DOCUMENT RESUME

ED 363 646 TM 020 669

| | |
|---|---|
| AUTHOR | Sireci, Stephen G.; Geisinger, Kurt |
| TITLE | Using Subject Matter Experts To Assess Content Representation: A MDS Analysis. |
| INSTITUTION | American Council on Education, Washington, DC. GED Testing Service. |
| PUB DATE | Apr 93 |
| NOTE | 66p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993). Part of a Doctoral Dissertation, Fordham University. |
| PUB TYPE | Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | Achievement Tests; Certified Public Accountants; Cluster Analysis; *Content Analysis; Content Validity; Correlation; Factor Analysis; Higher Education; Item Analysis; Junior High Schools; *Licensing Examinations (Professions); Mathematical Models; Matrices; *Multidimensional Scaling; *Research Methodology; Social Studies; Standardized Tests; Test Construction; *Test Items |
| IDENTIFIERS | *Experts; Similarities; Subject Content Knowledge; *Subject Specialists |

ABSTRACT

Various methods used to assess the content of a test are reviewed, and a new procedure designed to improve on these methods is presented. The two tests considered are a professional licensure examination, the auditing section of the Uniform Certified Public Accountant Examination, and an educational achievement test, a nationally standardized social studies achievement test. Previous methods have generally been empirical, using factor analysis or multidimensional scaling (MDS) to analyze the inter-item correlation matrix derived from examinee responses, or subjective, using the data provided by subject matter experts (SMEs) to determine whether items represent content areas that the test purports to measure. A method has previously been proposed that uses MDS to discover dimensions obtained from the analysis of ratings by SMEs of the similarity of items comprising a test. This study expanded that method by using 2 groups of SMEs (15 for each test) to evaluate the content of the 2 tests studied. Correlation and cluster analyses results suggest that the content structure of a test can be evaluated adequately by analyzing item similarity data provided by SMEs. Results further suggest that the MDS procedure should be used to supplement analyses of item relevance data rather than replace them. Six figures and 18 tables present analysis findings. (Contains 23 references.) (SLD)

# Using Subject Matter Experts to Assess Content Representation: A MDS Analysis[1]

Stephen G. Sireci

American Council on Education: GED Testing Service

Kurt Geisinger

SUNY Oswego

Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 15, 1993, Atlanta.

[1] The authors would like to thank Michael Green, Susan Menelaides, and the Research in Social Studies Education SIG of AERA for their help in providing the subject matter experts used in this study; Jim Blum and Paul Williams for providing the test instruments; Thanos Patelis for writing the data input program, and Lynn Shelley for her assistance in the preparation of the item similarity booklets. This research was part of the first author's dissertation research while at Fordham University. Substantial improvements to the technical quality of this report were made by his dissertation committee: Kevin Moreland, Warren Tryon, and John Walsh. This research was supported in part by a Dissertation Research Award from the American Psychological Association.

## Introduction

Developers of educational and psychological tests must demonstrate that their tests adequately measure the domain of content they purport to measure. This reasonable and fundamental requirement is commonly accepted by test specialists. Though there is agreement that a test must demonstrate adequate representation of the content domain tested, there are few procedures available to assess whether a test accomplishes this goal.

The purpose of this study was to review previous methods used to assess the content of a test and to evaluate a new procedure designed to improve upon these methods. The strengths and limitations of the new procedure were evaluated by analyzing the content representation of two tests using both the new and previous methods, and comparing the types of information provided. The two tests used in this study were a professional licensure examination (the Auditing section from the Uniform Certified Public Accountant Examination) and an educational achievement test (a nationally standardized social studies achievement test).

### Previous methods used to evaluate test content

Previous methods used to evaluate test content can be generally classified as either empirical or subjective. Empirical methods analyze item response data (i.e., examinees' responses to test items) to discover the underlying structure of these data. Subjective methods analyze the data provided by subject matter experts (SMEs) who rate the relevance of the test items to the content objectives specified in the test blueprint.[2]

### Empirical methods

Empirical methods typically use factor analysis or multidimensional scaling (MDS) to analyze the inter-item correlation matrix derived from examinees' responses to test items (e.g., Henrysson, 1971; Napior, 1972; Oltman, Stricker, & Barrows, 1990). The resulting factors or dimensions are compared with the structure of the content domain specified in the test blueprint. Though these methods are apparently objective, they are criticized as being irrelevant for content assessment because the degree of relevance of an item to its corresponding content domain is a concept that is independent of examinees' performance on the item. Variables inherent in item response data, such as item difficulty, the ability level and variability of the population tested, motivation, guessing, differential item functioning, social desirability, etc., affect factor or dimensional solutions, but are irrelevant to the assessment of content representation. While such analyses are relevant in construct validation or criterion-related studies, they are not central to evaluations of test content (Messick, 1989).

[2] The descriptive labels "empirical methods" and "subjective methods" are used to distinguish between methods employing *ratings* of test items from those analyzing item score data. These labels are not meant to imply that empiricism is not used in the judgmental methods or that the empirical methods are free from the subjective interpretations of the investigator.

Subjective methods

Subjective methods for evaluating test content use SMEs to determine whether the items that comprise a test represent the content areas the test purports to measure. The SMEs review the test items and make judgments regarding the appropriateness of each item for measurement of the content domain it purports to measure as defined by the test blueprint. There are several variations of the subjective method (see Crocker, Miller, & Franks, 1989; or Osterlind, 1989, for a review of some of these methods). The variations differ in the ratings required by the SMEs, and in whether or not they are informed of the content area specifications (blueprint specifications) of the test items.

Examples of subjective methods for evaluating test content are provided by Hambleton (1980, 1984), and Aiken (1980). Hambleton's method involved having SMEs rate, along a 3-point scale, the extent to which an item measured each of the content areas of the test. This task provided a "item-objective congruence index" that reflected the SMEs' perceptions of how well each item was matched to its blueprint classification. Hambleton (1984) also suggested that if SME ratings were measured along longer Likert-type scales, the mean congruence rating for each item, averaged over the SMEs, would provide a straightforward descriptive index of the SMEs' perceptions of the fit of an item to its designated content area. Aiken (1980) also provided an index derived from SMEs' ratings for evaluating the relevance of an item to a particular content domain. This index accounts for the number of categories used to rate each item and the number of judges that responded to each category. The equation for Aiken's validity index *v*, is

$$V = \frac{\sum_{i=1}^{c-1} i n_i}{N(c-1)} \qquad (1)$$

where *c* is the number of categories on the item relevance rating scale, *i* is the weight given to each category, $n_i$ is the number of judges who rated the item into the *i*th category, and *N* is the total number of SMEs. The lowest category is given a weight, (or *i*-value) of zero, the next category is given a weight of 1, etc, and the highest category is given a weight of *c-1*. Aiken provided a formula for evaluating the significance of the validity index when a large number of SMEs are used. This formula provides a normal deviate (z-score) for the index and the probability of obtaining the z-score is obtained from a standard normal z-table. The formula for deriving a normal deviate from the Aiken index is

$$Z = \frac{N(c-1)(2V-1) - 1}{\sqrt{\frac{N(c-1)(c+1)}{3}}} \qquad (2).$$

Like other subjective methods used to evaluate test content (c.f. Lawshe, 1975; Morris and Fitz-Gibbon, 1978), the methods of Hambleton and Aiken provide SME-based indices of the overall quality of each test item that can be averaged to provide an index of the overall content quality of the test.

Limitations of subjective methods. Though the subjective methods provide indices that can be used to evaluate the content representation of a test, they have two major limitations. First, because the SMEs are informed of the content areas that comprise the test blueprint, they cannot provide an independent assessment of the content structure of the test. Rather, the structure created by the test developers (the test blueprint) is imposed upon their ratings. Thus, these methods may tend to support implicitly the content structure of the test as defined by the test developers. The second major limitation is that these methods do not evaluate the overall content structure of the test. Because each item is rated in isolation (i.e., independent of the other items), no information regarding the global content structure of the test is provided.

### The MDS Method

Sireci and Geisinger (1992) used multidimensional scaling (MDS) to discover dimensions obtained from the analysis of SMEs' ratings of the similarity of all items comprising a test. Rather than being asked to judge the relevance of the items to the content areas specified in the blueprint, the SMEs were asked to judge the *similarity* of all possible pairings of test items (c.f. Tucker, 1961). This procedure prevented the SMEs from being influenced by an a priori knowledge of the content structure of the test as defined by the test developers. The results of this study provided several dimensions that were deemed relevant to the content structure of the tests and were congruent with the blueprint specifications of the test.

Although the results of the Sireci and Geisinger study illustrated that MDS analyses of SMEs' ratings of item similarity could provide information pertaining to the content structure of a test, the study was limited in several ways. First, the generality of the procedure was not evaluated. Only three SMEs were used and the test investigated was a locally-developed classroom test. Second, the MDS dimensions were interpreted subjectively without any supplementary analyses to support their interpretations. Third, traditional means for analyzing SMEs perceptions of content relevance were not used. This precluded the MDS method from being compared with the traditional methods on the same test instrument. To redress these shortcomings, the authors suggested that future research employ more SMEs, focus on different types of tests, and supplement the MDS analyses with more traditional analyses of item relevance data.

## Introduction to the Present Study

The purpose of the present study was to expand the method for evaluating test content proposed by Sireci and Geisinger (1992) and compare it with the more traditional methods. Two groups of SMEs were employed to evaluate the content of two different tests. The item similarity ratings provided by the SMEs were analyzed using MDS to uncover the structure of these data.

## Method

### Instruments

The two tests evaluated in this study were the Auditing section from the May, 1990 Uniform CPA Examination and a form of a nationally-standardized social studies test designed to measure achievement in social studies for junior high school students. The Auditing section was comprised of 60 multiple-choice items and four essays. Due to the difficulty in comparing essay-format questions with multiple-choice format questions, the essay questions were excluded from review. Because it would be an insurmountable task for the SMEs to rate the similarity of all 60 multiple-choice items to one another (1,770 paired comparisons would be necessary), and because it would be difficult to inspect multidimensional configurations containing 60 points, a 40-item subset of the 60-item test was used. The 40-item subset was chosen from the original 60-item test by: first, eliminating nine re-use items (items that were administered previously and were selected for repeated use because of their desirable content and statistical characteristics); second, eliminating five other items that involved different item formats than the other items (one item had an accompanying figure, and the other four items were so-called "K-type" items[3]); and third removing six other items randomly so that the percentage of items representing each content area in the original 60-item test was maintained in the 40-item subset.

