DOCUMENT RESUME

ED 320 934 TM 015 161

AUTHOR	Ackerman, Terry
TITLE	An Evaluation of the Multidimensional Parallelism of the EAAP Mathematics Test.
PUB DATE	Apr 90
NOTE	34p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; College Bound Students; College Entrance Examinations; *High School Students; Item Response Theory; *Mathematics Tests; Scores; *Standardized Tests; *Test Format; Test Items
IDENTIFIERS	*ACT Mathematics Usage Test; Dimensionality (Tests); *Parallel Test Forms

ABSTRACT

The issue of parallel forms is of paramount importance for producers of standardized tests. With increasing emphasis being placed on standardized test results it is necessary that each student achieve the same standard score regardless of the form he or she was administered. In the case of the American College Testing (ACT) Assessment Program, five different forms are administered annually. Each form is purported to be m asuring tne same skills with the same degree of accuracy. The parallelism of two forms of the Mathematics Usage Test from the Enhanced ACT Assessment Program (EAAP) that were administered to 2,500 examinees during the fall of the 1989-90 academic year was examined. Unlike traditional studies of parallelism, this study illustrated several new techniques that can be used to study parallelism from a multidimensional item response theory (MIRT) perspective. Based on the MIRT analysis, the tests did not appear to be multidimensionally parallel, although conditions that could alter this conclusion are discussed. Recognizing the fact that most tests assess multiple abilities, it is proposed that the methodology can provide a basis for future research. Nine tables and eight figures, as well as an appendix that illustrates two test items, are included. (Author/SLD)

An Evaluation of the Multidimensional Parallelism of the EAAP Mathematics Test

Terry Ackerman
University of Illinois

Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1990.

Abstract

The issue of parallel forms is of paramount importance for producers of standardized tests. With increasing emphasis being placed on standardized test results it is necessary that each student achieve the same standard score regardless of the form he or she was administered. In the case of the ACT Assessment Program, five different forms are administered annually. Each form is purported to be measuring the same skills with the same degree of accuracy. This study examines the parallelism of two forms of the Mathematics Usage Test that were administered during the fall of the 1989-90 academic year. These forms are part of the revised or enhanced ACT Assessment. Unlike tradition studies of parallelism this study illustrates several new techniques which can be used to study parallelism from a multidimensional item response theory perspective. Recognizing the fact that most tests assess multiple abilities, it is hoped that the methodology presented in this study can provide a basis for future research in the assessment of multidimensional parallelism.

In 1987-88 the Mathematics Usage Test of the ACT Assessment Program underwent a substantial revision. The former test contained 40 items that were to be completed in 50 minutes. The test did not measure achievement in any particular high school course, but rather sampled content from courses taught in grades seven through eleven. Items were classified into six categories: Arithmetic and Algebraic Operations; Arithmetic and Algebraic Reasoning; Geometry; Intermediate Algebra; Number and Numeration Concepts; and, Advanced Topics. The new Mathematics Usage Test from the Enhanced Act Assessment Program (EAAP) contains 60 items and has an administration time of 60 minutes. Based upon an analysis of information from three sources--published educational objectives for grades 7 - 12, a review of state approved textbooks in courses for grades 7 - 12, and, consultation with curriculum experts from both secondary and postsecondary levels -- a new dual classification of categories is used. Items are classified into four content areas and three skill areas. The new content areas are Pre-Algebra (PA), Elementary Algebra (EA), Intermediate Algebra and Coordinate Geometry (IA/CG), Plane Geometry (PG) and Trigonometry (T). The skill categories include basic skills, applications, and analysis. Unlike the previous Mathematics Usage Test on which a single score was reported, the new test will report three subtest scores, PA/EA, IA/CG, and PG/T, as well as a total score. The first administration of the EAAP was in the fall of 1989.

As with all new standardized tests which employ several forms the question of parallelism is paramount. Because of the importance placed upon EAAP scores, it is necessary that all forms be as similar as possible. High school students taking different forms on different test dates need to be assured that all forms are equal in difficulty and precision and that one would expect to achieve the same score regardless of which form was taken. Reckase, Davey, and Ackerman (1989) examined the parallelism of different forms of the former Mathematics Usage Test. They used a set of new multidimensional item response theory (MIRT) procedures to determine whether the test forms were multidimensionally parallel. The purpose of this paper is to apply the same procedures to the first two forms of EAAP Math (Forms 39B & 39F) to determine whether the forms are parallel in a multidimensional sense even though that was not a specific requirement in the test construction process.

Design of the study

Reckase, Davey, and Ackerman (1989) outlined a five step procedure that could be used to assess the multidimensional parallelism of a test which is assumed to be measuring multiple dimensions. The steps are: (1) determine the dimensionality of the latent ability space and substantively identify the m - dominant dimension(s); (2) calibrate the m -

dimensional structural (item) MIRT parameters; (3) rotate the solutions so that the reference composite (Wang, 1987) for the two forms coincide; (4) compare the amount of information that each form provides over the m-dimensional ability space; and (5) compare the expected true scores for each form over the m-dimensional ability space. If two forms are multidimensionally parallel then each form defines the same latent ability space such that a randomly chosen individual would have the same expected true score, no matter which form the individual takes, measured with the same degree of precision, no matter where the individual lies in the m-dimensional ability space.

To assess the dimensionality of the latent space each form was factor analyzed at both the subtest level and at the item level. At the subtest level a principal factor analysis was performed on the Pearson product-moment correlations between the number correct scores for each of the three subtests. At the item level, a principal factor analysis was performed on the matrix of interitem tetrachoric correlations. The principal factor analysis was performed using the largest correlation with the item as the communality estimate. Results were substantively reviewed to identify or define the ability dimensions. Similar to the Reckase et al. study, a two-dimensional solution was thought to span the latent ability space defined by the two EAAP Mathematics Test forms.

The MIRT calibration was performed using the NOHARM program (Fraser, 1983). This program estimates the structural parameters for n items using the multidimensional normal ogive model given by

$$P(u_{ij} = 1 \mid a_i\, d_i\, c_i\, \theta_j) = c_i + (1 - c_i) \int_{\gamma_{ij}}^{\infty} \phi(a_i'\theta_j + d_i)d\theta$$

where a_i is the nx2 matrix of discrimination parameters;
d_i is a vector of length n of difficulty parameters;
c_i is a vector of length n of guessing parameters;
θ_j is the two-dimensional latent ability vector.
$\phi(\,)$ is the normal density function

The structural parameters of this model can be converted to estimates of the logistic compensatory model by dividing them by 1.7.

NOHARM estimates only the discrimination and difficulty parameters; c_i is fixed by the user. It is assumed that the guessing parameter in the unidimensional IRT model is similar to the c_i parameter in the MIRT models. Thus, the c_i vector for each form was obtained from a unidimensional IRT calibration using the computer program BILOG (Mislevy & Bock, 1984).

In two-dimensional estimation there exists a rotational indeterminacy. NOHARM attempts to remove this indeterminacy, in part, by assuming that the first item loads only on the first ability dimension. That is, the user is required to select an item which most typifies the first ability dimension and place its vector of responses first. NOHARM sets the a_2 parameter of item one to zero. Based upon the factor analysis results and a substantive review of the test items two similar items, one from each form, were selected to define the first dimension. In a sense, this could be considered a "preequating" of the two forms. For the purposes of this study two plane geometry application items, one from each form, were selected to "anchor" the first dimension for their respective forms. It was initially thought that the first dimension could be characterized as a spatial problem solving dimension and that algebraic symbol manipulation defined the second dimension.

To further align the two dimensional ability axes, the reference composite (Wang, 1987) was determined for each test. The reference composite, determined from the first eigenvector of the $a'a$ matrix, represents the line onto which the two-dimensional ability plane would be mapped during a unidimensional IRT calibration run. The location of this line in the two-dimensional ability plane helps the researcher determine the composite of abilities that is being represented by a unidimensional ability estimate. Once the forms have been properly aligned their information and test characteristic surfaces can be determined and compared between forms.

