DOCUMENT RESUME

ED 328 621 TM 016 173

AUTHOR	Sireci, Stephen G.; And Others
TITLE	Applying Empirical Analyses to the Evaluation of Test Content.
PUB DATE	Nov 90
NOTE	44p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, November 1, 1990).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cluster Analysis; *Content Analysis; Content Validity; Evaluators; *Interrater Reliability; Item Response Theory; Mathematical Models; *Multidimensional Scaling; Multiple Choice Tests; *Test Content
IDENTIFIERS	*Dimensional Analysis; Empirical Analysis; Experts; *Similarity Ratings

ABSTRACT

Although some researchers have argued against use of the term "content validity," the ability of a test item to adequately represent the domain of knowledge tested continues to be an issue of paramount importance in test construction. The present paper reviews previous analyses of test content and proposes a new empirical method for evaluating the content representativeness of a test. The proposed empirical method evaluates the content of a test by determining if similarity ratings of expert judges reflect the content structure specified in the test blueprint. Three expert judges rated the similarity of items on a 30-item multiple-choice test of study skills. The test was designed to assess the knowledge acquired by students at the end of a five-session study skills course. The test blueprint specified six content areas: study habits, time management, classroom learning, textbook learning, preparing for taking examinations, and taking examinations. The similarity data were used in a multidimensional scaling procedure to determine the dimensionality of the data. A subsequent cluster analysis was performed to determine whether the item clusters corresponded to the arrangement of items in the test blueprint. The results indicate a strong correspondence between the similarity data and the arrangement of items in the original test blueprint. Advantages of using item similarity data as an alternative to item response data are provided. Five data tables and six figures are included. A 29-item list of references is provided. (Author/TJH)

Applying Empirical Analyses to the Evaluation of Test Content[1]

Stephen G. Sireci

Kurt F. Geisinger

Soonmook Lee

Fordham University

A paper presented at the 1990 Annual Meeting of the Northeastern Educational Research Association: November 1, 1990

Ellenville, New York

[1] The authors would like to thank Bruce Biskin, Herman Friedman, Samuel Messick, Thanos Patelis, and Howard Wainer for their helpful comments on the work as it progressed. Gratitude is also extended to James O. Ramsay for his comments and loan of the MULTISCALE program. The responsibility for any errors or omissions are ours alone.

Abstract

Although some researchers have argued against use of the term "content validity," the ability of test items to adequately represent the domain of knowledge tested continues to be an issue of paramount importance in test construction. The present paper reviews previous analyses of test content and proposes a new empirical method for evaluating the content representativeness of a test. The proposed empirical method evaluates the content of a test by determining if similarity ratings of expert judges reflect the content structure specified in the test blueprint. Three expert judges rated the similarity of items on a thirty--item multiple-choice test of study skills. The similarity data were utilized in a multidimensional scaling procedure to determine the dimensionality of the data. A subsequent cluster analysis was performed to determine whether the item clusters corresponded to the arrangement of items in the test blueprint. The results indicated a strong correspondence between the similarity data and the arrangement of items in the original test blueprint. Advantages of using item similarity data as an alternative to item response data are provided.

Index terms: cluster analysis
content domain sampling
content validity
multidimensional scaling
similarity data
test construction

The term "content validity" has traditionally been used to refer to how well the items on a test represent the underlying domain of skill or knowledge tested (Thorndike, 1982). However, use of this term has been criticized by several theorists who describe validity as a unitary concept (Fitzpatrick, 1983; Guion, 1977; Messick, 1975, 1989a, 1989b; Tenopyr, 1977). The term "Content Validity" has become controversial since many current psychometricians define validity in terms of inferences derived from tests scores, while studies of content validation rarely employ item or test score data. Rather, a test's content is usually "validated" through more subjective methods such as ratings of test items by content experts (Osterlind, 1989).

While many psychometricians admonish use of the term "content validity," they reassert the importance of a test to adequately sample its underlying content domain . Guion (1978) employs the term "content domain sampling" to describe this representation, while Fitzpatrick (1983) recommends use of the term "content representativeness." Thus, regardless of the terminology expressed,the ability of a test to represent its underlying content domain continues to be an issue of paramount importance in test construction.

This paper presents a new approach of evaluating and assessing test content. In this approach, the data gathered from the content experts evaluating the test are analyzed with the techniques of multidimensional scaling and cluster analysis. The goal of the analyses are to determine the degree to which the test blueprint holds up to the independent ratings of test items by judges knowledgeable of the content domain. Since this procedure focuses on analysis of item content rather than test scores, the terms "content representation" and "content relevance" will be used to encompass the psychometric concerns traditionally associated with the term content validity.

I. Previous Research in Content Assessment:

R.L. Thorndike (1982) states that traditional studies of test content have generally been qualitative in nature. Frequently, content domain experts are employed to evaluate each test item to determine the content area to which it corresponds, and its relevance to the underlying domain. Although Thorndike concludes that "There has been relatively little attempt to develop quantitative indices of content validity" (p. 185), several methods have been established to quantify the subjective data gathered from content domain experts.

Crocker, Miller, and Franks (1989), and Osterlind (1989) provide informative summaries of "quantitative" methods of assessing the content representativeness of a test. The content indices reported by these authors are similar in two ways: 1) they all provide at least one quantitative summary of subjective data gathered from domain content experts, and 2) the content experts employed rate each test item in terms of its relevance and/or match to specified test objectives. The major differences between the methods surveyed are in the specific instructions given to the judges (content domain experts), and whether or not an item is allowed to correspond to more than one objective.

For example, Morris and Fitz-Gibbon (1978) provide an index of content coverage that involves a four-stage process: 1) A judgement is made to identify the objectives that match each item 2) the importance of each of the relevant objectives is ascertained, 3) the appropriateness of the item format is rated, and 4) the appropriateness of the estimated difficulty level of the item is evaluated. Three indices result from this analysis: the index of coverage, index of relevance, and a "grand average." This grand average represents an averaging of the importance, format, and difficulty ratings.

Hambleton (1980), provides a more straightforward assessment of an item's match to a specified objective. This "item-objective congruence index" is construed for criterion referenced tests where each item is linked to a single objective. The index reflects content domain experts' ratings, along a three-point scale (-1,0,+1), of the extent to which an item measures its specified objective in relation to its correspondence with the other objectives of the test. Hambleton (1984) provided a variation of this procedure designed to reduce the demand on the content domain experts.