The Social Studies Test was comprised of 40 multiple-choice items; all 40 items were included in the analysis. The blueprint

[3] "K-type" items include three columns of information in the item stem and require examinees to choose which columns correspond to the correct answer (e.g., "I only," "II only," "I and II only," "I, II, and III").

of the Auditing Examination was comprised of four content areas: professional responsibilities, internal control, evidence and procedures, and reporting. The blueprint of the Social Studies Test was comprised of seven content areas: geography, economics, history, political science,interrelated disciplines, sociology/anthropology, and applied social studies. The content area blueprints for these two tests are presented in Tables 1 and 2, respectively.

Insert Table 1 About Here

Insert Table 2 About Here

Subjects

Thirty SMEs provided the data for this study. Fifteen SMEs were used to evaluate the Auditing Examination and another fifteen were used to evaluate the social studies test. The auditing SMEs were required to be licensed Certified Public Accountants (CPAs) with at least three years experience performing audits. Six of the auditing SMEs were female (40%) and nine were male (60%). The social studies SMEs were required to be State-certified to teach social studies and have at least three years experience in teaching social studies at the junior high school level. Ten social studies SMEs were female (67%) and five were male (33%).

Procedure

Item similarity ratings. The Auditing and Social Studies SMEs completed the same tasks. Each SME was given a booklet containing all possible pairings of the 40 test items. Thus, each booklet had 780 item pairings. Beneath each item pair was a ten-point Likert scale with the anchor for the point of 1 labeled "highly similar" and anchor for the point of 10 labeled "highly dissimilar." The SMEs were instructed to read each item pair and make a judgment regarding the similarity of the two items in terms of the (auditing or social studies) knowledge being measured. They were instructed to circle their rating on the ten-point scale printed below the item pair. The SMEs were not provided with any further criteria on which to make their ratings. This ambiguity in instruction was used to avoid biasing their ratings in favor of the test blueprint. After the SMEs had completed the 780 item similarity ratings, they were asked to complete a brief questionnaire that asked them to list the criteria they used to judge the item similarities.

Relevance ratings. Following completion of the item similarity ratings and follow-up questionnaire, the SMEs rated the relevance of each of the test items to each of the content areas of the test blueprint. These ratings were also made along a ten-point scale where 1="not at all relevant," and 10="highly relevant." Each SME was given a content area description sheet that described each of the content areas of the test blueprint. This step was completed so that data typical of previous methods of content evaluation could be compared to the present method.

The procedures described above resulted in the following data collected from each SME: 1) a lower triangular matrix of item similarity ratings, 2) a list of criteria used in making the item similarity ratings, and 3) a rectangular matrix of item relevance ratings.

Data Analyses

MDS Analyses. The item similarity data for each group of SMEs were analyzed separately. The data for each group of SMEs were analyzed using the INDSCAL (Carroll and Chang, 1970) model of the ALSCAL program of SPSSX (Young, Takane, & Lewyckyj, 1978). The INDSCAL MDS model represents a generalization of the classical multidimensional scaling model (CMDS) developed by Torgerson (1958), and expanded by Shepard (1962), and Kruskal (1964). In the INDSCAL model, each subject's dissimilarity matrix is multiplied by a vector of weights (**w**) consisting of elements $\mathbf{w}_{ka}$ that represent the relative emphasis subject *k* places on dimension *a*. The distances between stimuli in the INDSCAL model are computed by incorporating this weighting factor into the Euclidean distance formula used by classical MDS: Thus, the INDSCAL model defines the distance between two objects *i* and *j* as:

$$d_{ijk}=\sqrt{\sum_{a=1}^{r} w_{ka}(x_{ia}-x_{ja})^2} \qquad (3)$$

where: $d_{ij}$=the Euclidean distance between points *i* and *j*, $X_{ia}$=the coordinate of point *i* on dimension *a* and *r*=the maximum dimensionality requested. The results from an INDSCAL analysis include a multidimensional configuration of the attributes rated, called the stimulus space, and a multidimensional configuration of the subjects known as the group space (subject space). The elements of $w_{ka}$ serve as the coordinates for each subject in the subject space. Scrutiny of the stimulus space allows for visual inspection of the perceived similarity among the items. Items that are proximal to each other exhibit greater similarity than items that are more distant. Similarly, inspection of the subject space allows for visual inspection of the similarity among the subjects.

Two- through six-dimensional INDSCAL solutions were obtained for each group. These analyses were performed to discover the structure of the item similarity data and to investigate differences among the SMEs.

Correlation and regression analyses. The item relevance data were averaged over the 15 SMEs in each group and analyzed together with the item coordinates resulting from the INDSCAL analyses. First, the correlations among the relevance data for each content area were correlated with the coordinates for each of the MDS dimensions. Subsequently, the item relevance data were regressed onto the INDSCAL coordinates. These analyses were performed to help interpret the INDSCAL dimensions and to discover whether the dimensions obtained were relevant to the content structure of the test.

Cluster analyses. To aid in the interpretation of the

INDSCAL dimensions, and to discover substantive groupings of items in the MDS space, the item coordinates resulting from the INDSCAL analyses were cluster-analyzed using hierarchical cluster analysis. The between-groups average linkage method (Johnson, 1967; Sokal and Michener, 1958) was used to form the clusters.

Traditional analyses. To compare the results of the MDS procedure with previous methods, the item relevance data were analyzed independently. The relevance data were averaged over the 15 SMEs and the mean relevance rating for each item on each objective was computed. In addition to this averaged data, the percentage of SMEs who rated each item most relevant to its blueprint content area was also calculated. Finally, Aiken's (1980) validity index was computed for each item.

## Results

### The Auditing Examination

#### MDS Results

For the analysis of the Auditing Examination, in addition to the 15 SMEs, the test developer also completed the similarity ratings. The data from the Auditing test developer were included in the INDSCAL analyses to provide a reference point in the subject space (i.e., to discover how close the ratings of the SMEs corresponded to those of an expert who was aware of the items' blueprint specifications). Before looking at the fit of the data to the entire group of SMEs, a general inspection of the congruence of the SMEs was conducted.

SME congruence. The congruence of the auditing SMEs was evaluated by the fit values (RSQ and STRESS) observed for each SME, inspection of the weirdness indices for each SME, inspection of the subject space, and re-evaluation of the data after removing the most "aberrant" SMEs. The RSQ (amount of variance in the data accounted for by the model) and STRESS (departure of the data from the model) values for each SME are presented in Table 3 for the six-dimensional INDSCAL solution. The values obtained by the test developer (SME #1) indicated slightly better fit than the average obtained for the entire 16-matrix population (STRESS=.12 and RSQ=.78 for the test developer, average STRESS=.14 and RSQ=.63 for all 16 matrices). SME numbers 6, 12, 13, and 14 exhibited relatively poor fit (RSQ below .50), while SME numbers 8 and 11 exhibited relatively good fit (RSQ above .80).

Insert Table 3 About Here

The STRESS and RSQ values for each SME indicate that there were some differences among the SMEs in terms of their fit to the INDSCAL solution. Inspection of the subject weights and weirdness indices for each SME confirmed this observation. The subject weights represent the relative emphasis each subject places on each dimension. The weirdness index reflects the degree to which a subject uses a particular dimension in proportion to the other dimensions. The subject weights and weirdness indices for each SME are presented in Table 4. None of the subjects exhibited exceptionally high weirdness indices, although the test developer (SME #1) and SME #8 had indices above .25. The test developer had a higher relative weight on Dimension 3 than the other SMEs and SME #8 had a higher relative weight on Dimension 4.

Insert Table 4 About Here

To inspect the differences among the SMEs visually, the subject weights for Dimensions 1, 3, and 4 are presented in Figure 1. These three dimensions are portrayed together because they illustrate the three dimensions on which the SMEs differed the most. Figure 1 illustrates the emphasis that the test developer (#1) placed on Dimension 3, and the emphasis that SME #8 placed on Dimension 4.

Insert Figure 1 About Here

Subset analyses. Because of the differences observed among the SMEs, separate INDSCAL analyses were conducted with SMEs #1, 6, 12, 13, and 14 removed individually and together. The fit and interpretability of these solutions were compared with the solution derived from analysis of the entire data set. These subset solutions did not exhibit substantial improvement in fit, and did not provide different interpretations of the stimulus space than those that were gleaned from analysis of the complete data set. For these reasons, all SMEs were retained in the subsequent analyses, and the complete 16-matrix data set was used for all auditing MDS analyses reported below.

Final MDS Solution

A six-dimensional INDSCAL solution was chosen as the appropriate MDS model for the Auditing SME similarity data. The six-dimensional model was chosen on the basis of fit to the data and interpretability. The values of RSQ and STRESS, averaged over the 16 SMEs, are presented in Table 5. These values indicate that a moderate amount of variation was present in these data that was not accounted for by the model. However, these values are not surprising given the relatively large number of matrices and stimuli.

Insert Table 5 About Here

Visual interpretation. All six INDSCAL dimensions were interpretable for the auditing data. All of the dimensions separated the items according to characteristics relevant to the auditing content domain. Two of the dimensions were directly related to the content areas designated in the test blueprint. The other four dimensions separated the items according to other aspects of auditing that were not specified in the content area groupings of the test blueprint.

Dimensions 1 and 4 were the two dimensions that most clearly separated items comprising the different content areas.