Data

The data used in this study was obtained from an equating administration of the EAAP tests in the fall of 1989. A random sample of 2500 response vectors for both Forms 39B and 39F were selected. Because each form was distributed in a spiralled fashion, the two groups of examinees are assumed to be randomly equivalent on ability.

Table 1 summarizes the five content specifications for each form. Each form is constructed to meet a specified distribution of p-values and biserial correlations. Items are selected if their p-values range from .3 to .8 and have a biserial correlation greater than .3. When forms are constructed content considerations are given top priority, followed by item difficulty specifications and then discrimination specifications.

Insert Table 1 about here

Attempts are made to match the mean difficulty of each subtest across all forms. That is, the mean p-value for the PA/EA items would be matched for each form; the mean p-value for the IA/CG items would be matched for each form, etc.

Results

A standard item analysis was performed at the test and subtest levels for both forms. From a Classical Test Theory perspective if forms are truly parallel the mean and standard deviation of the obtained scores would be equal for each subtest across forms. Likewise the reliabilities would be similar for each subtest across forms. The results of the item analyses are reported in Table 2.

Insert Table 2 about here

Each of the three subtests in 39F appear to be easier than their counterparts in 39B. This fact is supported both by the larger subtest mean scores and larger mean p-values for 39F when compared with 39B. Discrimination values however, favor 39B. In each subtest the mean biserial correlation coefficient is larger in 39B than in 39F. KR-20 reliability estimates are slightly larger for each subtest in Form 39B. The IA/CG subtest appeared to be the most difficult of the three subtests as examinees answered less than half of the items correctly on the average in both forms. The PA/EA subtest was the easiest subtest in both forms.

A more detailed analysis of how such differences occur can be seen in Tables 3 and 4 which display the distribution of p-values and biserial correlations by subtest for each form. The marginal distributions for each p-value range reveal almost twice as many 39F items in the p-value range of .60 - .79, and four times as many 39F items in the .80 - 1.00 range.

Insert Tables 3 & 4 about here

Differences are not quite as pronounced in the marginal distributions for each biserial correlation range, although it appears clear that 39B contains more items in the higher discrimination ranges than 39F. Further evidence that 39B is more difficult can be seen in the cumulative percentage frequency polygon for each group displayed in Figure 1. The 39F examination group, represented by the dotted line, clearly out-performed their 39B counterparts over the entire raw score range.

Insert Figure 1 about here

To examine the dimensionality and relationship among the subtests for each form the Pearson product-moment correlations between the number correct scores for each subtest and total test were computed. These are given in Table 5

Insert Table 5 about here

There is a remarkable similarity between the inter-subtest correlations within each form. The subtests for 39B show a stronger linear relationship on the average than subtests for 39F. (≈ .78 to .70, respectively). In both forms the IA/CG subtest correlates most highly with the total test score. The magnitude of all correlations are uniformly large. Table 6 shows the results of the principal factor analysis of the correlation values listed in Table 5. These results suggest that each form is dominated by a strong first factor that accounts for 81% of the variance in 39B and 76% of the variance in 39F. The IA/CG subtest loads highest on the first factor for each form. The largest percentage of variance that is not explained by the first factor for both forms indicates that the two forms are definitely not unidimensional.

Insert Table 6 about here

Confirmation of the multidimensionality of each form was provided by the principal factor analysis of the inter-item tetrachoric correlation matrix. Before the factor analysis was conducted the matrix was corrected for guessing (cf. Carroll, 1945) using the BILOG guessing parameter estimates. The factor analysis results are reported in Table 7.

Insert Table 7 about here

These results suggest that there are at least two identifiable factors in each test. The first two factors in 39B account for 54.20% and 7.18% of the total variance respectively. The first two factors in 39F account for 46.73% and 6.1% respectively. The third factor accounts for about 3.6% in both forms. The rate at which the eigenvalues decrease does not provide a clear indication as to how many factors are significant. Because two factors were substantively identifiable, it was decided to analyze the two tests from a two-dimensional perspective.

The test forms were scrutinized to identify two items, i.e., one per form, that appeared to be measuring the same spatial ability. The items that were selected were Item 56 from 39B and Item 51 from 39F. Each of these items are classified as a PG application item. These items are listed in Appendix A. The response vector for each of these items was moved to the first item position for the calibration run of each form. This process enabled the estimation program NOHARM to fix or "define" the first dimension in each calibration run as spatial ability. It was hypothesized that the second dimension would be defined by an algebraic symbol manipulation skill and thus would be typified by PA/EA type items, although there was no such constraint imposed upon the NOHARM program.

Once the two forms were calibrated the reference composite was then determined for each subtest and the total test. These results are provided in Table 8. The reference composite for the total forms are quite similar. The angle between the positive θ_1 axis and the reference composite for 39B is 40.53° and for 39F, 38.73°. Thus it appears that the two "anchor items" measured somewhat similar skills. The order of the reference composites from the θ_1 axis was expected to be PG/T, IA/CG and PA/EA, respectively. This was not the case. The order of the reference composites for 39B starting from the positive $\theta 1$ axis was PG/T, PA/EA, and IA/CG. The order of the reference composites for 39F starting from the positive θ_1 axis was IA/CG, PG/T, and PA/EA.

Insert Table 8 about here

Thus the IA/CG subtest "flip-flopped" about the PG/T and PA/EA reference composites. That is, in 39B it lies closest of the three subtest reference composites to the θ_2 axis and in 39F it lies closest to the θ_1 axis. These results were somewhat puzzling so two additional analyses were conducted in an effort to explain the curious results.

The first analysis was simply to look at the item classifications and see if the number of items in each skill classification differed between the two forms. (The number of items in each content classification is identical because of the construction procedures.) It was found that 39B contained eleven items classified as basic skill items, five as application items, and two as analysis items. This was quite different from 39F which had seven basic skill items, nine application items, and two analysis items. The items in the other subtests were almost identical in the number of items classified into each skill category. This result caused some suspicion about the true definition of each of the dimensions. That is, perhaps the dimensions should have been classified according to skill levels and not according to content classifications.

The second analysis that was conducted to provide insight into the different ordering of the subtest reference composites was to plot the item vectors. In an item vector plot each item is represented by a vector. The length of the vector represents the amount of discrimination in the direction of the vector. The tail of the vector is orthogonal to the $p = .5$ equiprobability contour line through the θ_1-θ_2 ability plane. The angle the vector makes with the θ_1 axis represents the direction of maximum discrimination. This direction helps the researcher to identify the θ_1-θ_2 composite which is being measured most accurately by the item. Using these plots it could be determined if the reference composite was representative of the subtest item vectors or whether the reference composite was influenced by outlier items which had extremely large discrimination values. The vector plots are displayed separately by subtest by form in Figure 2 and Figure 3.

Insert Figures 2 & 3 about here

To provide a more accurate framework for comparison these plots were made after the solution of 39F was rotated so that its reference composite coincided with the reference composite of 39B. Note this required only a slight orthogonal rotation of only 1.8°.

These results raised more questions than answers. First of all, only the PG/T subtest for 39F appeared to contain an extremely heterogenous group of item vectors. Item 51 has a vector which lies only 1.8° above the positive θ_1 axis, whereas six items lie only a few degrees to the left of the negative θ_2 axis. The remainder of the items for this subtest lie close to a 45° angle from the positive θ_1 axis. Thus this particular subtest also clearly demonstrates a confounding of difficulty and dimensionality. That is, the easier items measure θ_2, whereas the difficult items tend to measure more of θ_1. Of the items which lie closest to θ_2 half of them are classified as basic skill items, half of them are application items. Further analysis of the IA/CG item vectors for 39F revealed that the items which were closest to the θ_2 axis tended to be classified as basic skill items. This was true for both forms.