Another noteworthy method of the assessment of test content is provided by Lawshe (1975). Lawshe quantified the subjective ratings of content domain experts in order to establish a measure of each item's relevance to a particular domain (in his case, job task domain). He employed a "Content Evaluation Panel" consisting of workers and supervisors knowledgeable of the job task domain to judge the relevance of each test item to a specified job. The proportion of judges who considered a given item "essential" to the performance of the job was calculated for each item. This proportion was labeled the "Content Validity Ratio" (CVR). Items with low CVR's were deleted from the test. Lawshe averaged the CVR's for all the retained test items to come up with an overall "Content Validity Index" which represented a quantitative index of the test's content relevance.

The preceding examples of "quantitative" indices of content representation demonstrate the traditional practices in the evaluation of test content. However, Crocker et. al., (1989), and Thorndike (1982) are quick to point out that these techniques are rarely used in practice. One reason for the lack of application is the procedures are often impracticable. For example, in Lawshe's (1975) example, a large number of domain content experts are required (175 in his first scenario, p. 569). Similarly, Cronbach's (1971) method (employing two teams of test constructors who develop alternate forms of the test), can be discounted in terms of its impracticality.

Another reason that these methods may lack popularity is that they may tend to implicitly support the content structure of the test. Presenting the judges with the content objectives of the test may bias the judges' ratings by imposing an ext nal structure upon their ratings. Indeed both Crocker et. al., (1989) and Osterlind (1989) recommend that the judges not be informed of the item-objective specifications of the test blueprint. Furthermore, Crocker et. al., (1989) point out that the item ratings often differ due to minor changes in the wording of the directions to the judges (pp. 181,182). Clearly, it would be beneficial to modify these procedures to avoid imposing the test blueprint on the content domain experts' judgements, and to gain economy of time, money, and manpower.

Analyses of item response data can be advanced as practical and economical methods of assessing the structure of a test. While factor analyses have been employed to discover if items which correspond to the same content area of the blueprint load on the same factor, the literature is replete with problems associated with factor analyzing dichotomous test data. Content analyses based upon test response data introduce confounds which are irrelevant to content representation and may bias such analyses. Item difficulty, the ability level and variability of the tested population, motivation, guessing, differential item functioning, social desirability, etc., are all variables that may affect the results of traditional factor analyses, but are irrelevant to assessment of content representativeness. The degree of relevance of an item to its corresponding content domain is a concept which is independent of the group's performance on the item. Analyses employing test response data allow the performance of the tested population to determine the relationships among the test items while ignoring inherent item characteristics. Although such analyses may be relevant in evaluations of construct or criterion-related validity, they are not central to evaluations of test content.

Napior (1972) expanded upon traditional analyses of test response data by applying multidimensional scaling and cluster analysis to the item-response correlation matrix. His procedure consisted of a classical multidimensional scaling of the items followed by a hierarchical cluster analysis of the items within the multidimensional space. Napior's intent was to identify unidimensional subsets of items within the test and to derive summary scores for each subset. Although Napior was not primarily concerned with test content, his procedure can be readily applied to the data obtained from content domain experts to provide information regarding the content structure of a test.

More recently, Oltman, Stricker, and Barrows (1990) applied Napior's (1972) method to item response data of the Test of English as a Foreign Language (TOEFL; Educational Testing Service, 1987). A primary purpose of the study was to investigate the utility of multidimensional scaling in the analysis of test structure. The authors concluded that the method was useful for analyzing test structure; however, the fact that an item difficulty dimension emerged illustrates the problem in using item response data in the assessment of content representation. The Oltman, et. al. study was not a direct assessment of the content relevance of the TOEFL; however, their findings provide evidence supporting the test blueprint and in supporting the construct validity of the test.

The current paper applies a procedure similar to Napior's (1972) method to data obtained from content expert ratings of the similarity of test items. Three content domain experts were employed as judges to rate the similarity of all items on a test to one another, according to their own criteria. The fit of the item similarity data to the test blueprint provided information useful in evaluating how well the test represented the defined content domain, and how well the test items corresponded to the content areas specified in the domain definition. The fact that the judges employed in this method were unaware of the content areas hypothesized in the test blueprint controls for a potential expectancy bias which may be unaccounted for in previous methods of content assessment (e.g., Lawshe, 1975). The next section describes the procedure, followed by a discussion of future implications of the method.

II. Method

1. Description of the Test

The content analysis was performed on a test of study skills (Sireci, 1988) constructed to test the knowledge acquired by students at the end of an five-session Study Skills course. This test is a thirty-item multiple choice exam keyed to the concepts and skills which were taught in the course. The blueprint of the test specified six content areas derived directly from the course syllabus.

The six content areas defined in the test blueprint were:

	Content Area	No. of Items
1.	Study Habits	4
2.	Time Management	4
3.	Classroom Learning	6
4.	Textbook Learning	6
5.	Preparing for Exams	5
6.	Taking Exams	5

The goal of the analysis was to discover how well the items on the SST represent the above six content areas.

2. Description of the Judges:

Three judges were employed to evaluate the similarity of the test items. Two of the judges had formerly taught a Study Skills course and were chosen for their knowledge of the subject domain. The third judge was a psychometrician with many years of experience in the construction and evaluation of educational tests. All the judges were ignorant of the blueprint of the test.

3. Procedure:

The Study Skills Test was distributed to each of the three judges independently. The original order of the items on the test was randomly scrambled using a Fortran random sorting program. This procedure was used to control for any order effect that may have influenced the judges' similarity ratings. The task of each judge was to "Judge how similar the test questions are to each other according to the following scale:" The scale presented to the judges was a 5-point Likert-type scale ranging from 1, "not at all similar," to 5, "extremely similar."

The judges were not given any criteria upon which to judge the similarity of the test items. This ambiguity in instruction was employed to avoid biasing the judges ratings in favor of supporting the test blueprint. The judges rated the similarity of every item pair and entered their ratings into a matrix. The matrix of all possible comparisons resulted in a 30 X 30 matrix. Since reciprocal comparisons were not requested, each judge provided a lower triangular matrix. The similarity ratings of all three judges were then used as input data for the multidimensional scaling (MDS) analyses.

A) Description of the MDS model:

Multidimensional Scaling is a set of mathematical procedures which attempt to reveal the underlying structure of a data set spatially, as in a map (Schiffman, Reynolds, and Young, 1981). The primary function of MDS is to convert measures of similarity (dissimilarity) to distances so that the relations between objects can be inspected visually. There are a variety of MDS models ranging from metric and nonmetric models, to those based on nonEuclidean distance. The analyses employed in this study were essentially nonmetric analyses defined in Euclidean space. There were two models employed; both models scaled the items using the Euclidean distance formula. The fundamental model employed was the Classical MDS model (Kruskal, 1964; Shepard, 1962). In this model, the distance between two objects, Dij, is defined according to the Pythagorean theorem:

$$D_{ij} = [\sum_{a=1}^{r} (X_{ia} - X_{ja})^2]^2 \quad (1)$$

where: D_{ij}=the Euclidean distance between points i and j,

X_{ia}=the coordinate of point i on dimension a,

$\sum$=summation over dimensions ranging from 1 to r;

r=the maximum dimensionality requested.