Dimension 1 distinguished between Reporting items and Internal Control items, and Dimension 4 distinguished between Internal Control and Evidence & Procedures items. The two-dimensional scatterplot of Dimension 1 versus Dimension 4 is presented in Figure 2. No dimension clearly distinguished the Professional Responsibilities items from the other content areas, but Dimension 2 did account for knowledge and application of professional standards and did pull most of these items in the same direction. Because these three dimensions were most relevant to the content structure of the test, they are plotted together in Figure 3.

Insert Figure 2 About Here

Insert Figure 3 About Here

Dimensions 3, 5, and 6 reflected other content characteristics of the items that were extraneous to the content areas listed in the test blueprint. Dimension 3 distinguished items measuring commonly-performed auditing procedures from those measuring extraordinary auditing procedures, Dimension 5 distinguished items related to the execution of the audit from those involved with planning or concluding the audit, and Dimension 6 distinguished items that measured higher-level auditing procedures from the more elementary auditing procedures.

The visual interpretation of the dimensions indicated that the dimensions were relevant to the content characteristics of the items. To validate these interpretations the content area relevance ratings were analyzed together with the six-dimensional INDSCAL stimulus (item) coordinates. The results of these analyses are reported below.

Correlation and regression results

The content area relevance ratings gathered from the SMEs represented the SMEs' ratings of the relevance of each item to each of the four content areas of the auditing test blueprint. The test developer, SME #1, was not asked to rate the items because he was aware of their content area designations. The correlations of the relevance ratings with the coordinates resulting from the six-dimensional INDSCAL solution are presented in Table 6.

Insert Table 6 About Here

The correlations between the averaged item relevance data and the six-dimensional coordinates from the INDSCAL solution supported the visual interpretations given to each dimension. Because Dimensions 1 and 4 were most related to the content areas of the test blueprint, they exhibited significant correlations with the content areas to which they corresponded. The coordinates for Dimension 1 (reporting versus internal control) correlated highly with the relevance ratings for the Reporting content area and the Internal Control content area. The coordinates for Dimension 4 (internal control versus evidence and procedures) correlated highly with the relevance ratings for the Internal Control content area and the Evidence and Procedures content area. The coordinates for Dimensions 2, 3, 5, and 6 each correlated significantly (at $\underline{p} < .05$) with the relevance ratings from at least one content area. However, because these dimensions did not relate directly to the blueprint content areas, the correlations were difficult to interpret. Though these correlations did not support the interpretations of these dimensions directly, they were not incongruent with the interpretations given to each dimension.

Regression analyses. The results of the multiple regression analyses, where the relevance ratings for each content area were regressed onto the item coordinates for all six dimensions, were consistent with the correlational results. A summary of these results is presented in Table 7. Though these results supported the dimensional interpretations generally, there were many more statistically significant beta weights obtained than were expected, given the interpretations ascribed to each dimension.

Insert Table 7 About Here

### Cluster analysis results

The cluster analysis of the MDS item coordinates revealed several substantive groupings of test items that related directly to the item groupings specified in the test blueprint. The results from this hierarchical cluster analysis are presented in Table 8. The items comprising the content areas of Reporting, Internal Control, and Evidence and Procedures were grouped together in the clustering solution. Some of the items comprising the Professional Responsibilities content area clustered together; however, these items tended to overlap with the other three content areas. Four of the clusters presented in Table 8 reflect the four content areas designated in the test blueprint: stages 30 (Evidence and Procedures), 34 (Reporting), 35 (Internal Control), and 36 (Professional Responsibilities). Given these four substantive clusters, several items were not joined with the other items comprising their blueprint content area.

Insert Table 8 About Here

## Results using traditional procedures

Analysis of the relevance data, averaged over the 15 SMEs, indicated that only seven of the forty items were not "matched" to their blueprint content area by the SMEs (i.e., these seven items had higher relevance ratings for one of the other three content areas). However, these seven items were still judged to be relevant to their blueprint content area. A summary of the averaged auditing content area relevance ratings is presented in Table 9.

Insert Table 9 About Here

Analysis of the percentage of SMEs who classified each item in accordance with its blueprint specification revealed that 18 items were classified correctly by all 15 of the SMEs, 10 items were classified correctly by 12 to 14 of the SMEs (80 to 93%), 8 items were classified correctly by only 8 to 11 SMEs (53 to 73%), and 4 items were classified correctly by less than half of the SME (5 or 6 SMEs). Of the four items that were matched by less than half the SMEs, two were Professional Responsibilities items, one was an Internal Control item, and one was an Evidence and Procedures item.

Analysis of the relevance data using Aiken's (1980) validity index identified fewer items as incongruous with their blueprint classification than did the preceding analysis. However, seven items had relatively low validity indices (below .70), and three of these items were included in the four items misclassified by at least half of the SMEs (reported above). The Aiken validity index for the auditing items are presented in Table 10.

Insert Table 10 About Here

## Social Studies Results

### MDS Results

SME congruence. The fit values of RSQ and STRESS for the 15 social studies SMEs are presented in Table 11. The fit values for these SMEs were generally lower than those observed for the auditing SMEs. SME #9 obtained the lowest value of RSQ (.187) and largest value of STRESS (.205), indicating that s/he was not rating the items in a manner similar to the other SMEs.

Insert Table 11 About Here

As with the auditing SMEs, the subject weights and weirdness indices were also evaluated to determine SME congruence. SME #5 had a relatively high weirdness index of .75 and emphasized Dimension 3, while SME #13 had a weirdness index of .61 and emphasized Dimension 1. A three-dimensional scatterplot, portraying Dimensions 1 through 3 of the social studies SME subject space, is presented in Figure 4. The scatterplot in Figure 4 clearly illustrates the emphasis SME #5 placed on Dimension 3, and the essential unidimensionality of SME #13's data.

Insert Figure 4 About Here

Subset analyses. Because SMEs #5, #9, and #13 appeared different from the other subjects, separate INDSCAL analyses were conducted to discover whether removing their data led to a more meaningful MDS solution. As with the subset analyses performed on the auditing data, none of the subset analyses led to substantial improvement in fit or interpretability compared to analysis of the complete data set. Therefore, all subsequent analyses were conducted using the complete, 15-matrix data set.

Final MDS Solution

A six-dimensional INDSCAL solution was chosen for the social studies item similarity data. The fit values for the two- through six-dimensional solutions, averaged over the group of SMEs, are presented in Table 12. The six-dimensional solution exhibited the best fit to the data and all six dimensions were interpretable.

Insert Table 12 About Here

Unlike the auditing data, the six-dimensional INDSCAL solution did not reflect the general content structure specified in the test blueprint. Five of the dimensions distinguished the test items according to their content characteristics. However, only three of these dimensions separated blueprint content areas: Dimension 1 separated the geography items from the other items, Dimension 2 separated the economics items from the others, and Dimension 4 separated the history items from the other test items. A two-dimensional scatterplot of Dimension 1 plotted against Dimension 2 is presented in Figure 5.

Insert Figure 5 About Here

Dimensions 5 and 6 configured the items according to content characteristics that were not accounted for in the test blueprint. Dimension 5 separated those items that dealt with culture, from the items that were less related to culture. Dimension 6 separated those items dealing with U.S.-specific information from items that had an international context. Dimension 3 was the only dimension that was not related to the content characteristics of the items. Rather, this dimension accounted for the cognitive levels measured by the items. Items that measured higher-level thinking skills were separated from those items measuring higher level skills.

The three dimensions that most closely reflected the content areas specified in the test blueprint (Dimensions 1, 2, and 4) are plotted together in Figure 6. The different symbols plotted in Figure 6 illustrate the content areas to which the items correspond. An inspection of Figure 6 illustrates that, using the three dimensions most representative of the structure of the test blueprint, substantial overlap among the content areas exists.

Insert Figure 6 About Here

## Correlation and regression results

As with the auditing data, the content area relevance data were averaged over the 15 SMEs and correlated with the six dimensional coordinates for each item. (Relevance ratings for the Interrelated Disciplines content area were not gathered because this was not a distinct content area; rather, items that were classified as Interrelated integrated two of the other six content areas.) The results of the correlation and regression analyses were generally consistent with the interpretations given to the dimensions. A summary of the correlational results is presented in Table 13. The statistically significant (at *p* <.01) correlations that were consistent with the interpretations given to the dimensions were: Dimension 1 (geography versus other) correlated with geography, dimension 2 (other versus economics) correlated with economics, Dimension 4 (other versus history) correlated with history, and Dimension 5 (cultural versus other) correlated with sociology/anthropology. Dimensions 3 (lower-level thinking skills versus higher-level thinking skills) and 6 (international versus national) did not correlate strongly with the relevance ratings for any content area.

Insert Table 13 About Here

Similar to the correlational analyses, the multiple regression analyses revealed significant regression weights for the dimensions that were deemed relevant to each content area.

Though the correlation and regression analyses supported the interpretations given to the dimensions, there were several statistically significant (at *p* <.01 or less) correlations and regression weights that were not consistent with the interpretations. A summary of the results of the regression analyses is presented in Table 14.

Insert Table 14 About Here

Cluster analysis results

The cluster analysis of the MDS item coordinates grouped together only the items comprising the economics and political science content areas. Though many of the items comprising the other content areas did group together, considerable overlap was observed. Three different cluster solutions failed to identify substantive groupings of items that represented the content areas of sociology/anthropology, applied social studies, and interrelated disciplines. The items comprising the geography and history content areas tended to group together, but the sociology/anthropology, applied social studies, and interrelated disciplines items obscured their visibility. The result of the cluster analysis of the social studies item coordinates are presented in Table 15.

Insert Table 15 About Here

Results using traditional procedures

Analysis of the relevance data averaged over the 15 SMEs indicated that the items comprising the history, political science, geography, and economics content areas were rated highly congruent to their blueprint content areas. The items comprising the content areas of sociology/anthropology, and applied social studies tended to be rated more relevant to a different content area. Analysis of the percentage of SMEs who classified each item in accordance with its blueprint specification revealed similar results.