The average item parameter estimates, average distance from the origin to the p = .5 equiprobability contour (denoted by D), and the average item vector angle with the positive θ_1 axis (denoted by α), are given for each subtest and the total test for each form in Table 9.

Insert Table 9 about here

It was interesting to see what effect the IA/CG differences would have upon the remaining analyses to investigate the multidimensional parallelism of the two forms. For an item the amount of information on the two-dimensional ability plane is a function of the item's discrimination vector. The location of the information is a function the item's difficulty and the angular composite of θ_1-θ_2 that is being measured. The item information values in ten directions from 0° to 90° (in increments of 10°) were computed for 49 points spread equally over the θ_1-θ_2 ability plane (cf. Reckase, 1986) This was done for each subtest in each form. Vectors representing the difference, 39F - 39B, were plotted at each of the 49 (θ_1,θ_2) points. Vectors were colored red if they were positive and black if they were negative. These information difference "clamshell" plots are shown in Figure 4.

Insert Figure 4 about here

The PA/EA plot shows that 39B consistently provides more information over the entire two-dimensional ability plane except for minute differences in the 90° direction in the upper portion of the first quadrant. This suggests that 39B has more precision which agrees with the higher KR-20 value in Table 2.

The IA/CG plot shows that the two tests are actually measuring the two ability dimensions differently. The first dimension is measured more accurately by 39F and the second dimension is measured more accurately by 39B. This result coincides with the average discrimination parameter estimates for the two forms provided in Table 9. This may be due to the fact that 39B contains more basic skill items and 39F contains more application items. The two forms appear to be measuring equally well in the 45° direction near the center of the ability plane.

The PG/T plot shows that 39F clearly measures better than 39B in the 45° to 90° sector throughout most of the ability plane except in the upper portion of the first quadrant. This also coincides with the fact that 39F PG/T items have, on the average, larger a_2 values as listed in Table 9.

The plot for the total test information difference reveals a diagonal band running from $\theta_1 = -3.0, \theta_2 = 3.0$ to $\theta_1 = 3.0, \theta_2 = -1.0$ in which 39B provides more information especially near the second dimension. The differences are noticeably smaller throughout the rest of the ability plane.

The last test of multidimensional parallelism was to examine the difference between the test characteristic surfaces for each subtest and the total test by form. The test characteristic surface represents the expected true score for each (θ_1,θ_2) point over the ability plane. If the two tests are multidimensionally parallel the surfaces would coincide. These surface difference plots are shown in Figures 5 - 8.

Insert Figures 5 - 8 about here

The plot of the difference between the test characteristic surface of the PA/EA subtests of 39F - 39B suggests that a higher true score would be expected on 39F for examinees lying in the second and third quadrants of the ability plane. Examinees would be expected to do better on 39B if they were in the fourth quadrant.

The IA/CC plot reveals somewhat the opposite results of the PA/EA plot. The differences between the forms for this subtest are also not as pronounced as in the PA/EA plot. Examinees located in the top third quadrant of the ability plane would be expected to perform better on 39B, whereas examinees in the lower corner of the fourth quadrant would be expected to achieve a higher score on 39F.

The PG/T plot is very similar to the PA/EA plot although the differences are of the same magnitude as the IA/CG differences. Examinees located in the upper corner of the second quadrant would be expected to perform better on 39F and those in the lower corner of the fourth quadrant better on 39B.

The plot of the difference between the 39F and 39B tests taken as a whole reveals that throughout most of the ability plane an examinee would be expected to perform better on 39F than on 39B. This difference increases from a difference of two to three true score units in he first and fourth quadrants; to about six units near the origin; to about ten units in the extreme second quadrant. Table 2 supports this finding because the mean score of 39F was almost four raw score units higher than that obtained for 39B.

Discussion

The purpose of this paper was to summarize a series of analyses that examined the multidimensional parallelism of two EAAP math forms 39B and 39F. It should be emphasized that the forms were constructed to be parallel based upon classical item statistics such as p-values and biserial correlations, and that the multidimensional parallelism analysis is from a MIRT perspective. However, both the classical analysis and the MIRT analysis do seem to be in agreement. Interestingly the MIRT analyses also seem to be sensitive to the different types of item content, and even more so to the differences in the skill classifications.

Results of the classical item analysis suggest that 39B is noticeably more difficult than 39F; in fact, about equally more difficult in each of the subtests. Yet the similarity of subtest reliability coefficients suggest that the degree of measurement precision is the same for both forms.

The results of this study illustrate one of the advantages of analyzing tests using MIRT. It has been suggested (Traub, 1983) that tests are multidimensional to some extent. Classical test theory (like unidimensional IRT) does not address this issue. Differences in mean test scores are assumed to reflect differences in abilities of the groups of examinees and/or the difficulty of the items. The hypothesis that groups perform differently because the tests are measuring different skills can never be tested directly. However, MIRT analyses enable researchers to examine the multidimensionality of items in detail, and to study the similarities and differences between the skills that groups of items measure. One example of this is the MIRT concept of a reference composite. If items are measuring the same composite of skills they should have similar reference composites. MIRT analysis of 39B and 39F revealed that the IA/CG reference composites are noticeably different. This suggests that a different composite of abilities may be measured at the subtest level (especially in the IA/CG subtest). Surprisingly these differences seem to balance themselves as the total test reference composite for each form differ by slightly more than one degree.

Examining the information provided by each form at the subtest level also reveals differences. The sign of the difference does not appear to be consistent over the two-dimensional ability plane from one subtest to the other. The differences seem to echo the classical results which suggest only slight SEM subtest differences between the two forms. Like the reference composites, the information differences tend to cancel themselves out when considered at the total test level for much of the ability plane.

Noticeable dissimilarities are seen in the plots of the differences of the test characteristic surfaces. Examinees in the upper right portion of the second quadrant are expected to do better on the PG/T and PA/EA subtests for 39F but poorer on the IA/CG subtest for 39B. At the total test level the differences which appear in each of the subtests are simply summed. Thus, the magnitude of the expected true scores differences is much greater, especially in the extreme second quadrant.

It is interesting that the information and test characteristic surface criteria appear to be in disagreement. Although such disagreement is dictated by the results presented in Table 2 and Figure 1. The information differences should be small since they relate closely to the KR-20 values, which are almost identical for each form. The test characteristic surfaces provide a crude two-dimensional version of the cumulative percentage frequency polygon, albeit on a different scale. The cumulative ogive curve for 39F is higher than the curve for 39B throughout the entire score scale. Therefore it should be expected that the test characteristic surface would rise more sharply for 39F than its 39B counterpart.

Thus, based upon the MIRT analyses the tests do not appear to be multidimensionally parallel. However, such a conclusion needs to be qualified with several important considerations. First of all, it remains to be determined as to what magnitude of difference in information or test characteristic surface constitutes an important difference. By examining the two forms using MIRT, we have magnified the differences. Thus we need to translate information differences on the two-dimensional ability plane into a more meaningful metric. One suggestion might be to compute the two-dimensional equivalent of relative efficiency and then translate any differences into a metric such as "n number of items need to be added to Form X to make it as informative as Form Y". Any examination of differences should also be weighted by an estimation of the density of the examinee population.

Secondly, we made the assumption that the forms were essentially two-dimensional. This could have been a mistake. That is, if the tests were in fact three dimensional and the three dimensions could clearly be identified, the results reported in this paper could be a distortion of the true solution. If three dimensions are required to completely define the latent ability space than comparisons would have to be made by examining the differences of solutions mapped onto reference composite planes instead of reference composites lines. Such an analysis may, in part, help explain the difference in the ordering of the reference composites. However, as the dimensionality of the latent space increases we must be careful not to sacrifice utility and understanding for "completeness" of solution. The inability to decide the appropriate dimensionality is a weakness of the MIRT approach.

