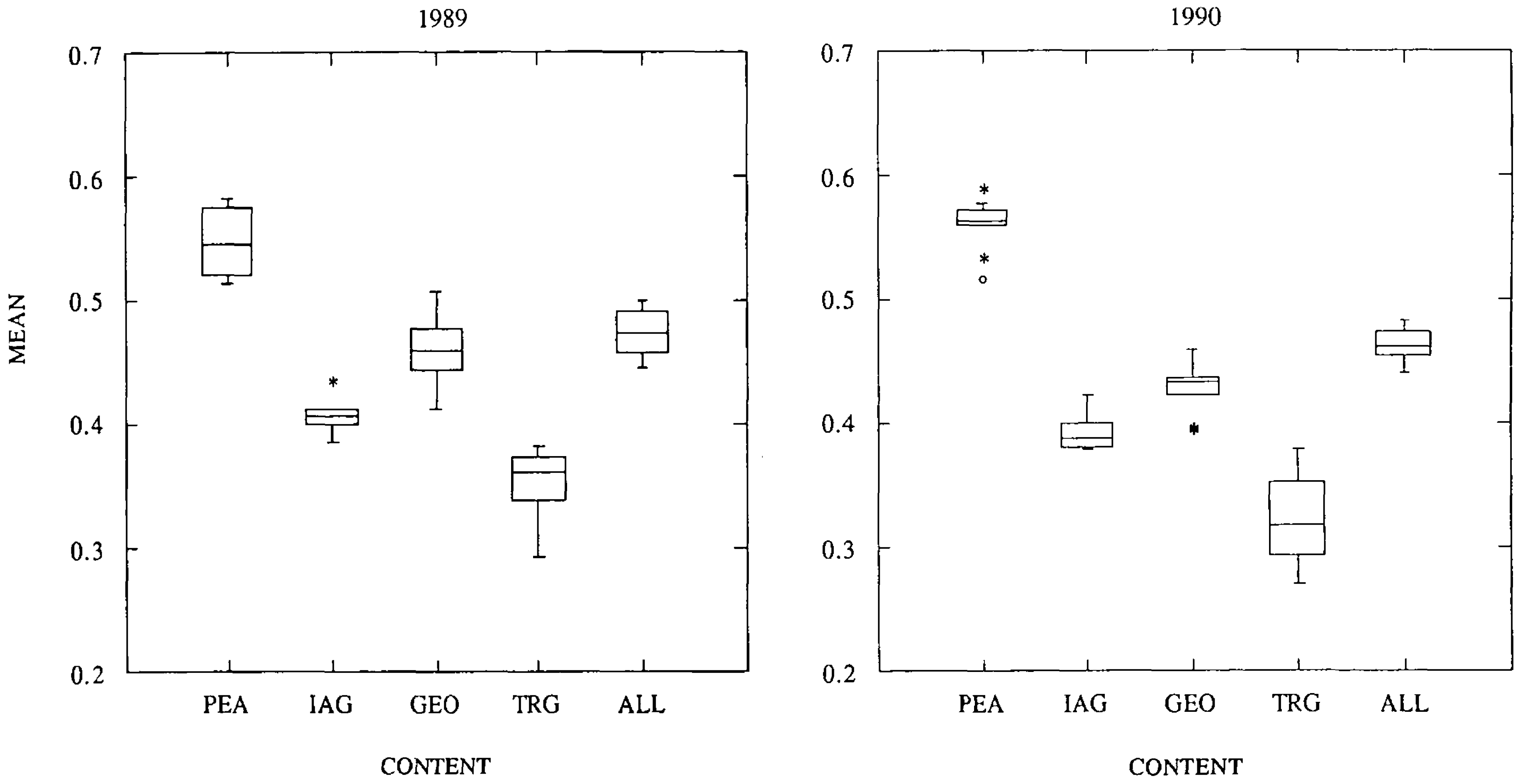
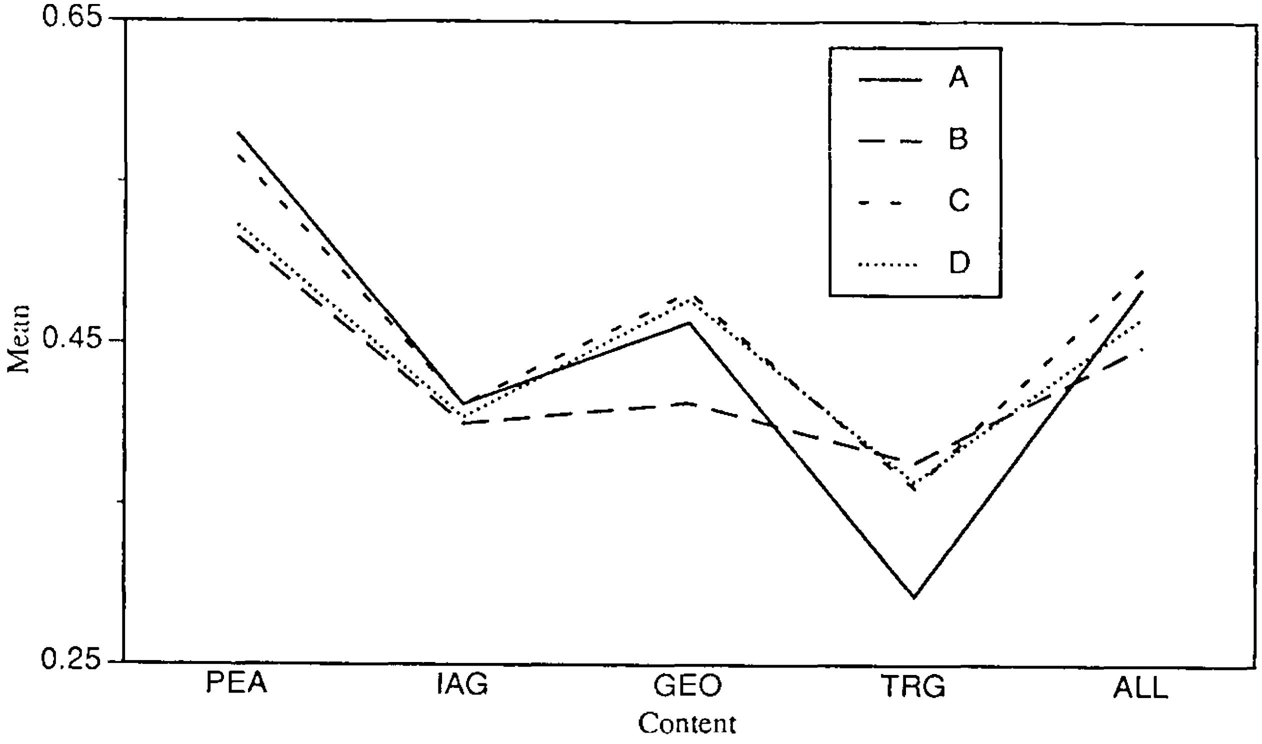


Figure 2