The other MDS model employed in this paper is the INDSCAL model (Caroll and Chang, 1970). INDSCAL is a weighted Euclidean distance model that represents an expansion of the Classical model in that each subject's dissimilarities are weighted by a factor wk corresponding to the relative emphasis subject k places on dimension r. Adding the weighting factor to equation (1) above gives us the INDSCAL model:

$$D_{ij} = [\sum_{a=1}^{r} w_{ka} (X_{ia} - X_{ja})^2]^{1/2} \quad (2).$$

The classical MDS and INDSCAL analyses were performed using the ALSCAL program in SPSSX (Takane, Young, and deLeeuw, 1977; Young, Takane, and Lewyckyj, 1978) on a VAX/VMS mainframe. ALSCAL is an "alternating least squares approach" (Young, et. al., 1978) to MDS that transforms the observed similarity data into distances which are then configured in the multidimensional space. The alternating least squares approach specifies a loss function which is minimized during the data transformation process. The data for all analyses were treated at the ordinal level and ties in the data were untied. Since the data obtained were measures of similarity, ALSCAL performed a simple transformation to convert them to measures of dissimilarity.

One other program, MULTISCALE (Ramsay, 1986) was used at one stage of the analysis to validate the ALSCAL analyses and confirm the selection of the final solution. MULTISCALE is a highly flexible MDS program that fits the distances to the dissimilarities through maximum likelihood estimation. The result of the maximum likelihood procedure provides a loglikelihood estimate that serves as an index of the relative fit of the distances to the dissimilarities. Since the absolute difference between the loglikelihoods of two models is distributed as a Chi Square statistic, this difference can be evaluated for significance. To test a significant improvement in fit from a more restrictive model to a more general model, the following test is employed:

$$X^2 = 2(L_n - L_u) \qquad (3)$$

where: L_n=loglikelihood for the more general model,

L_u=loglikelihood for the more restrictive model.

The X^2 value is evaluated at the degrees of freedom equal to the difference between the number of parameters incorporated in the two models.

B) MDS Analyses:

There were two types of MDS analyses conducted on the data. The first was an INDSCAL (Carrol and Chang, 1970) analysis employed to determine the relative consensus among the judges. The second analysis was a "Classical" MDS (Kruskal, 1964; Shepard, 1962) analysis which was performed on a single data matrix resulting from an averaging of the three individual matrices.

Step 1: Investigation of Inter-judge
Consensus: An Application of INDSCAL

Since each content domain expert (judge) provided a matrix of item similarity measures, there were three matrices analyzed INDSCAL. It was hoped that the judges would exhibit similar item configurations and dimension weights so that the stability of the solution could be verified. Unfortunately, the judges were somewhat different in their similarity ratings. Table 1 illustrates the STRESS (departure of data from the model), and RSQ (proportion of variance accounted for by the model). Table 2, illustrates the dimension weights for each of the three judges.

Table 1: STRESS, and RSQ Values for INDSCAL Analysis

Dimension		Average	Judge1	Judge2	Judge3
2	RSQ	.777	.760	.711	.861
	STRESS	.206	.213	.236	.162
3	RSQ	.807	.756	.753	.912
	STRESS	.164	.185	.185	.110
4	RSQ	.844	.803	.797	.931
	STRESS	.123	.136	.140	.088
5	RSQ	.859	.828	.825	.924
	STRESS	.101	.106	.111	.082

Table 2: Subject Weights from INDSCAL Analysis

5 Dimensional Solution:	Dim1	Dim2	Dim3	Dim4	Dim5
Judge 1	.5359	.3785	.3315	.3354	.4180
Judge 2	.4493	.4935	.4473	.3733	.2012
Judge 3	.5991	.6365	.2601	.2887	.0938
Overall[2]	.2826	.2640	.1259	.1117	.0747
4 Dimensional Solution:					
Judge 1	.4502	.5238	.3968	.4100	
Judge 2	.4682	.4300	.5276	.3388	
Judge 3	.6359	.5510	.4405	.1697	
Overall	.2754	.2543	.2099	.1039	
3 Dimensional Solution:					
Judge 1	.5927	.5537	.3139		
Judge 2	.6213	.4562	.3987		
Judge 3	.5783	.5866	.4827		
Overall	.3572	.2863	.1635		
2 Dimensional Solution:					
Judge 1	.5991	.6334			
Judge 2	.6774	.5024			
Judge 3	.6572	.6548			
Overall	.4166	.3608			

[2] Refers to the overall importance of each dimension, described as the proportion of variance accounted for by the dimension. The sum of the overall importance weights equals RSQ.

The STRESS and RSQ values listed in Table 1 indicate that there is a moderate degree of error unaccounted for in the data; especially for judges 1 and 2. The third judge is the least aberrant exhibiting consistently lower values of STRESS, and higher values of RSQ. Table 2 provides a more thorough comparison of the judges. In the four and five-dimensional solutions, it can be seen that the third judge has relatively smaller weights on the highest dimensions as compared to those of the other judges. In contrasting these weights with the subject weights obtained in the three-dimensional solution, it appears that the judges are using three similar dimensions in rating the similarity of the test items; the addition of a fourth or fifth dimension distinguishes between the first two judges and the third judge. The addition of a fourth or fifth dimension appears to be contributing information regarding individual differences among the judges. However, since only three judges were employed, and no data were gathered on the potential differential characteristics of the judges, an investigation of these differences could only be speculative.

Figure 1 displays selected dimensions of the subject space for the three-dimensional solution and the four-dimensional solution. In comparing these configurations, it can be seen that the judges are relatively similar in a three dimensional solution. Since the judges appear to be using three common dimensions, and since individual subject matrices often contain more error than an averaged matrix, the data matrices of the three judges were averaged to create a single matrix. This