Thirdly, the assumption has been made from the outset that the groups are randomly equivalent in ability. Although one can never tell, it must be realized that five percent of the time the two randomly selected groups will be significantly different in ability.

It is very important for the reader to realize that differences between 39B and 39F are minimized through equating before any results are reported. The analyses reported in this study were done before any equating was performed. Thus, despite some of the differences that were identified, after equating it is expected that an examinee who may have taken 39B would receive the same standard score that he would have achieved on 39F.

Obviously research of the examination of multidimensional parallelism will continue. Ideally we would like to find analyses that can distinguish between different types of content or skill levels and can help us predict examinee performance to a greater degree than is now possible. Perhaps the "magnifying" power of MIRT analyses can help us achieve this greater level of a uracy. We must, however, be able to understand and interpret for the practitioner, what we are looking at and why it is important.

References

The American College Testing Program (1989). Preliminary Technical Manual for the Enhanced ACT Assessment. Iowa City, 'A: Author.

Carroll, J.B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10, 1-19.

Fraser, C. (1983). NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Mislevy, R.J., & Bock, R.D. (1984). BILOG: Item Analysis and test scoring with binary logistic models. Scientific Software, Inc. Mooresville,

Reckase, M.D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M.D., Davey, T., & Ackerman, T.A. (1989, March). Similarity of the multidimensional space defined by parallel forms of a mathematics test. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Traub, (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Wang, M. (1987, April). Estimation of ability parameters from response data to items that are precalibrated with a unidimensional model. Paper presented at the annual meeting of the American Educational Research Association. Washington D.C.

Table 1. Specifications for the ACT Mathematics Test

Description of the test. The Mathematics Test is a 60-item, 60-minute test that is designed to assess the mathematical skills that students have typically acquired in courses taken up to the beginning of grade 12. The test presents multiple-choice items that require students to use their reasoning skills to solve practical problems in mathematics. Knowledge of basic formulas and computational skills are assumed as background for the problems, but complex formulas and extensive computation are not required. The material covered on the test emphasizes the major content areas that are prerequisite to successful performance in entry-level courses in college mathematics.

The items included in the Mathematics Test cover three skill areas: (1) basic skills, (2) application, and (3) analysis. Basic skills includes items that can be solved by performing a familiar sequence of operations in a familiar setting. Application items can be solved by performing a familiar sequence of operations, but the solution will not be routine. Analysis items require a student to know why the familiar sequence of operations yields a solution, under what conditions it will not yield a solution, or how to examine all the cases that can arise within the restrictions stated in the stem of the item.

Four scores are reported for the ACT Mathematics Test: a total test score based on all 60 items, a subscore in Pre-Algebra/Elementary Algebra based on 24 items, a subscore in Intermediate Algebra/ Coordinate Geometry based on 18 items, and a subscore in Plane Geometry/Trigonometry based on 18 items.

Content of the test. Items are classified according to the five content categories. These categories and the approximate proportion of the test devoted to each are given below.

Content Area	Proportion of Test	Number of Items
Pre-Algebra and Elementary Algebra	.40	24
Intermediate Algebra and Coordinate Geometry	.30	18
Plane Geometry	.23	14
Trigonometry	.07	4
Total	1.00	60

a. **Pre-Algebra.** Items in this category are based on operations with whole numbers, decimals, fractions, and integers. They also may require the solution of linear equations in one variable.

b. **Elementary Algebra.** Items in this category are based on operations with algebraic expressions. The most advanced topic in this category is the solution of quadratic equations by factoring.

c. **Intermediate Algebra and Coordinate Geometry.** Items in this category are based on graphing in the standard coordinate plane or on other topics from intermediate algebra such as operations with integer exponents, radical expressions and rational expressions, the quadratic formula, linear inequalities in one variable, and systems of two linear equations in two variables.

d. **Plane Geometry.** Items in this category are based on the properties and relations of plane figures.

e. **Trigonometry.** Items in this category are based on right triangle trigonometry, graphs of the trigonometric functions, and basic trigonometric identities.

Table 2. Summary statistics of the EAAP Mathematics Forms 39B and 39F.

	FORM	$\bar{x}$	s_x	$\bar{p}$	s_p	$\bar{r}_{bis}$	s_r	KR-20	SEM
PA-EA	B	12.57	4.88	.52	.17	.50	.13	.81	2.09
(24)*	F	13.84	4.68	.58	.17	.46	.07	.80	2.13
IA	B	6.99	3.92	.39	.14	.51	.10	.80	1.75
(18)*	F	7.94	3.90	.44	.15	.48	.09	.78	1.83
PG-T	B	7.26	3.72	.40	.16	.45	.11	.79	1.70
(18)*	F	8.53	3.70	.47	.20	.36	.10	.77	1.77
TOTAL	B	26.83	11.57	.45	.17	.49	.12	.92	3.27
(60)*	F	30.32	11.01	.51	.19	.47	.10	.91	3.30

*Number of items.

Table 3. p-value distribution by subtest and total test.

		Range				
Subtest	Form	.00-.19	.20-.39	.40-.59	.60-.79	.80-.100
PA/EA	B	1	5	11	6	1
	F		4	9	9	2
IA/CG	B	1	8	8	1	
	F	1	5	9	3	
PG/T	B	2	5	9	2	
	F	8	4	4	2	
Total	B	4	18	28	9	1
	F	1	17	22	16	4

Table 4. Biserial distribution by subtest and total test.

		Range				
Subtest	Form	.00-.19	.20-.29	.30-.39	.40-.49	.50-.100
PA/EA	B	1		4	7	12
	F		1	2	14	7
IA/CG	B			4	2	12
	F		1	3	6	8
PG/T	B	1		2	9	6
	F	1	3		3	11
Total	B	2		10	18	30
	F	1	5	5	23	26

Table 5. Pearson product-moment correlations between subtests.

			Subtest		
		Form	IA/CG	PG/T	Total
	PA/EA	B	.77	.78	.91
		F	.71	.71	.89
Subtest	IA/CG	B		.78	.92
		F		.70	.91
	PG/T	B			.85
		F			.81

Table 6. Eigenvalues and first factor loadings for each subtest for each math form.

Eigenvalue	Form B	Form F
1	2.42	2.29
2	.37	.47
3	.21	.24

Subtest	First Factor Loading	
PA/EA	.89	.86
IA/CG	.93	.92
PG/T	.87	.84

Table 7. The first five eigenvalues from a principal factor analysis of the interitem tetrachoric correlation matrix for Form 39B and 39F.

	Form	
Eigenvalues*	B	F
1	32.52	28.04
2	4.31	3.66
3	2.20	2.19
4	2.00	1.77
5	1.77	1.55

*Note: Form 39B had 13 eigenvalues greater than 1.0; Form 39F had 12 eigenvalues greater than 1.0.

Table 8. The angle of the reference composite with the θ_1 axis for each subtest.

	Form	
Subtest	B	F
PA/EA	43.69	49.08
IA/CG	48.09	35.11
PG/T	27.13	40.53
Total	40.53	38.73

Table 9. Summary statistics of the two-dimensional MIRT item parameter estimates.