Table 16 summarizes the mean relevance ratings by listing, for each content area, the number of items belonging to the content area that were rated highly-relevant. If the highest relevance rating for an item corresponded to the blueprint area of the item, it was placed in the "Rated First" column; if the second-highest rating corresponded to the blueprint area of the item, it was placed in the "Rated Second" column, etc. If an item was rated highly relevant to two content areas, it was considered classified as "interrelated disciplines." The summary provided in Table 16 indicates that the averaged relevance ratings were highly congruent with the blueprint specifications for the items comprising the geography, economics, history, political science, and interrelated disciplines content areas. However, the averaged ratings for the items comprising the sociology/anthropology and applied social studies content areas were less congruous with the blueprint specifications.

Insert Table 16 About Here

A crosstabulation of the areas rated most relevant for each item appears in Table 17. The cells along the diagonal of Table 17 indicate the number of items that were rated in accordance with their blueprint specifications (i.e., were rated most relevant to their specified content area). The cells off the diagonal illustrate the "misclassified" items. An inspection of Table 17 indicates that, similar to the results of the cluster analysis, the sociology/anthropology items were classified as either geography or history, and that the applied social studies items were classified as geography, history, or political science.

Insert Table 17 About Here

The summarization of the item relevance data using the averaged ratings indicated that 35% (14/40) of the items may not be congruent to their specified objectives. The majority of these 14 items (11) comprise the sociology/anthropology or applied content areas. Thus, the averaged relevance rating data indicate that these two subdomains are not represented adequately by these items. Analysis of the proportion of SMEs who classified each item correctly provided similar conclusions.

The Aiken validity indices computed for the social studies tests, and were tainted by knowledge of how the structure of the test was defined by the test developers.

Using both item similarity data and item-to-content-area relevance data to evaluate the content structures of the two tests provided greater information than the information provided by either procedure alone. The MDS and cluster analyses of the item similarity data provided an unbiased illustration of how the SMEs perceived the content structure of the test. This unbiased illustration was not provided by the traditional analysis of the item relevance data. However, the correlation and regression analyses of the item relevance data and MDS stimulus coordinates were useful in interpreting the MDS configurations. Furthermore, the traditional analyses of the item relevance data did identify some aberrant items that were not identified by the MDS/cluster analyses. The results of this study indicate that the MDS procedure should be used to supplement analyses of item relevance data rather than replace them.

Comparing the utility across test types

In comparing the utility of the procedure with respect to the two different tests, some differences were noted. The results for the Auditing Examination appeared more congruent with their blueprint specifications than were the results for the social studies test. Several reasons may account for this observation. First, there were only four content areas defined in the auditing test blueprint and so there were fewer content areas to be represented than in the analysis of the social studies items provided results similar to the percentage of SMEs classifying each item correctly. The items comprising the geography, economics, history, and political science content areas exhibited higher values than the items comprising the sociology/anthropology content area. The Aiken validity indices for the items are presented in Table 18.

Insert Table 18 About Here

## Discussion

The results of this study demonstrated that the content structure of a test can be evaluated adequately by analyzing item similarity data provided by SMEs. The configurations of test items resulting from the MDS analyses provided an unbiased illustration of how the SMEs perceived the content structure of the test. When these configurations were compared with the blueprint specifications for the tests, the homogeneity of the content areas, and overlap between content areas, were illustrated.

### Comparing the procedure to previous methods

The results of the traditional analyses of the item-to-content area relevance data were useful for identifying those content areas that were poorly represented by their constituent items, and those items that did not conform to their blueprint specifications. However, these results did not provide any information regarding the underlying content structure of the data. Because fewer content areas were involved, the potential for overlap between content areas was reduced. Thus, the seven-area social studies blueprint was a more difficult blueprint to verify than was the four-area auditing blueprint.

A second reason for the differences noted between the two application areas may stem from the fact that contextual dependencies existed among many of the social studies items, while no such dependencies were present among the auditing items. Many of the social studies items corresponded to a single visual graphic (i.e., map, figure, or time line). Some of the items that corresponded to the same visual graphic belonged to different content areas. Though the interpretations of the MDS dimensions did not indicate that any dimension was related to the contextual dependencies among the items (i.e., no dimension grouped items according to a particular visual graphic, nor to a group of graphics), some subtle effects of these contextual dependencies may have been present that were not detected by the MDS and cluster analyses.

Because of the complex nature of the domain of social studies, it may be argued that substantial overlap *should* exist among the subdomains defined in the test blueprint. Should test developers agree with this theoretical position, then the test blueprint should be modified to allow the items to correspond to more than one content area. Though such an approach may present problems if objective scores are to be derived for each content area, it may provide a more accurate representation of the content domain. Perhaps the test developers intended to represent the generality of the social studies domain by including the general content areas of applied social studies and interrelated disciplines. The heterogeneity of these two content areas was evident; but because the items comprising these two content areas were not explicitly linked to other content areas, it was difficult to evaluate how well these two subdomains were represented.

The particular social studies test used in this study may also explain the difference in results noted between the two tests. The social studies test was designed to provide norm-referenced information regarding student achievement. Conversely, the Uniform CPA Examination is referenced to comprehensive practice analyses of Certified Public Accountants. Had a more criterion-referenced social studies test been used, the results for the two tests may have been more similar.

Limitations of the procedure

Though the procedure proposed here improves upon previous methods, it has two major limitations. First, the procedure places substantial burden on the SMEs, especially when a large number of test items is involved. Second, when a test is comprised of a large number of content areas, it becomes increasingly difficult to evaluate the test blueprint using MDS. More dimensions are needed to identify the content structure and high-dimensional MDS solutions are difficult to interpret; especially when substantial differences exist among the SMEs.

To redress the limitations of the current procedure, some directions for future research are suggested. To reduce the demands on the SMEs, future research should explore incomplete MDS designs (e.g., Spence, 1982; 1983), where the SMEs are required to rate subsets of item pairings, and sorting procedures where the SMEs are required to sort the items into a limited number of piles (according to their similarity). If SME congruence is not a concern, the similarity data can be averaged over the SMEs to provide a single matrix for CMDS analysis. To test the adequacy of the test blueprint further, confirmatory MDS procedures (e.g., Borg & Lingoes, 1980) should also be investigated.

### Implications for test validity

To ensure that the content domain is represented adequately, test developers must demonstrate that: 1) the content specifications define the domain adequately, 2) the test blueprint represents the content specifications adequately, and 3) the test items are representative of the test blueprint. Given the paucity of publications regarding evaluations of test content (Sireci & Geisinger, 1992), it appears that test developers demonstrate only the first and third aspects of content representation.

To develop test specifications, developers of licensure tests typically conduct practice analyses to discover the knowledge and skills that are required for practice in the profession. For educational tests, the test specifications are typically constructed from surveys and reviews of national curricula and basal textbooks, and obtaining expert consensus regarding the subject matter to be tested. Though practice analyses, curricular reviews, and expert consensus help to ensure adequate definition of the content domain, they do not ensure the representation of the domain by the test itself. Therefore, they fall short of the requirements of content-related evidence for validity set forth in the AERA/APA/NCME *Standards* (1985):

> The first task for test developers is to specify adequately the universe of content that a test is intended to represent, given the proposed uses of the test ... Another important task is to determine the degree to which the format and response properties of the sample of items or tasks in a test are representative of the universe. (pp. 10-11)

It is hoped that this paper will remind test developers of the importance of ensuring the appropriateness of test content, in relation to the purpose of the test. It is further hoped that the procedure presented here will prove useful in helping test developers better evaluate the content representation of their tests.

## References

Aiken, L.R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40, 955-959.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Borg, I., and Lingoes, J.C. (1980). A model and algorithm for multidimensional scaling with external constraints on the distances. Psychometrika, 45, 25-38.

Carroll, J.D. and Chang, J.J. (1970). An analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 238-319.

Crocker, L.M., Miller, D., and Franks E.A. (1989) Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education, 2, 179-194.

Hambleton, R.K. (1980). Test score validity and standard setting methods. In R.A. Berk (ed.),Criterion-referenced measurement: the state of the art. Baltimore: Johns Hopkins University Press.

Hambleton, R.K., (1984). Validating the test score. In R.A. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press, pp. 199-230.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.) Educational measurement (2nd ed., pp. 130-159). Washington, D.C.: American Council on Education.

Johnson, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241-254.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling. Psychometrika, 29, 1-27, 115-129.

Lawshe, C.H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563-575.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C.

American Council on Education.

Morris, L.L., and Fitz-Gibbon, C.T. (1978). How to measure achievement. Beverly Hills: Sage.

Napior, D. (1972) Nonmetric multidimensional techniques for summated ratings. In Shepard, R.N.; Romney, A.K.; and Nerlove S.B. (eds.), Multidimensional scaling: Volume 1: Theory. New York: Seminar Press.

Oltman, P.K., Stricker, L.J., and Barrows, T.S. (1990). Analyzing test structure by multidimensional scaling. Journal of Applied Psychology, 75, 21-27.

Osterlind, S.J. (1989). Constructing test items. Norwell, MA: Academic Press.

Sireci, S.G., and Geisinger, K.F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. Applied Psychological Measurement, 15, 17-31.

Sokal, R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, 38, 1409-1438.

Spence, I. (1982). Incomplete experimental designs for multidimensional scaling. In R.G. Goledge & J.N. Rayner (Eds), Proximity and preference: problems in the multidimensional analysis of large data sets. Minneapolis: University of Minnesota Press.

Spence, I. (1983). Monte carlo simulation studies. Applied Psychological Measurement, 7, 405-426.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: Wiley.

Tucker, L.R. (1961). Factor analysis of relevance judgments: an approach to content validity. In A. Anastasi (Ed.). Testing problems in perspective: Twenty-fifth anniversary volume of topical readings from the invitational conference on testing problems, (pp. 577-586) Washington, D.C.: American Council on Education (1966).

Young, F.W., Takane, Y., & Lewyckyj, R. (1978). ALSCAL: a nonmetric multidimensional scaling program with several difference options. Behavioral Research Methods and Instrumentation, 10, 451-453.