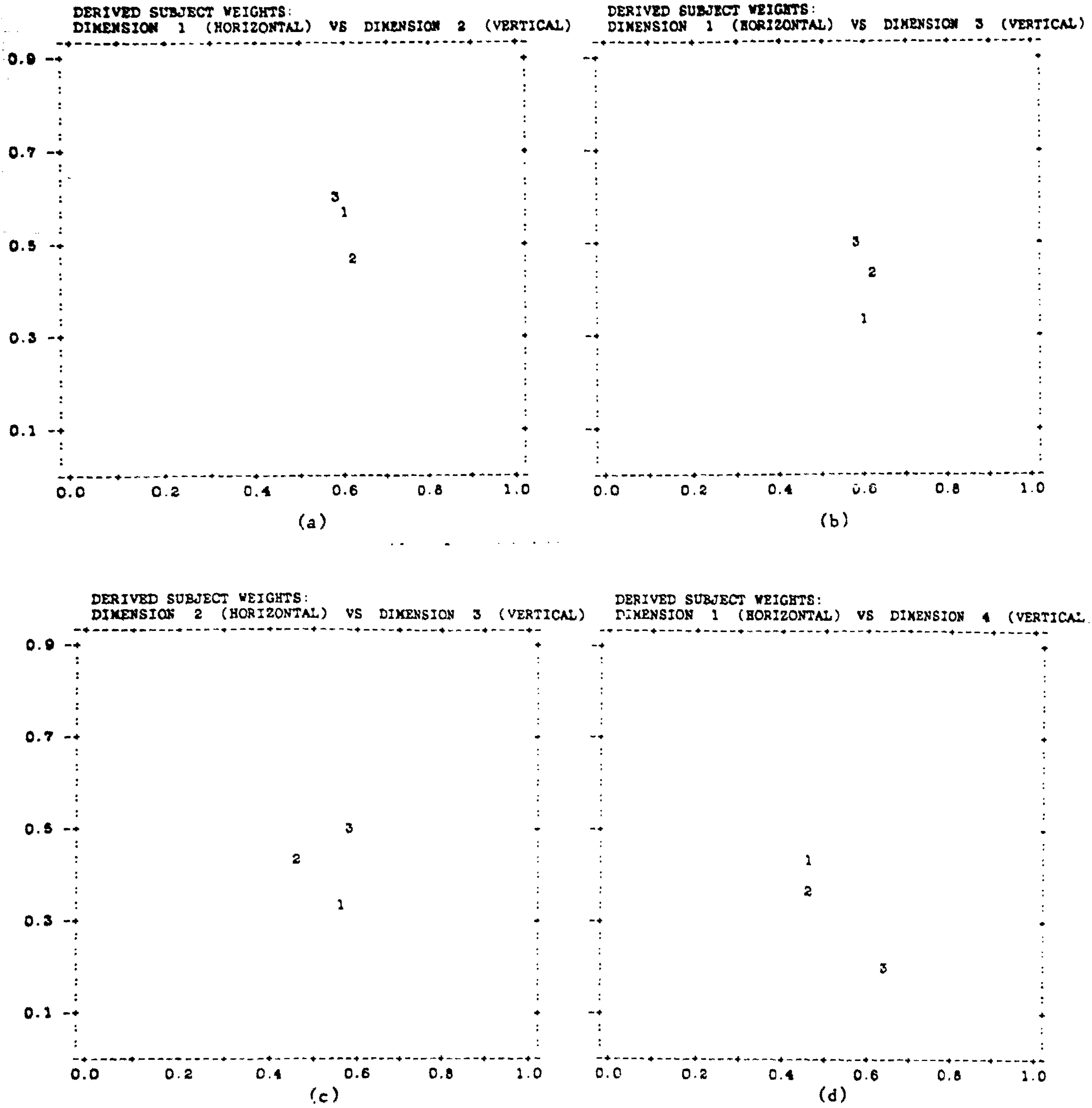


FIGURE 1: Subject space for 3-Dimensional solution (a through c), and dimensions 1 and 4 from the 4-Dimensional solution (d).

averaged item similarity matrix was entered into a classical MDS analysis to determine the appropriate dimensionality of the data. From the results of the INDSCAL analysis, it appears that a three-dimensional solution is minimally necessary.

Step 2: Investigation of Item Similarity: Classical MDS.

ALSCAL allows for a maximum of six dimensions to be fit to the data, and so six solutions were obtained ranging from the one through the six-dimensional solution. In order to identify the best-fitting solution, the fit measures RSQ (proportion of variance in the data accounted for by the model) and STRESS (departure of data from the model) were inspected for each model. Kruskal and Wish (1978) suggest that appropriate models should yield RSQ values greater than .90, and STRESS values between .05 and .10. Table 3 provides the RSQ and STRESS values obtained from the CMDS analyses.

The results from Table 3 indicate that the four, five, or six dimensional models are plausible solutions in terms of STRESS and RSQ. Kruskal and Wish (1978) recommend plotting the STRESS and RSQ values across incremental models to detect the points at which the acceleration or deceleration of these values occur. They assert that the dimensions over which these "elbows" occur are indicative of the appropriate dimensional