1989 Content Area Means from Table 2



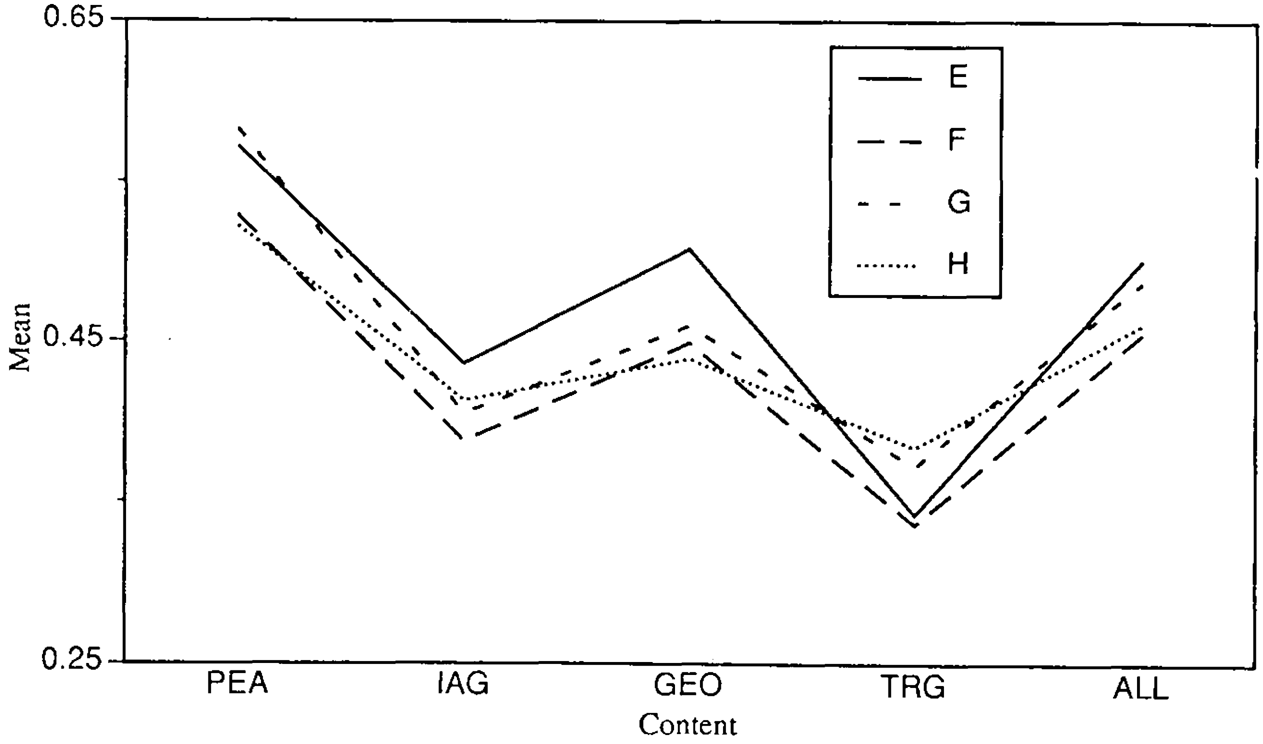
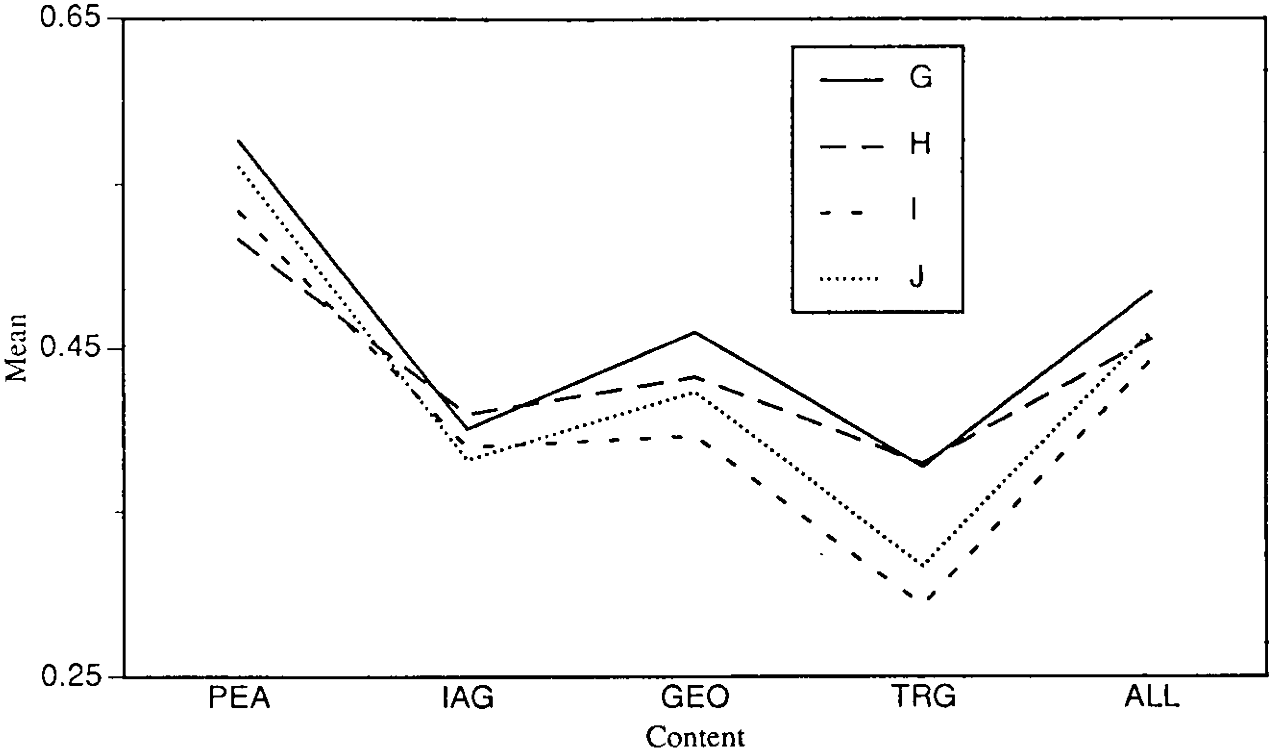