Table 1

Content Area Blueprint for Auditing Items

| Content Area | Item # | Total |
|---|---|---|
| **Professional Responsibilities** | 46,47,48,49,51,52,53,56,57,58 | 10 |
| **Internal Control** | 26,28,29,30,32,35,36,37,38,39,41,42,43 | 13 |
| **Evidence & Procedures** | 1, 2, 3, 4, 5, 7, 10 | 7 |
| **Reporting** | 11,12,15,16,19,20,21,22,23,24 | 10 |
| **Total** | | **40** |

Table 2
Content Area Blueprint for Social Studies Test

| Content Area | Item # | Total # Items |
|---|---|---|
| Geography | 1, 7, 14, 15, 22, 36 | 6 |
| Economics | 18, 19, 20, 30, 31, 32 | 6 |
| History | 11, 23, 34, 38, 40 | 5 |
| Political Science | 5, 10, 13, 28 | 4 |
| Socio./Anthro. | 2, 9, 17, 25, 26, 37 | 6 |
| Interrelated | 3, 4, 8, 12, 21, 24, 33 | 7 |
| Applied | 6, 16, 27, 29, 35, 39 | 6 |
| Total: | | 40 |

Table 3

STRESS and RSQ Values for Auditing'SMEs:

Six-Dimensional INDSCAL Solution

| SME | STRESS | RSQ | SME | STRESS | RSQ |
|---|---|---|---|---|---|
| 1 | .12 | .78 | 9 | .14 | .62 |
| 2 | .12 | .76 | 10 | .14 | .61 |
| 3 | .16 | .52 | 11 | .09 | .84 |
| 4 | .14 | .62 | 12 | .18 | .33 |
| 5 | .11 | .78 | 13 | .18 | .32 |
| 6 | .17 | .42 | 14 | .17 | .41 |
| 7 | .13 | .69 | 15 | .13 | .68 |
| 8 | .07 | .93 | 16 | .12 | .74 |

Table 4

Auditing SME Subject Weights and Weirdness Indices

| | | Subject Weights | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | Dimension | | | |
| SME # | Weirdness | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | .39 | .48 | .23 | .56 | .29 | .25 | .20 |
| 2 | .18 | .53 | .43 | .34 | .21 | .29 | .21 |
| 3 | .06 | .40 | .31 | .24 | .27 | .27 | .24 |
| 4 | .08 | .45 | .30 | .28 | .27 | .29 | .31 |
| 5 | .13 | .63 | .28 | .27 | .28 | .27 | .28 |
| 6 | .09 | .35 | .26 | .21 | .26 | .26 | .23 |
| 7 | .10 | .53 | .38 | .26 | .28 | .24 | .26 |
| 8 | .28 | .51 | .38 | .44 | .50 | .23 | .14 |
| 9 | .09 | .48 | .34 | .24 | .26 | .28 | .26 |
| 10 | .11 | .44 | .28 | .27 | .30 | .24 | .33 |
| 11 | .07 | .50 | .37 | .36 | .38 | .33 | .27 |
| 12 | .16 | .27 | .22 | .20 | .22 | .20 | .27 |
| 13 | .13 | .30 | .23 | .17 | .22 | .22 | .23 |
| 14 | .13 | .31 | .27 | .20 | .26 | .25 | .26 |
| 15 | .10 | .50 | .36 | .25 | .27 | .28 | .30 |
| 16 | .13 | .53 | .31 | .41 | .24 | .26 | .27 |

Table 5
Average STRESS and RSQ Values for Auditing Data

Two- Through Six-Dimensional INDSCAL Solutions

| Dimension | STRESS | RSQ |
|---|---|---|
| 6 | .140 | .627 |
| 5 | .161 | .613 |
| 4 | .186 | .611 |
| 3 | .224 | .597 |
| 2 | .294 | .580 |

Table 6
Auditing Coordinate/Relevance Rating Correlations
(N=40 Items)

| Dimension: | Prof. | Internal | Evidence | |
|---|---|---|---|---|
| Interpret. | Resp. | Control | & Proc. | Reporting |
| 1: Reprt. vs. Int. Control | .16 | -.71** | -.50** | .87** |
| 2: Knowldg. of vs. Applct. of | -.48** | .50** | .36* | -.38* |
| 3: Common vs. Extraordinary | -.31* | .00 | .21 | .34* |
| 4: Int. Cntrl. vs. Evd.& Proc. | -.59** | .41** | -.42** | .24 |
| 5: Supplmntry vs. Field | -.15 | -.33* | .16 | .13 |
| 6: Senior vs. Entry-level | .23 | .19 | -.39** | .06 |

**$p$ < .01   *$p$ < .05

Table 7

Summary of Auditing Regression Analyses

| | Content Area | | | |
|---|---|---|---|---|
| Dimension | Prof. Resp. | Int. Cntrl. | Evd./Proc. | Report. |
| 1: Reprt/Int.Cntrl. | | *** | *** | *** |
| 2: Knldg/Applic. | *** | | * | *** |
| 3: Common/Extra. | ** | *** | * | *** |
| 4: IntCntl/Ev&Proc. | *** | *** | *** | *** |
| 5: Suppl./Field | | *** | *** | |
| 6: Senior/Entry | | *** | ** | |

* $\underline{p} \leq .05$ ** $\underline{p} \leq .01$ *** $\underline{p} \leq .001$

NOTES: Asterisks indicate significance of regression weights for regression of item relevance ratings across the stimulus coordinates for each dimension.

Table 8
Hierarchical Cluster Analysis on 6-D INDSCAL Coordinates for Auditing

```
Stage                          Items Clustered
 1              F B
 2                                                                L K
 3              F B A
 4                                        W V
 5                                                                        * $
 6                                                          Q P
 7                                                      U R
 8              F B A 9
 9                                    Y X
10                                                          Q P O
11                                                                    N I
12                              ! J
13                                                      U R Q P O
14                                                  4 2
15          & C
16                                                                                Z 1
17                      E 8
18                                                4 3 2
19                            @ ! J
20          & C F B A 9
21                                                                            T M
22                                            7 5
23                                                                L K N I
24                                                                            T M Z 1
25      G D
26                                        W V 6
27                                                                        * $ T M Z 1
28          & C F B A 9 E 8
29                          S @ ! J
30                                            7 5 3 4 2
31        H & C F B A 9 E 8
32                                    Y X W V 6
33                                                      U R Q P O L K N I
34      G D H & C F B A 9 E 8
35                                                      U R Q P O L K N I * $ T M Z 1
36                          S @ ! J Y X W V 6
37                                            7 5 3 4 2 U R Q P O L K N I * $ T M Z 1
38      G D H & C F B A 9 E 8 S @ ! J Y X W V 6
39      G D H & C F B A 9 E 8 S @ ! J Y X W V 6 7 5 3 4 2 U R Q P O L K N I * $ T M Z 1
```

| Content Area | Item Symbols |
|---|---|
| **Professional Responsibilities** | V, W, X, Y, Z, !, @, $, &, * |
| **Internal Control** | I, J, K, L, M, N, O, P, Q, R, S, T, U |
| **Evidence & Procedures** | 1, 2, 3, 4, 5, 6, 7 |
| **Reporting** | 8, 9, A, B, C, D, E, F, G, H |

Table 9

Summary of Averaged Auditing Content Area Relevance Ratings

| Content Area | # of Items | # Rated First | %First | # Rated Second | % 1st or 2nd |
|---|---|---|---|---|---|
| Professional Responsibilities | 10 | 7 | 70% | 3 | 100% |
| Internal Control | 13 | 11 | 85% | 2 | 100% |
| Evidence and Procedures | 7 | 6 | 86% | 1 | 100% |
| Reporting | 10 | 9 | 90% | 1 | 100% |

Table 10

Aiken's Validity Indices for Auditing Items

| Content Area | Item | Value | |
|---|---|---|---|
| **Professional Responsibilities** | 31 | .96**** | |
| | 32 | .95**** | |
| | 33 | .81**** | |
| | 34 | .87**** | |
| | 35 | .64* | |
| | 36 | .60 | |
| | 37 | .81**** | |
| | 38 | .96**** | |
| | 39 | .72** | |
| | 40 | .93**** | Average=.83 |
| **Internal Control** | 18 | .88**** | |
| | 19 | .59 | |
| | 20 | .93**** | |
| | 21 | .93**** | |
| | 22 | .96**** | |
| | 23 | .83**** | |
| | 24 | .58 | |
| | 25 | .88**** | |
| | 26 | .88**** | |
| | 27 | .79*** | |
| | 28 | .81**** | |
| | 29 | .69* | |
| | 30 | .85**** | Average=.82 |
| **Evidence & Procedures** | 1 | .79*** | |
| | 2 | .74** | |
| | 3 | .64* | |
| | 4 | .73** | |
| | 5 | .93**** | |
| | 6 | .93**** | |
| | 7 | .93**** | Average=.81 |
| **Reporting** | 8 | .98**** | |
| | 9 | .99**** | |
| | 10 | .97**** | |
| | 11 | .97**** | |
| | 12 | .99**** | |
| | 13 | .69* | |
| | 14 | .94**** | |
| | 15 | .98**** | |
| | 16 | .90**** | |
| | 17 | .94**** | Average=.94 |

***p < .001 **p < .01 *p < .05

Table 11
STRESS and RSQ Values for Social Studies SMEs

Two- Through Six-Dimensional INDSCAL Solutions

| SME | STRESS | RSQ | SME | STRESS | RSQ |
|---|---|---|---|---|---|
| 1 | .15 | .56 | 9 | .21 | .19 |
| 2 | .14 | .64 | 10 | .11 | .79 |
| 3 | .15 | .57 | 11 | .11 | .80 |
| 4 | .11 | .80 | 12 | .08 | .89 |
| 5 | .11 | .88 | 13 | .10 | .91 |
| 6 | .15 | .59 | 14 | .13 | .70 |
| 7 | .10 | .85 | 15 | .15 | .59 |
| 8 | .10 | .80 | | | |