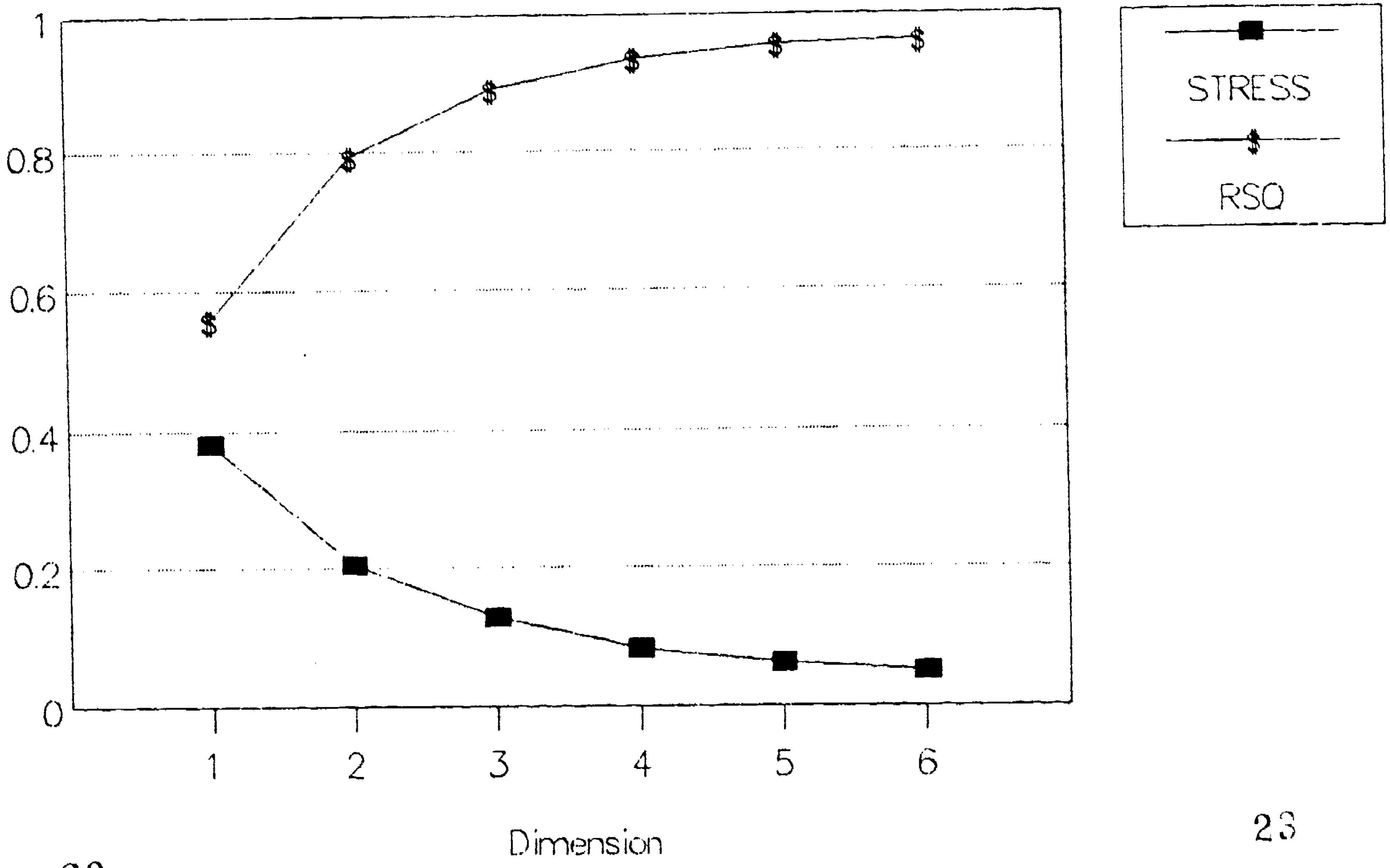
Table 3: CMDS STRESS and RSQ Values

Dimensional Solution	STRESS	RSQ
6	.050	.963
5	.061	.956
4	.082	.936
3	.127	.892
2	.204	.792
1	.383	.558

solution. The values of STRESS and RSQ have been plotted in Figure 2. It is difficult to detect an "elbow" for the STRESS curve, however, it can be seen that there is little improvement in STRESS after the five-dimensional solution. An elbow in the RSQ curve appears above the third dimension, and a slight bend in the curve appears above the four-dimensional solution. This inspection of Figure 2 implies that the three or four-dimensional solutions appear plausible.

In order to determine which model is more appropriate, the spatial configurations of each model were examined. Schiffman, Reynolds, and Young (1981) state that, "Dimensions which cannot be interpreted probably do not exist" (p. 12). The interpretation of our MDS solutions is greatly enhanced by our a priori knowledge of the stimuli; namely, our test blueprint. If there are dimensions that spatially represent the test items

Figure 2: Elbow plot from CMDS



in a way congruent with the test blueprint, then the dimensions are interpretable and the content relevance of the test is supported. In other words, if the judges, ignorant of the test blueprint, rated the similarities of the test items in such a way that the items clustered together similarly to their arrangement in the test blueprint, then there is strong evidence that the items are corresponding to the same content area. In order to determine the appropriate dimensional solution (model), and to see how well the MDS configurations match the test blueprint, arrangement of the items in the multidimensional space was inspected.

The four and five-dimensional solutions each revealed five clusters corresponding to five of the six content areas of the test blueprint (see Figures 3, 4, 5, and 6). Table 4 identifies the item symbols necessary to interpret the spatial configurations. If the predicted content areas of Time Management and Study Habits are merged, the MDS solutions perfectly reflect the test blueprint. The identification of the five content areas of the test appear within the first four dimensions of both models. The fifth dimension is not readily interpretable in the five-dimensional solution. Thus, the five-dimensional solution can be dismissed in terms of lack of interpretability, and lack of substantial improvement of RSQ and STRESS.

Given our knowledge of the characteristics of the test items (their blueprint specifications), the four dimensions are readily interpretable. Figures 3, 4, and 5 provide labels for the four dimensions derived from the configuration of the test items. The first dimension separates items relating to preparatory behaviors from items relating to active behaviors (e.g. "study habits" separated from "taking exams"). The second dimension distinguishes exam-specific behaviors from other behaviors (i.e., "taking exams" and "exam preparation" separated from all other content areas). The third dimension separates "textbook learning" from "classroom learning", while the fourth dimension differentiates "taking exams" from "exam preparation." Thus the four dimensions can be interpreted as follows: Dimension 1: Preparatory behaviors versus action behaviors, Dimension 2: Exam-related behaviors versus other behaviors, Dimension 3: Classroom learning versus textbook learning, Dimension 4: Taking Exams versus preparing for exams.

It is interesting to note that item 21, which is an item designed for the "taking exams" content subdomain, (labeled "L" in the MDS configurations), is not adequately separated from the "exam preparation" category. Similarly, item 6 can be viewed as an outlier (e.g., see Figure 6). Their relative positions in the configuration may indicate that these items are misclassified or poor, respectively.

Table 4: Item Symbols for MDS and Cluster Analyses

Content Area	Item Symbol
Study Habits	9, I, N, P
Time Management	2, 5, Q, M
Classroom Learning	4, B, C, E, G, O
Textbook Learning	1, 3, 8, A, F, R
Prep for Exams	S, 6, 7, H, U
Taking Exams	D, J, K, L, T

Though the four-dimensional solution appears to be the appropriate solution in terms of fit and interpretability, an argument can be made that the three-dimensional solution is equally plausible, based on the results of the previous INDSCAL analyses. In order to determine if the four-dimensional solution is indeed superior to the three dimensional solution, the averaged item similarity matrix was reanalyzed using the MULTISCALE MDS computer program (Ramsay, 1986). The MULTISCALE program was utilized to verify the findings of the ALSCAL solutions and to provide a statistical test of the difference in fit between the three and four-dimensional solutions. The Classical MDS analyses were run for the three and four dimensional models using a power transformation to fit the log of the dissimilarities to the log of the distances.