	Form	PA/EA	IA/CG	PG/T	Total
μ_{a1}	B	.87	.92	1.07	.94
	F	.50	1.12	.91	.81
σ_{a1}	B	.53	.37	.60	.52
	F	.54	.58	.64	.64
μ_{a2}	B	.77	1.01	.63	.80
	F	.77	.82	.94	.84
σ_{a2}	B	.60	.41	.34	.50
	F	.29	.37	.29	.33
μ_d	B	-.63	-1.23	-1.31	-1.01
	F	-.23	-1.16	-1.00	-.74
σ_d	B	1.42	1.02	1.47	1.37
	F	1.11	1.33	1.60	1.40
μ_D	B	.29	.79	.82	.60
	F	-.03	68	.52	.35
σ_D	B	.86	.57	.71	.78
	F	.88	.63	.96	.90
μ_α	B	39.95	47.94	32.93	40.24
	F	62.12	38.96	51.15	51.88
σ_α	B	15.38	10.65	15.13	15.20
	F	14.49	12.04	20.63	18.63
μ_c	B	.17	.17	.17	.17
	F	.17	.19	.18	.18
σ_c	B	.05	.05	.06	.06
	F	.05	.05	.06	.05

Figure Captions

Figure 1. Cumulative percentage frequency polygon for Form 39B and Form 39F.

Figure 2. Two dimensional MIRT item vector plots for Form 39B subtests.

Figure 3. Two dimensional MIRT item vector plots for Form 39F subtests.

Figure 4. Vector plots illustrating the difference in test information between the subtests for Forms 39F and 39B at 10° intervals at selected points on the two-dimensional ability plane.

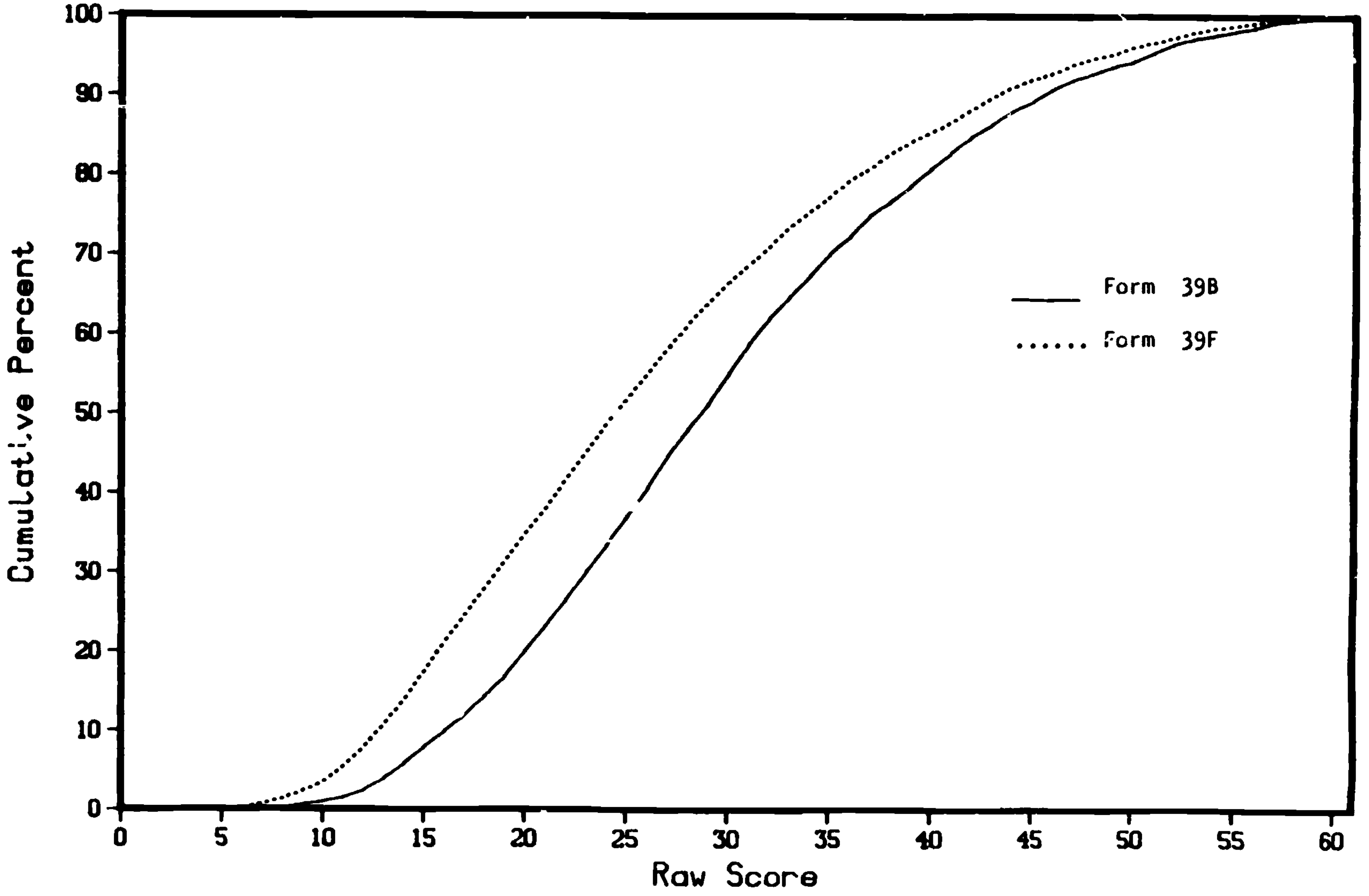
Figure 5. A surface and contour plot illustrating the difference between the test characteristic surface of Forms 39F and 39B for the PA/EA subtest.

Figure 6. A surface and contour plot illustrating the difference between the test characteristic surface of Forms 39F and 39B for the IA/CG subtest.

Figure 7. A surface and contour plot illustrating the difference between the test characteristic surface of Forms 39F and 39B for the PG/T subtest.

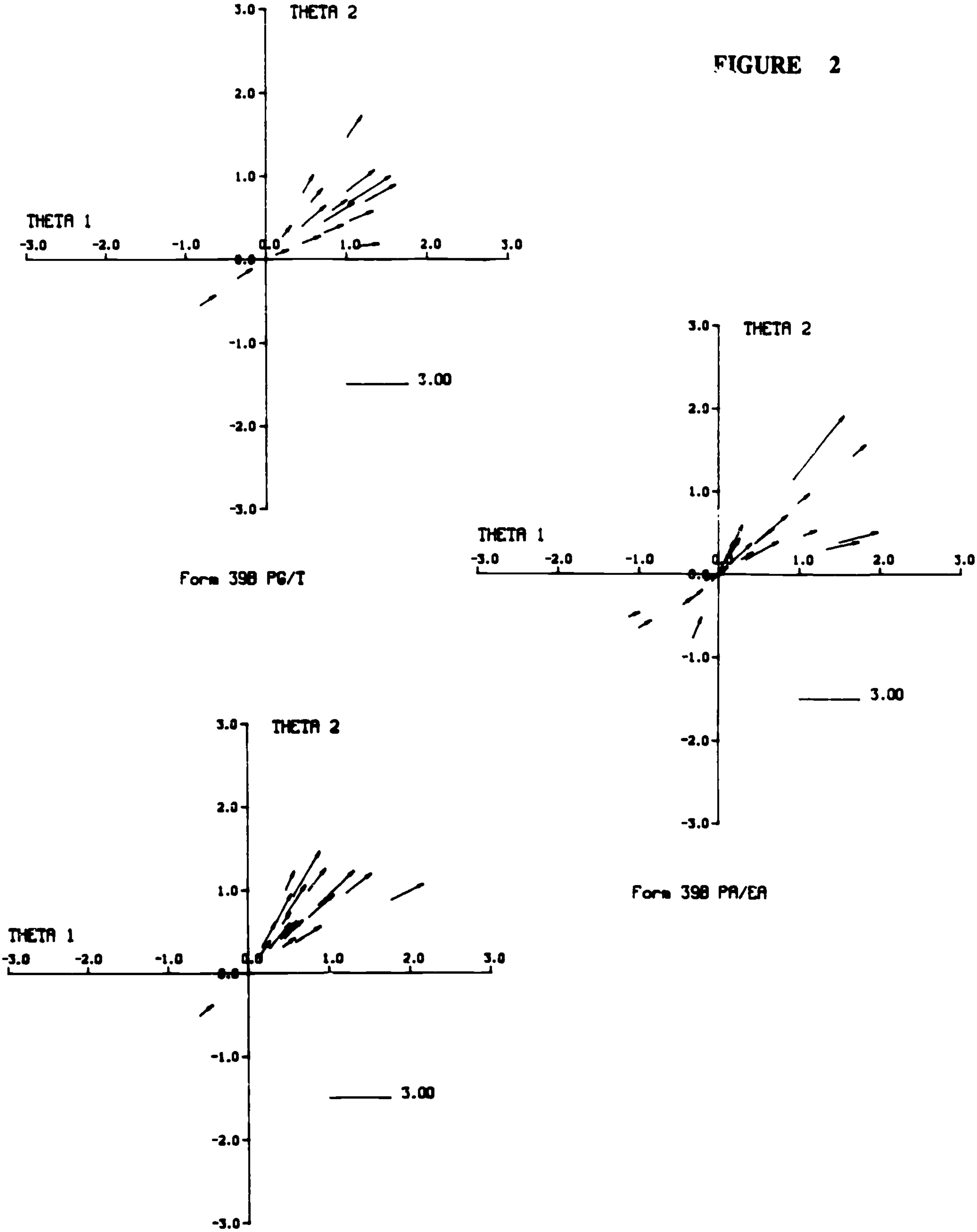
Figure 8. A surface and contour plot illustrating the difference between the test characteristic surface of Forms 39F and 39B.