Table 12

Average STRESS and RSQ Values for Social Studies Data

Two- Through Six-Dimensional INDSCAL Solutions

| Dimension | STRESS | RSQ |
|---|---|---|
| 6 | .13 | .70 |
| 5 | .15 | .70 |
| 4 | .17 | .69 |
| 3 | .21 | .67 |
| 2 | .26 | .65 |

Table 13
Social Studies Coordinate/Relevance Rating Correlations
(N=40 Items)

| Dimension (Interpretation) | Geog. | Econ. | Hist. | Pol. Sci. | Soc.-Anth. | App. | Comm. |
|---|---|---|---|---|---|---|---|
| Dimension 1 (other/geog) | .84** | -.12 | -.12 | -.68** | -.34 | -.57** | -.63** |
| Dimension 2 (other/econ) | .30 | -.79** | .23 | .36 | .10 | -.05 | .10 |
| Dimension 3 (low-/high-thnk) | -.23 | -.09 | .28 | .37* | -.20 | -.27 | -.11 |
| Dimension 4 (other/hist) | .12 | .42* | -.70** | .01 | -.43* | -.33 | -.12 |
| Dimension 5 (cultr/other) | -.49** | -.39* | .08 | .40* | .61** | .34 | .22 |
| Dimension 6 (intern/u.s.) | .20 | -.10 | .01 | -.27 | .26 | -.17 | -.31 |

**p < .01 *p < .05

Table 14
Summary of Social Studies Regression Analyses

| | Content Area | | | | | |
|---|---|---|---|---|---|---|
| Dimension | Geog. | Econ. | Hist. | Pol.Sc. | Soc./Ant. | Appl. |
| 1: Geog./Other | *** | * | | *** | * | *** |
| 2: Other/Econ. | *** | *** | | *** | | |
| 3: Low-T/High-T | ** | | * | * | * | ** |
| 4: Other/Hist. | | *** | *** | | ** | ** |
| 5: Cultr/Other | *** | *** | | ** | *** | |
| 6: Intrnt/U.S. | ** | | | * | * | |

* $p \leq .05$ ** $p \leq .01$ *** $p \leq .001$

NOTE: Asterisks indicate significance of regression weights for regression of item relevance ratings across the stimulus coordinates for each dimension.

Table 15: Results of Social Studies 6-D Cluster Analysis

| Stage | Items: Q P * N M O @ L $ C R D B 6 S A 5 W V U T K J I & ! Z Y X F G H E 8 3 9 2 7 4 1 |
|---|---|
| 1 | U T |
| 2 | V U T |
| 3 | J I |
| 4 | 41 |
| 5 | Z Y |
| 6 | ! Z Y |
| 7 | 8 3 |
| 8 | W V U T |
| 9 | A 5 |
| 10 | K J I |
| 11 | N M |
| 12 | ! Z Y X |
| 13 | D B |
| 14 | 7 4 1 |
| 15 | 92 |
| 16 | @ L |
| 17 | D B 6 |
| 18 | $ C |
| 19 | H E |
| 20 | S A 5 |
| 21 | * N M |
| 22 | Q P |
| 23 | ! Z Y X F |
| 24 | O @ L |
| 25 | G H E |
| 26 | R D B 6 |
| 27 | 9 2 7 4 1 |
| 28 | * N M O @ L |
| 29 | G H E 8 3 |
| 30 | G H E 8 3 9 2 7 4 1 |
| 31 | R D B 6 S A 5 |
| 32 | & ! Z Y X F |
| 33 | W V U T K J I |
| 34 | $ C R D B 6 S A 5 |
| 35 | & ! Z Y X F G H E 8 3 9 2 7 4 1 |
| 36 | Q P * N M O @ L |
| 37 | W V U T K J I & 1 Z Y X F G H E 8 3 9 2 7 4 1 |
| 38 | Q P * N M 0 @ L $ C R D B 6 S A 5 |
| 39 | Q P * N M 0 @ L $ C R D B 6 S A 5 W V U T K J I & 1 Z Y X F G H E 8 3 9 2 7 4 1 |

| Content Area | Item Symbol |
|---|---|
| Geography | 1, 7, E, F, M ! |
| Economics | I, J, K, U, V, W |
| History | B, N, Y, $, * |
| Political Science | 5, A, D, S |
| Socio./Anthro. | 2, 9, H, P, Q, @ |
| Interrelated | 3, 4, 8, C, L, O, X |
| Applied | 6, G, R, T, Z, & |

Table 16

Summary of Averaged Social Studies Content Area Relevance Ratings

| Content Area | # of Items | # Rated First | %First | # Rated Second | # Rated Third | % Rated in top 3 |
|---|---|---|---|---|---|---|
| Geography | 6 | 5 | 83% | 1 | 0 | 100% |
| Economics | 6 | 5 | 83% | 1 | 0 | 100% |
| History | 5 | 5 | 100% | 0 | 0 | 100% |
| Pol. Sci. | 4 | 4 | 100% | 0 | 0 | 100% |
| Soc./Anth. | 6 | 0 | 0% | 2 | 2 | 67% |
| Interrel. | 7 | 6 | 86% | N/A | N/A | 86% |
| Applied | 6 | 1 | 17% | 0 | 1 | 33% |

Table 17
Crosstabulation of Blueprint Specifications

and Highest Relevance Ratings for Each Item

| | Geog | Econ | Hist | PlSc | Soc | Appl | Intr | **Total** |
|---|---|---|---|---|---|---|---|---|
| Geog | **5** | | | | 2 | 2 | 1 | 10 |
| Econ | | **5** | | | | | | 5 |
| Hist | 1 | 1 | **5** | | 4 | 2 | | 13 |
| PlSc | | | | **4** | | 1 | | 5 |
| Soc | | | | | | | | 0 |
| Appl | | | | | | **1** | | 1 |
| Intr | | | | | | | **6** | 6 |
| **Total** | 6 | 6 | 5 | 4 | 6 | 6 | 7 | 40 |

Table 18

Aiken's Validity Indices for Social Studies Items

| Content Area | Item | Value | | |
|---|---|---|---|---|
| **Geography** | 1 | .86*** | | |
| | 7 | .99*** | | |
| | E | .87*** | | |
| | F | .97*** | | |
| | M | .79*** | | |
| | ! | .98*** | | |
| | | | Average: | .91 |
| **Economics** | I | .99*** | | |
| | J | .99*** | | |
| | K | .99*** | | |
| | U | .70** | | |
| | V | .73** | | |
| | W | .70** | | |
| | | | Average: | .85 |
| **History** | B | .90*** | | |
| | N | .76*** | | |
| | Y | .96*** | | |
| | $ | .93*** | | |
| | * | .78*** | | |
| | | | Average: | .87 |
| **Polit. Sci.** | 5 | .97*** | | |
| | A | .98*** | | |
| | D | .91*** | | |
| | S | .96*** | | |
| | | | Average: | .96 |
| **Soc./Anthro.** | 2 | .29** | | |
| | 9 | .65* | | |
| | H | .59 | | |
| | P | .46 | | |
| | Q | .41 | | |
| | @ | .48 | | |
| | | | Average: | .48 |
| **Applied** | 6 | .10*** | | |
| | G | .00*** | | |
| | R | .24*** | | |
| | T | .06*** | | |
| | Z | .03*** | | |
| | & | .04*** | | |
| | | | Average: | .08 |