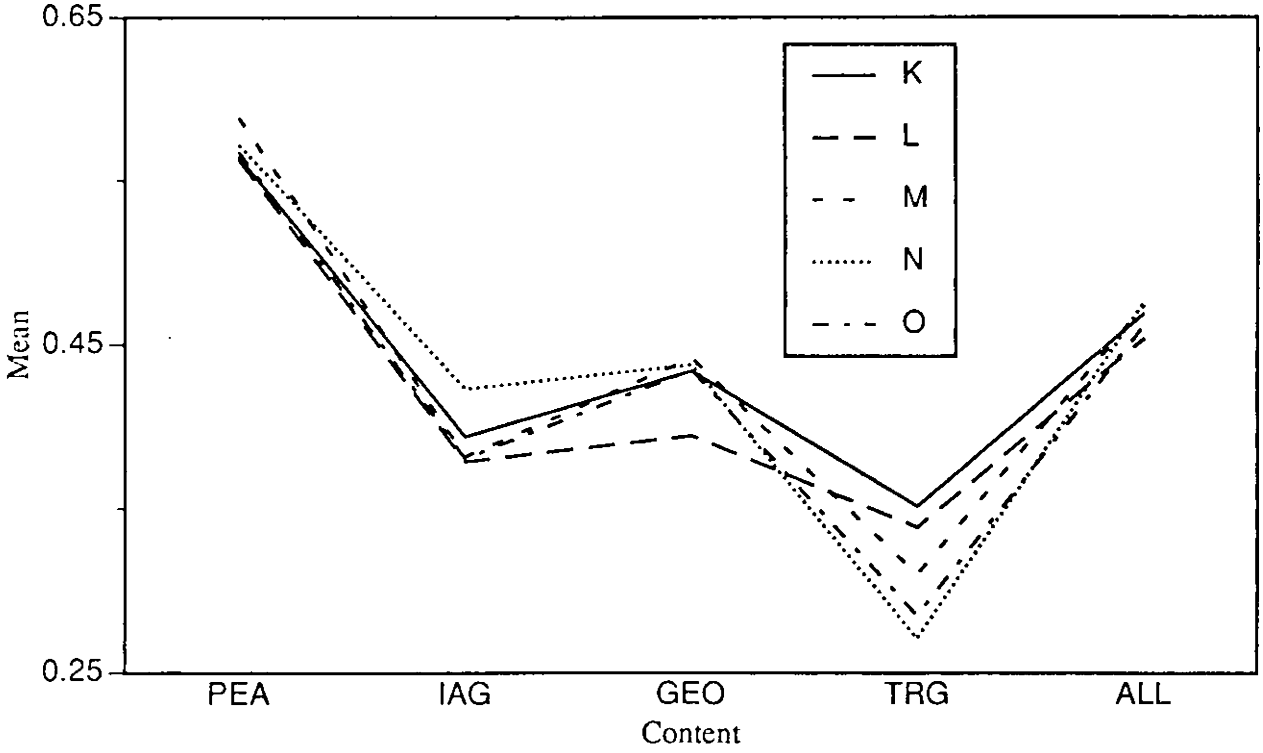


Figure 3
1990 Content Area Means from Table 3


Mean
PEA
IAG
GEO
TRG
ALL
Content
G
H
I
J


Mean
PEA
IAG
GEO
TRG
ALL
Content
K
L
M
N
O