FIGURE 1



100
90
80
70
60
50
40
30
20
10
0
Cumulati.ve Percent
0
5
10
15
20
25
30
35
40
45
50
55
60
Raw Score
Form 39B
Form 39F

FIGURE 2

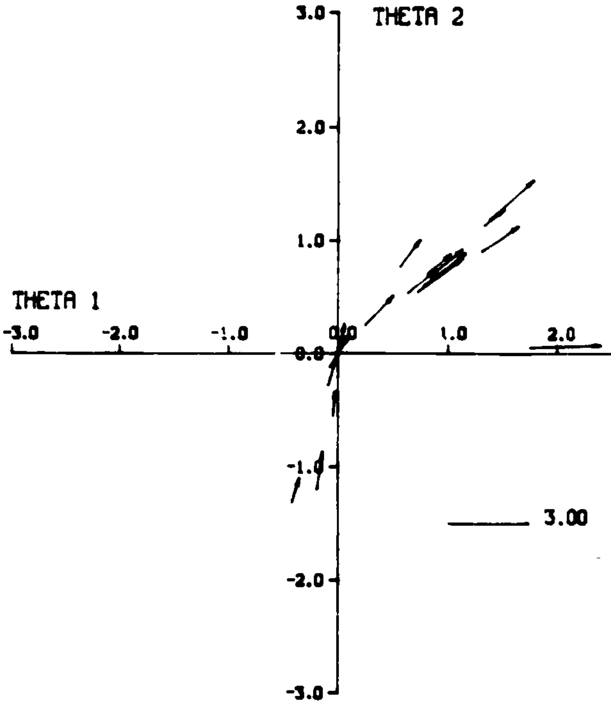


Form 398 PG/T

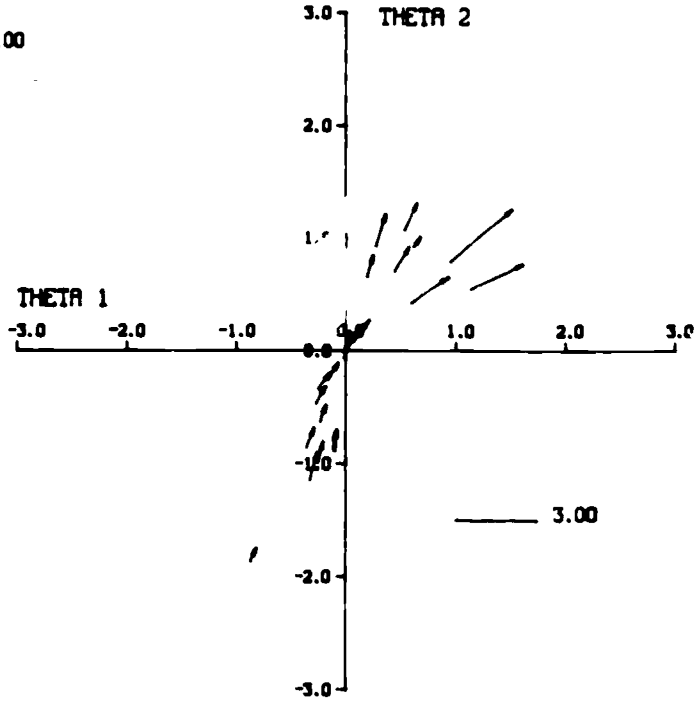
Form 398 PA/EA

Form 398 IA/CG

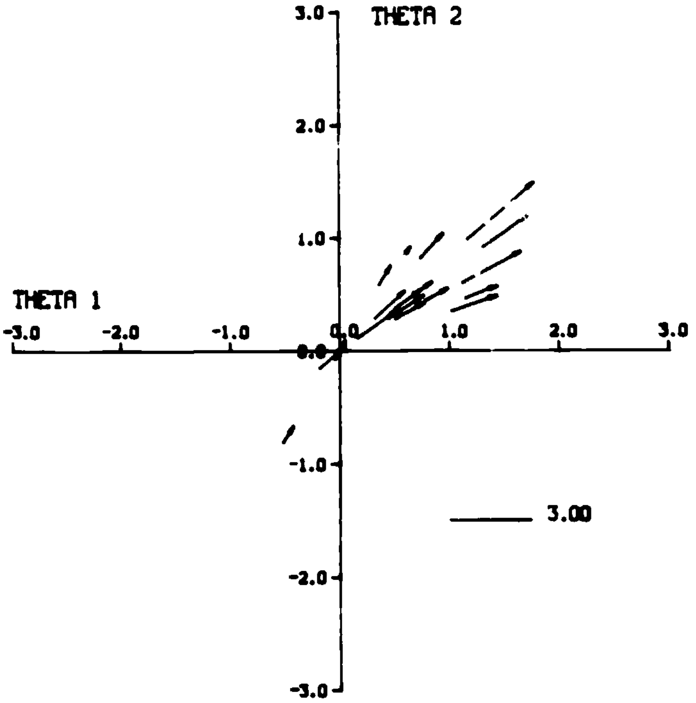
FIGURE 3



Form 39F PG/T

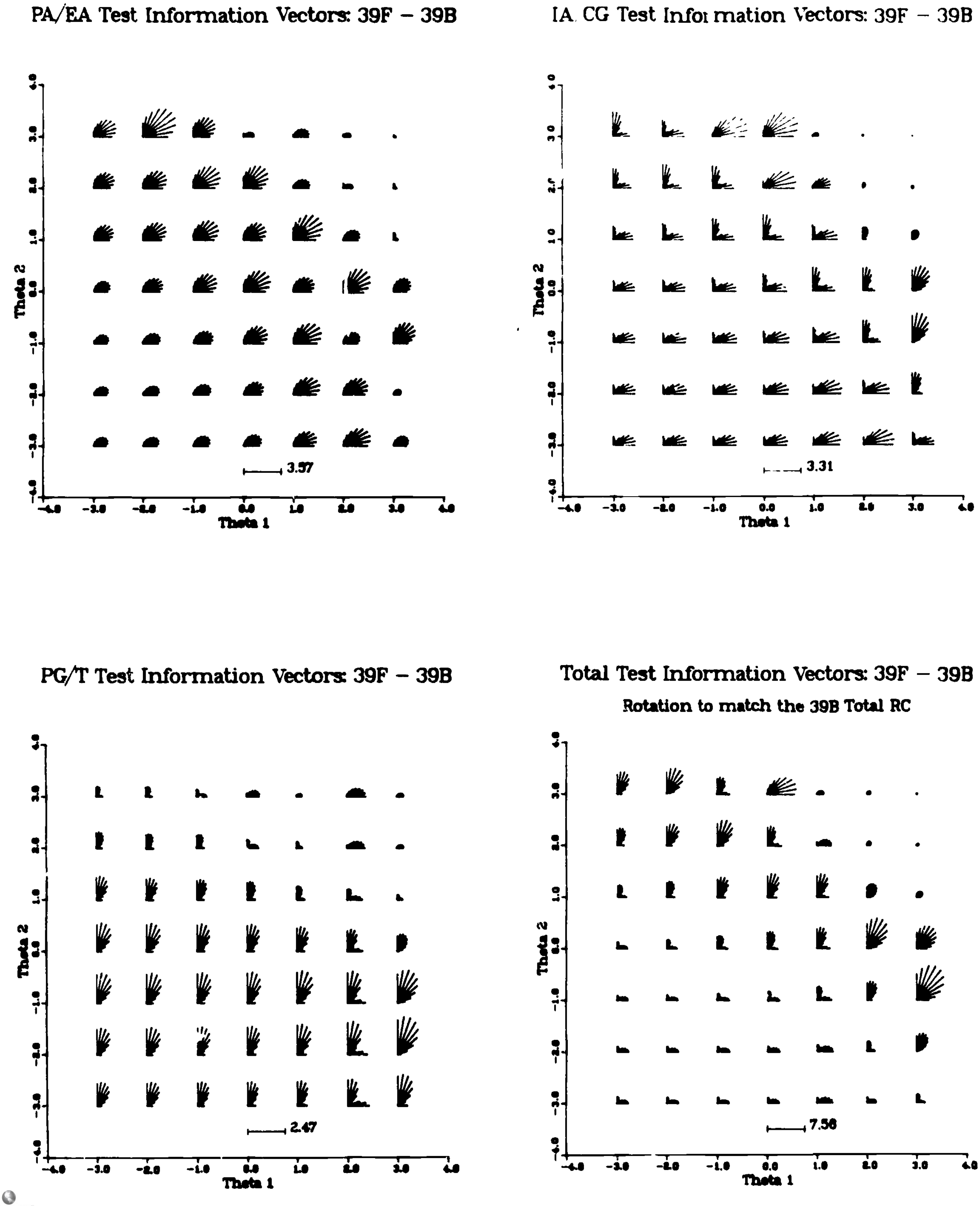