***p < .001 **p < .01 *p < .05

Figure 1: Three-Dimensional Auditing Subject Space

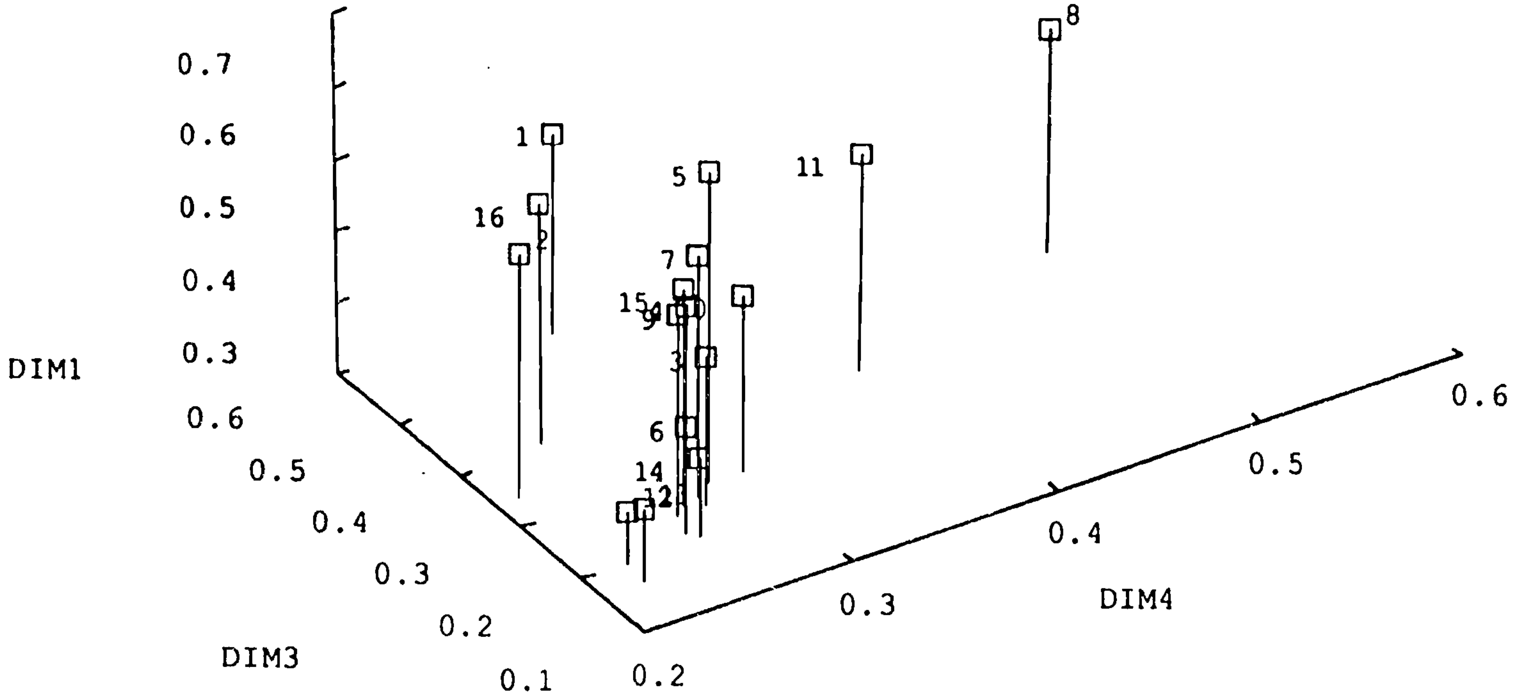
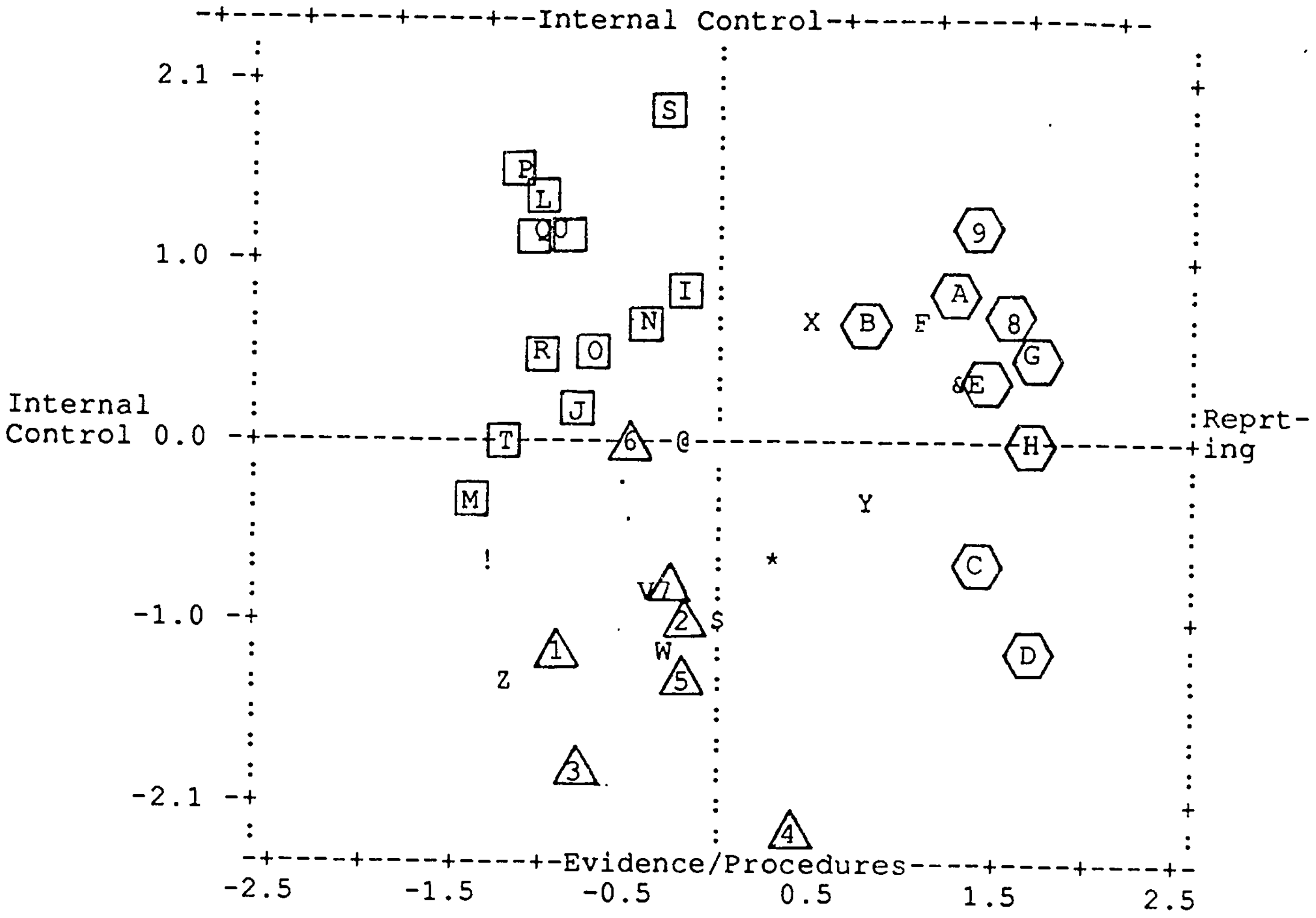


Figure 2: INDSCAL Stimulus Configuration for Auditing

Dimension 1 (Horizontal) Versus Dimension 4 (Vertical)



| Content Area & Symbol | Item Symbols |
|---|---|
| **Professional Responsibilities** | V, W, X, Y, Z, !, @, $, &, * |
| **Internal Control** ( □ ) | I, J, K, L, M, N, O, P, Q, R, S, T, U |
| **Evidence & Procedures** ( △ ) | 1, 2, 3, 4, 5, 6, 7 |
| **Reporting** ( ⬡ ) | 8, 9, A, B, C, D, E, F, G, H |