The configurations obtained through MULTISCALE were highly similar to those obtained in ALSCAL. The loglikelihood Chi

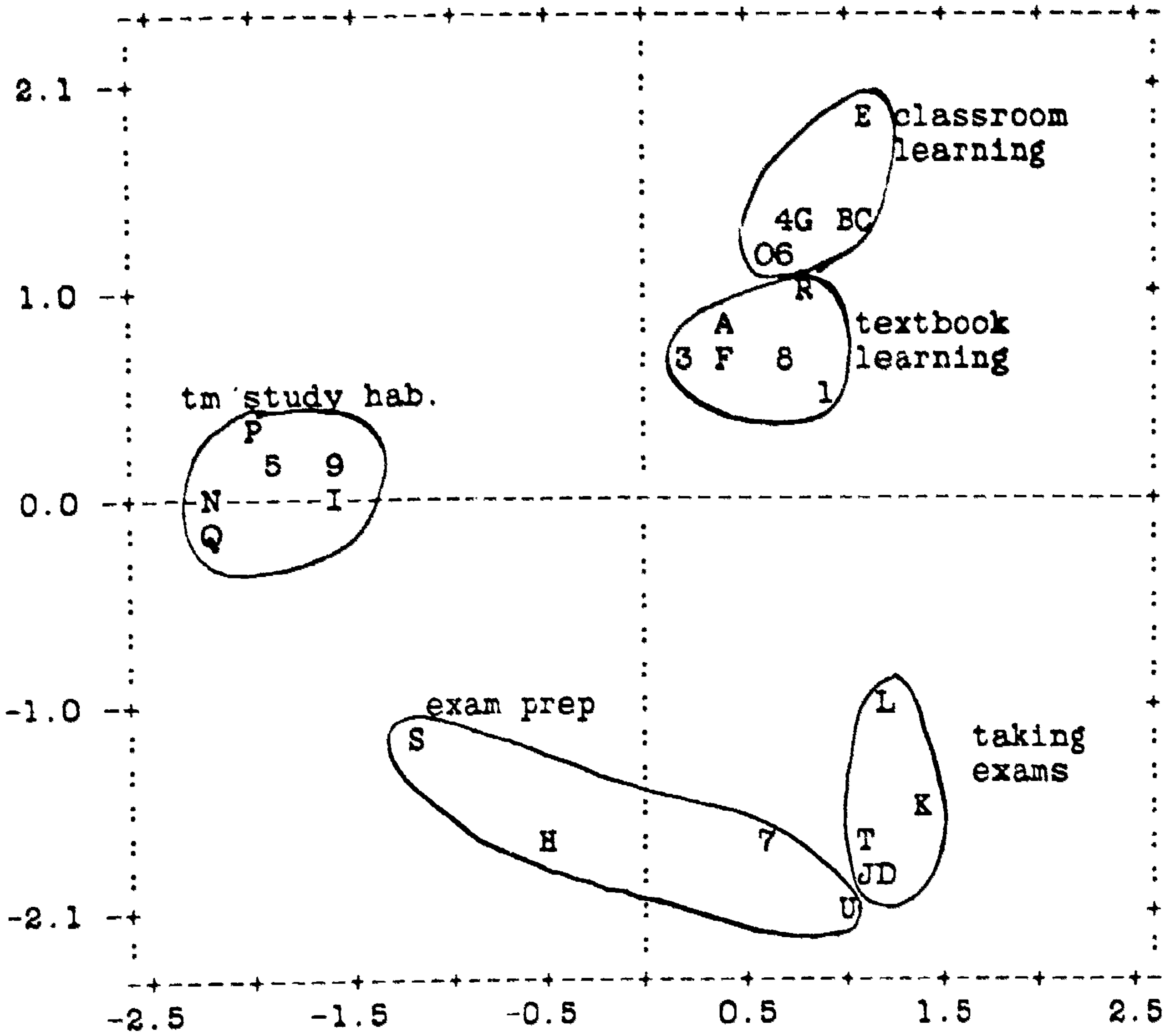


FIGURE 3: Five-Dimensional CMDS Solution

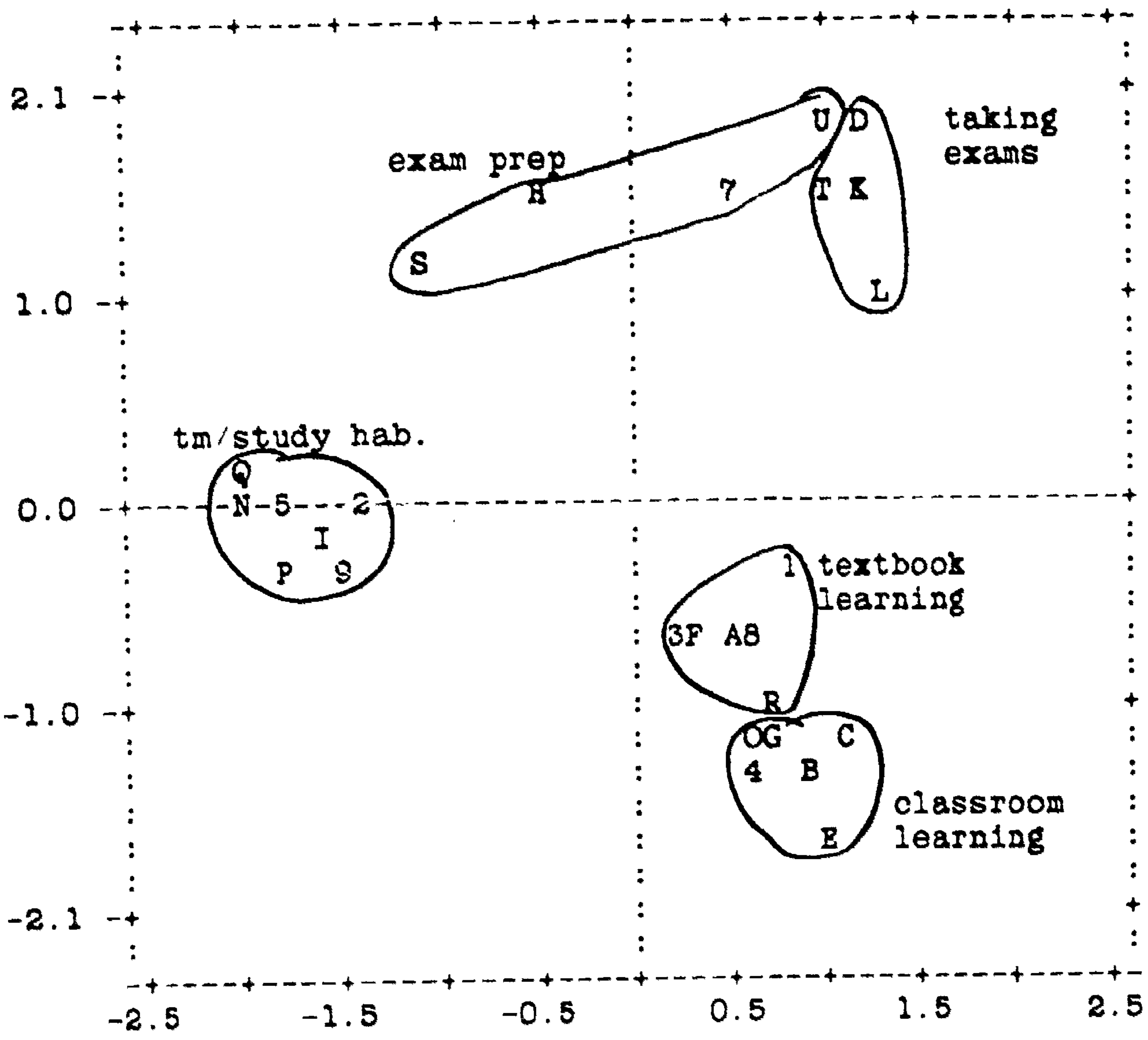


FIGURE 4: Four-Dimensional Solution: Dim 1 Preparations versus Actions (horizontal) plotted against Dim 2 Exam-specific versus Other, (vertical).