Form 39F PA/EA



Form 39F IA/CG

FIGURE 4



PA/EA Test Information Vectors: 39F – 39B

IA CG Test Infoı mation Vectors: 39F – 39B

PG/T Test Information Vectors: 39F – 39B

Total Test Information Vectors: 39F – 39B

Rotation to match the 39B Total RC

FIGURE 5

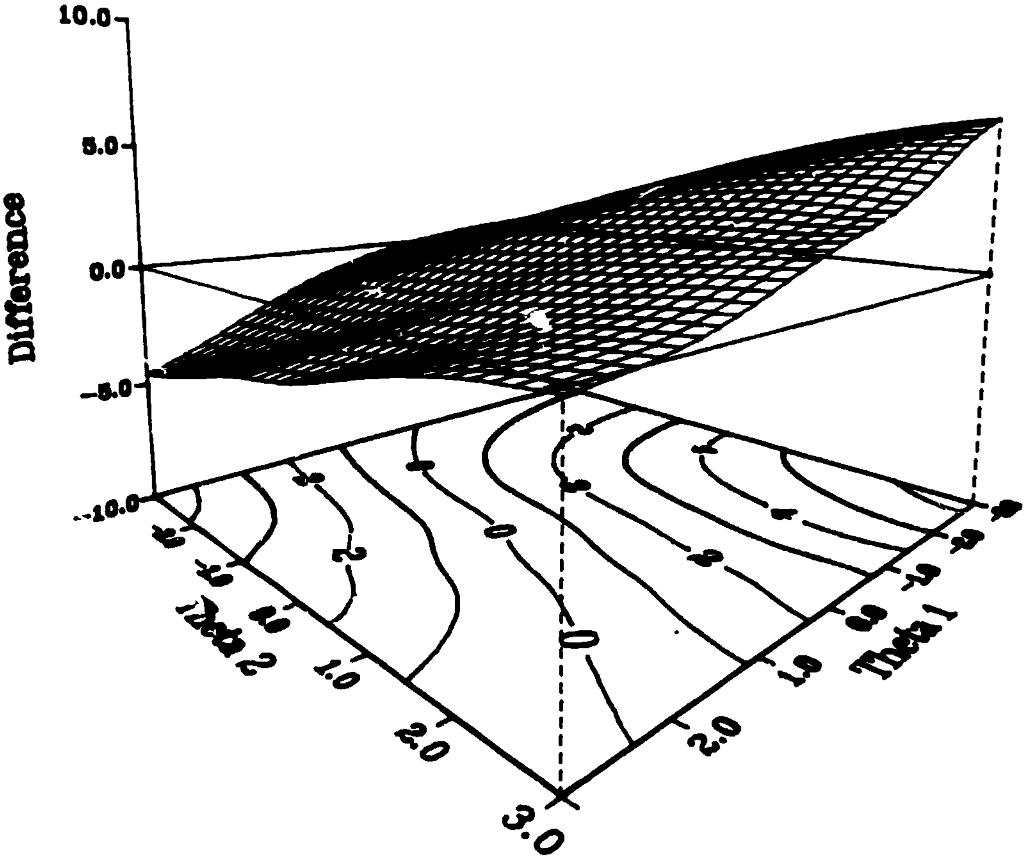

10.0
5.0
0.0
-5.0
-10.0
Difference
Theta 2
1.0
2.0
3.0
Theta 1
1.0
2.0
-1.0
-2
0
0
2
-4
4