Figure 3: 3-D Subspace of Stimulus Configuration for CPA Data

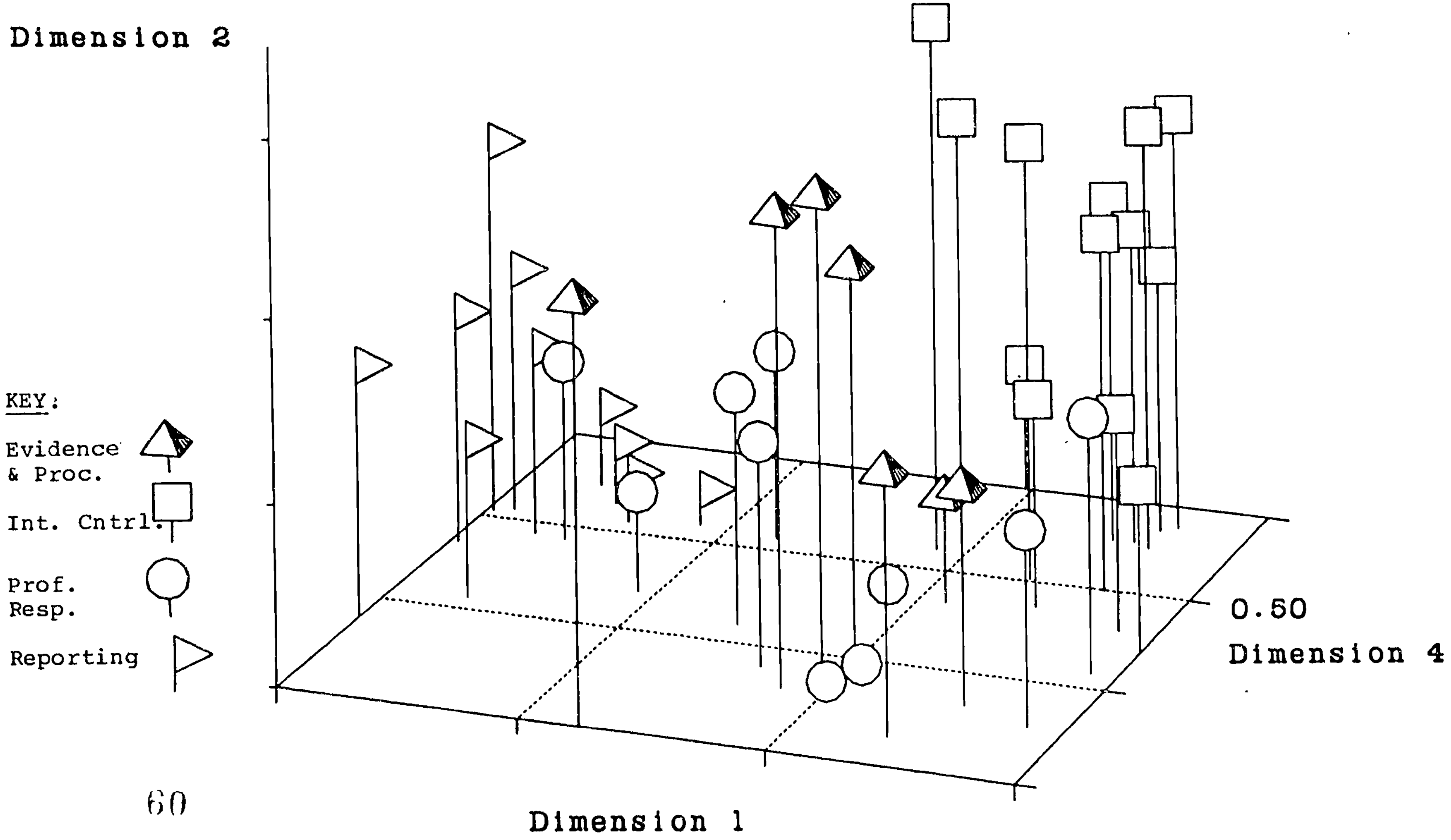


Figure 4: Three-dimensional Subject Space for Social Studies Data

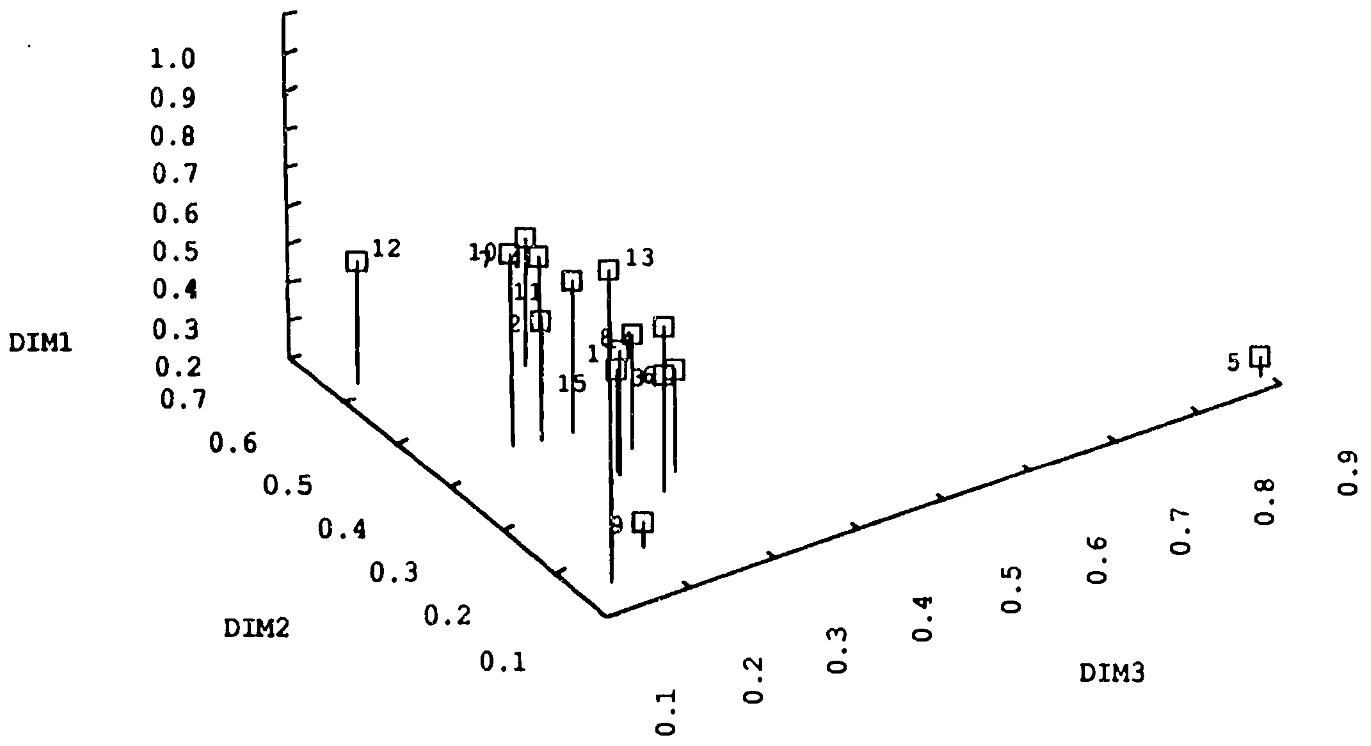


Figure 5: INDSCAL Stimulus Configuration for Social Studies

Dimension 1 (horizontal) versus Dimension 2 (vertical)

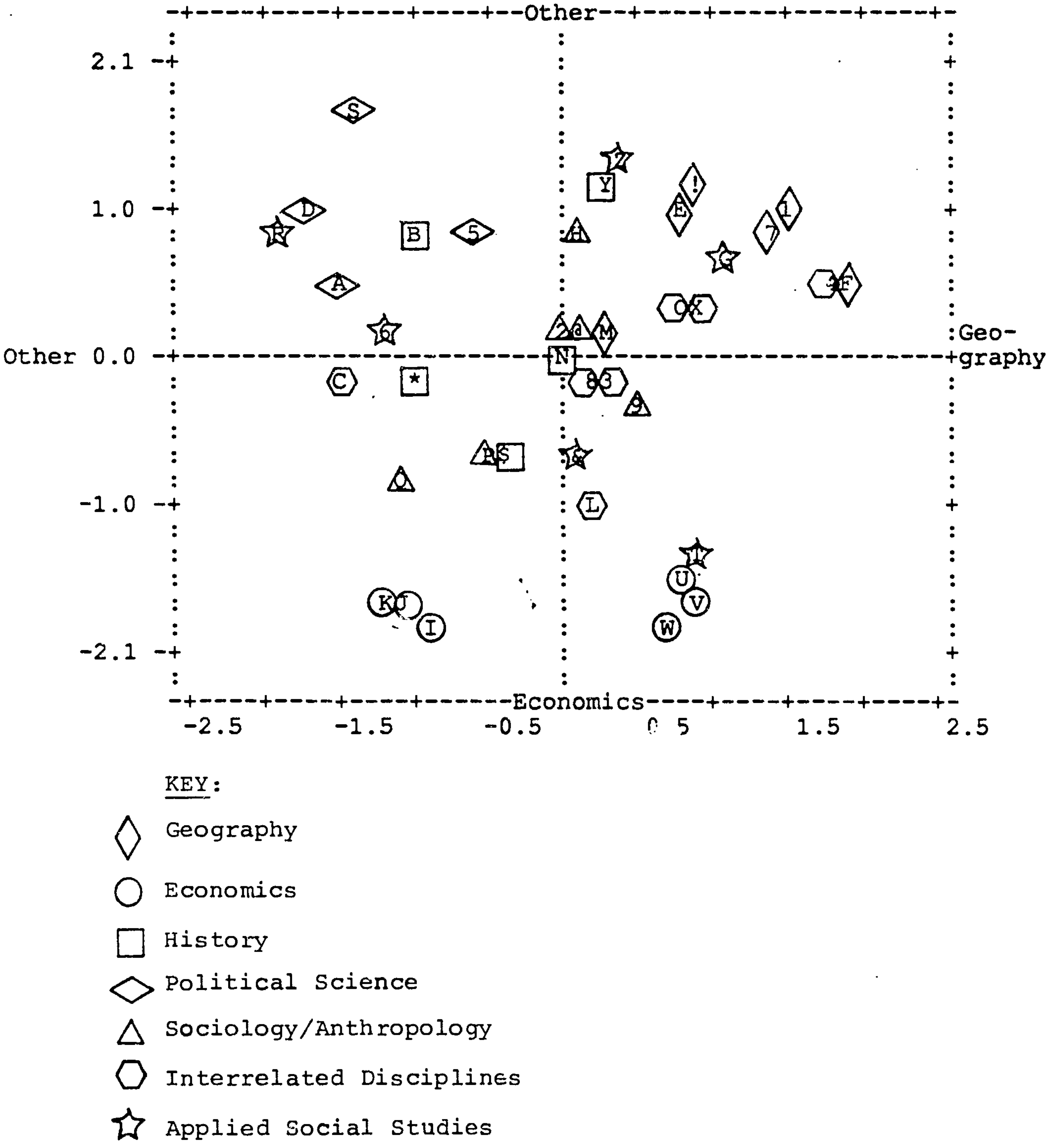
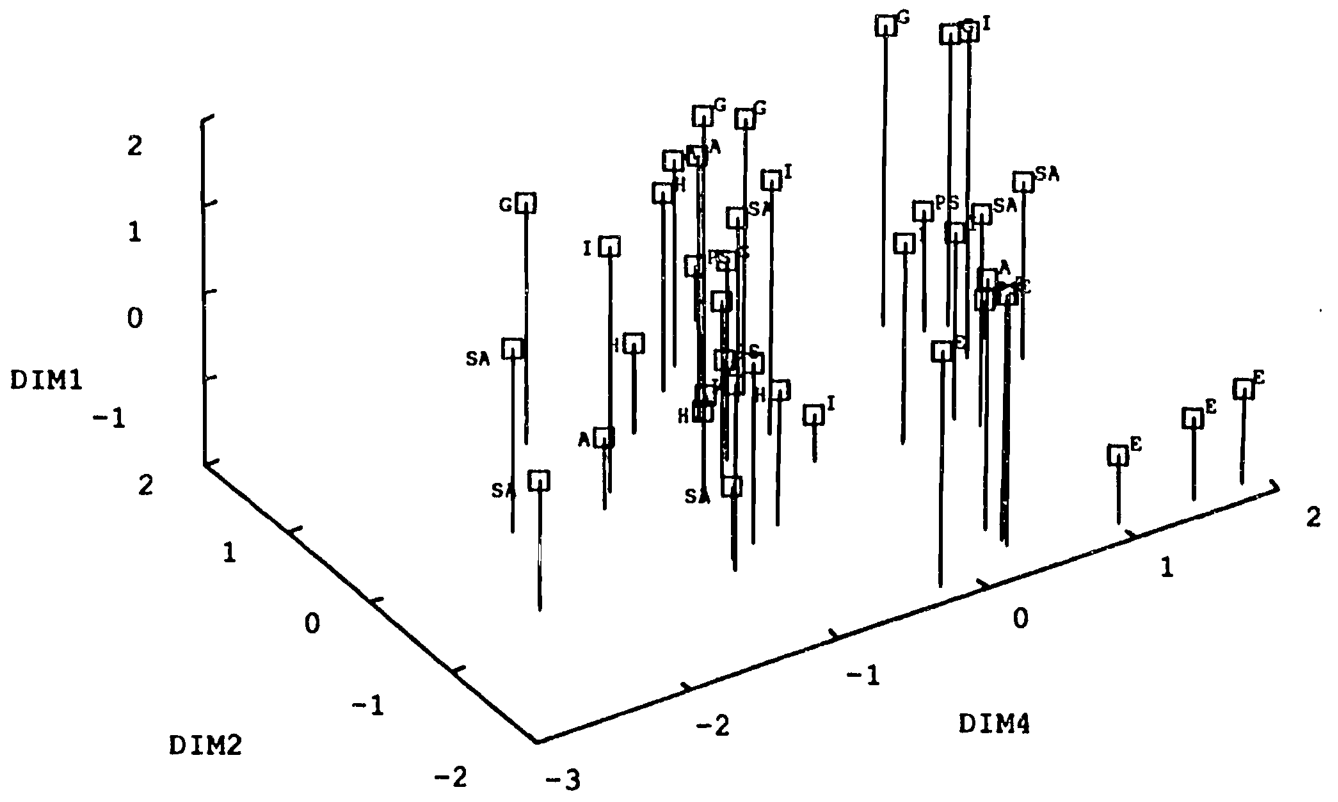


Figure 6 : Social Studies Stimulus Configuration: Dimensions 1, 2, and 4



KEY: G = Geography, E = Economics, H = History, PS = Political Science
SA = Sociology/Anthropology, I = Interrelated, A = Applied