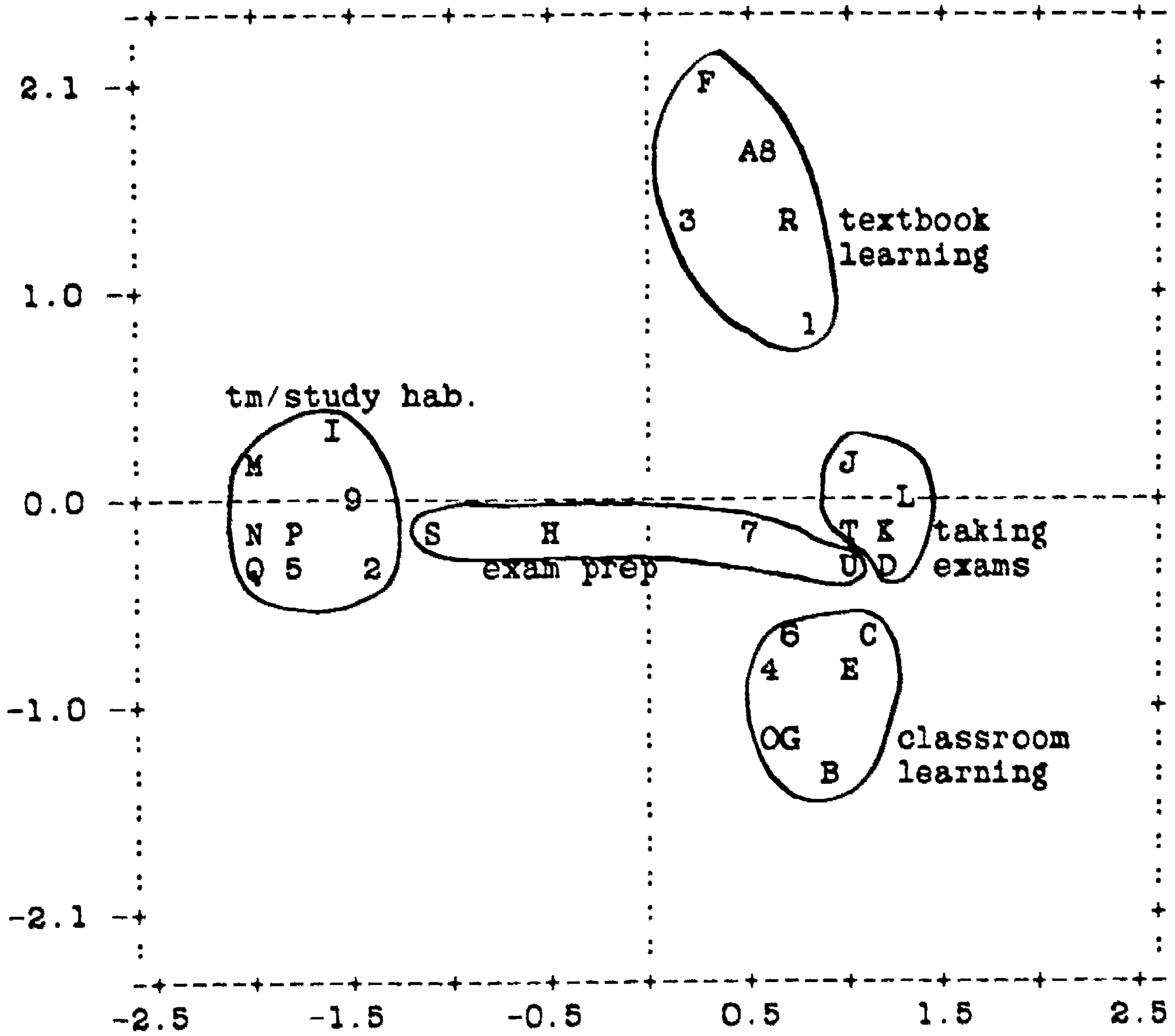
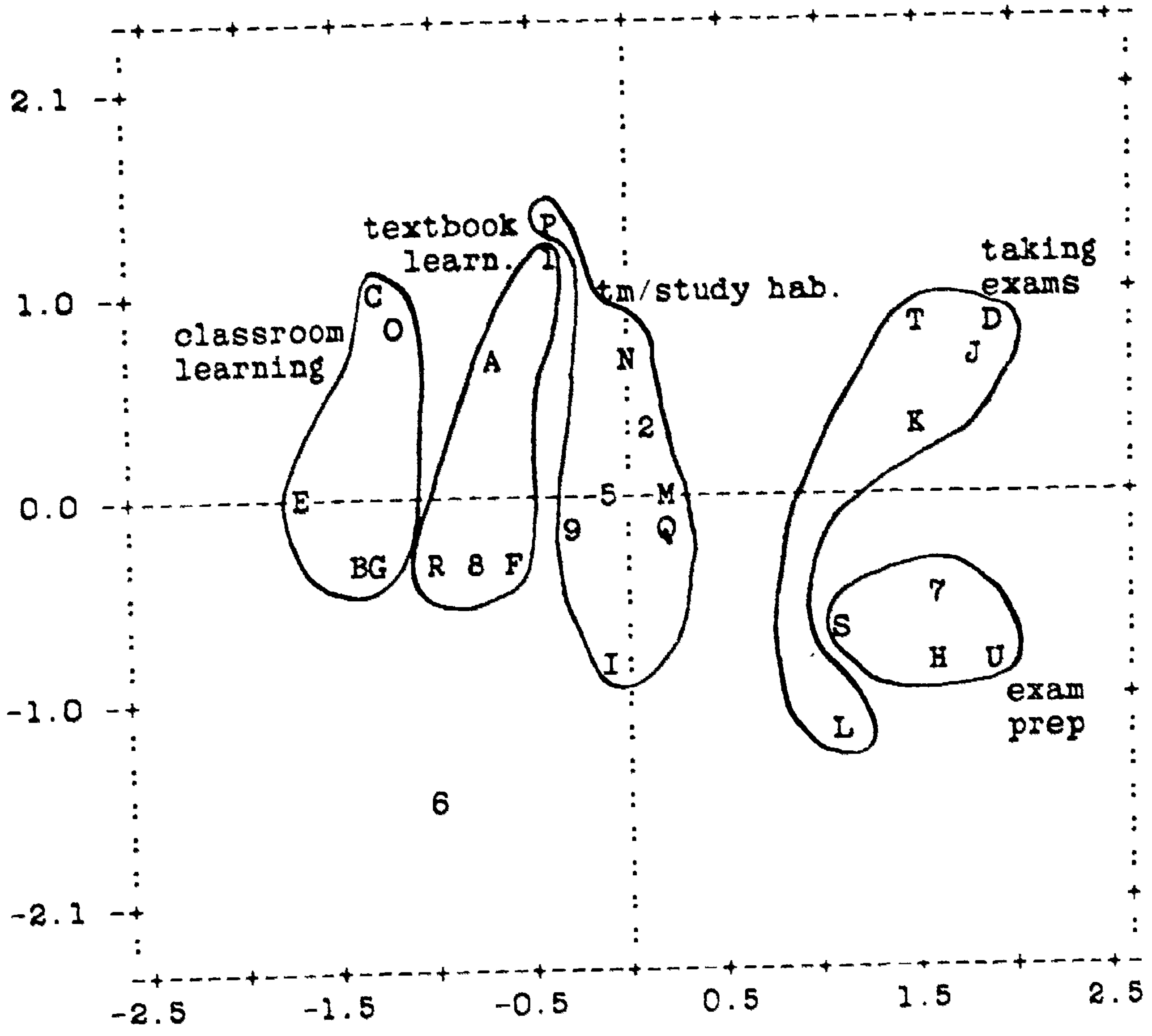