FIGURE 6

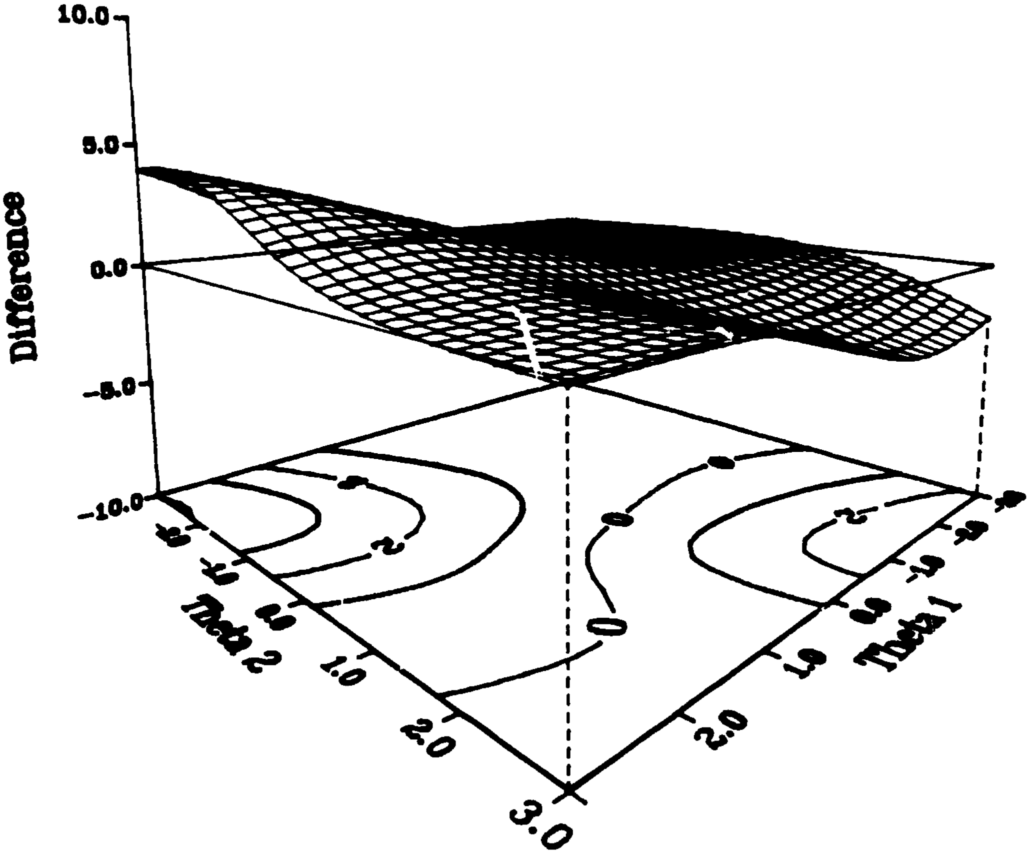

10.0
5.0
0.0
-5.0
-10.0
Difference
-1.0
0.0
1.0
2.0
3.0
Theta 2
-2.0
-1.0
0.0
1.0
2.0
Theta 1
-2
2
0
0
0
2

FIGURE 7

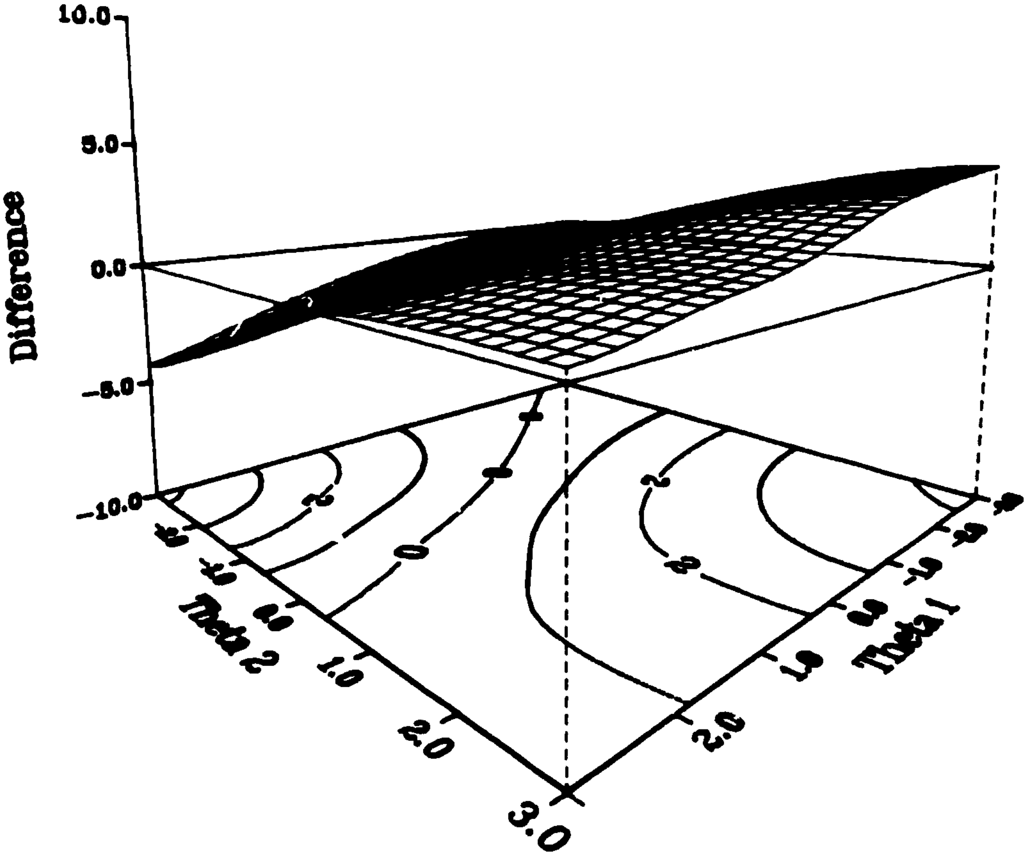
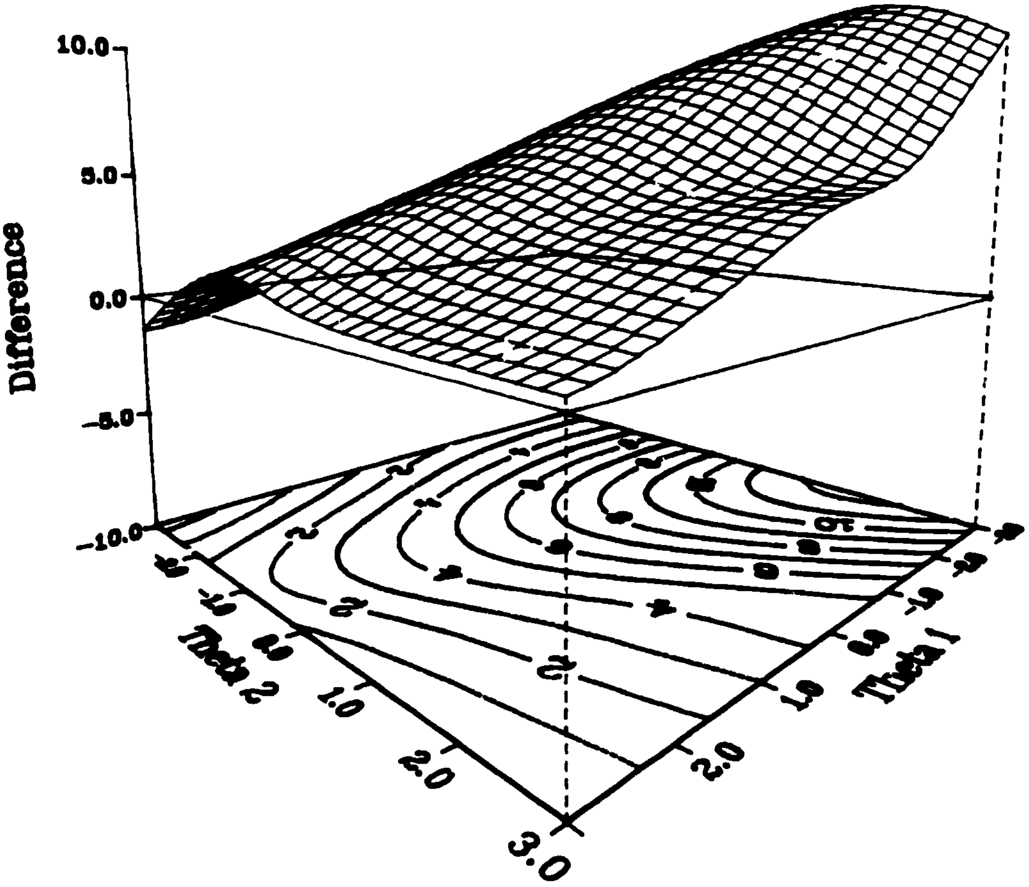


FIGURE 8



FORM 39B - Item 56

A rectangular box has a length of 10 feet, a width of 6 feet, and a height of 4 feet. The longest straight rod that could fit in this box would have to go along the diagonal between opposite corners. How many feet long is this distance?

F. $2\sqrt{13}$

G. $2\sqrt{29}$

H. $2\sqrt{34}$

J. $2\sqrt{38}$

K. $4\sqrt{15}$

FORM 39F - Item 51

In the figure below, square *ABCD* has sides of length 4 unit*s*, and *M* and *N* are midpoints of $\overline{AB}$ and $\overline{CD}$, respectively. What is the perimeter, in units, of quadrilateral *AMCN*?