FIGURE 5: Four-Dimensional Solution: Diml (horizontal) against Dim 3 Classroom Learning versus textbook learning (vertical).



FIUGRE 6: Four-Dimensional Solution: Dim 2 (horizontal) against Dim 4 Taking Exams versus Preparing for exams (vertical).

Square test yielded a Chi Square of 116 with 26 degrees of freedom (111 parameters in the 4-D model minus 85 parameters in the 3-D model), which was significant at p<.001. Thus, the four-dimensional model is a better representation of the averaged item similarity matrix than is the three-dimensional model.

The preceding MDS analyses of the item similarity data represent a common approach for analyzing similarity data and interpreting the results. However, Kruskal and Wish (1978) suggest the use of a neighborhood or pattern approach to facilitate the interpretation of multidimensional configurations. Neighborhood approaches to interpreting data structure focus on the local proximity of items (large similarities), whereas MDS approaches focus on distances (large dissimilarities). As Kruskal and Wish put forth, "...neighborhood interpretations can be used to supplement and clarify dimensions rather than compete with them" (p. 45). In order to obtain a neighborhood interpretation, a hierarchical cluster analysis was performed on the disparity data. A hierarchical cluster analysis is the method of neighborhood interpretation advocated by Kruskal and Wish (1978), and also employed by Napior (1972) and Oltman et. al., (1990).

Step 3: Cluster Analyses:

In order to initiate the most effective cluster analysis, the coordinates obtained in the four-dimensional ALSCAL solution served as the input data for a hierarchical cluster analysis using the average linkage method. Through this method, the similarities among the items could be discovered in relation to the pertinent dimensions (within the multidimensional space). Napior (1972) advocates performing cluster analyses within the multidimensional space, rather than directly on the raw similarity data, so that the subtle and complex relations among the items can be uncovered. The dimensions in our model which separate the content areas of the test should reveal these more subtle and complex relations.

The hierarchical cluster analysis, employing the average linkage method, groups items into homogeneous categories by defining the (squared Euclidean) distance between two clusters as the average of the distances between all intercluster pairs. Items which are closest in terms of squared Euclidean distance (differences between dimension coordinates), are joined to form clusters. The solution is hierarchical in that an item's cluster membership can not change; once two items are joined within a cluster, they cannot be separated to join other clusters.

Table 5: Results of 4-dimensional Cluster Analysis

Clusters Representing Content Areas

Level	Items Clustered	Content Areas
7	22, 26, 5, 2	Time Management
20	16, 14, 11, 12, 24, 4	Classroom Learning
21	22, 23, 25, 9, 18, 26, 5, 2	Time Mgt/Study Habits
22	28, 17, 30, 7	Exam Preparation**
23	1, 10, 3, 8, 27, 15	Textbook Learning
24	13, 19, 20, 29, 21	Taking Exams

**Item 6 removed from content area

The results of the cluster analysis are presented in Table 5. With the exception of item 6, the above results exactly mirror the test blueprint consisting of five content areas. There is no overlap at all between content areas with regard to the individual items. It is interesting to note that item 6 was often an outlier in the MDS spatial configurations. It is also interesting to note that the first two content areas that merged were Time Management and Study Skills. It was the decision to combine these two content areas that advocated the acceptance of the four-dimensional MDS model. At the level where the data are condensed into five clusters (level 24), all five content areas of the test are represented as clusters (with the exception of item 6).

In figures 3, 4, 5, and 6, the items circled in the MDS space are items which formed clusters in the hierarchical cluster analysis. Napior (1972) recommended this method of circling the clusters in the MDS configurations in order to classify the scaled stimuli. It is relevant to note that the clusters are proximal to each other in multidimensional space, and that all five remaining content areas are represented as clusters.

III. Discussion

It is proposed that the procedure described above may be of significant utility in the evaluation of test content. It provides information regarding the efficacy of the test blueprint based upon judgements from domain experts unaware of the explicit content areas specified in the blueprint. The multidimensional scaling procedure can determine the appropriate number of dimensions needed to describe the similarity data, while the cluster analysis can determine whether or not the obtained dimensions are congruent with the arrangement of items specified in the test blueprint. Taken together, the two techniques provide a neutral means for testing the adequacy of the original test blueprint independent of test response data.

Potential Benefits of the Procedure:

There are several advantages in using MDS and cluster analysis in the evaluation of item similarity data. The analyses do not involve the administration of the tests and do not require a large number of domain experts, which leads to economy of time, persons and money. The method is much more economical in comparison to many of the more traditional methods (e.g., Lawshe, 1975). It should be noted that the present method employed only three judges. It is probable that more than three judges should be employed, but more than seven is probably not necessary. Osterlind (1989), for example recommends 4-5 judges for test of moderate size). Future research should employ larger groups of judges to determine if the dimensions and clusters are consistent across samples comprising different numbers of judges. The quality of the judges employed is another pertinent concern. For this method to be successful, it is imperative that the judges be knowledgeable in the specified content domain and that they are representative of the domain of all possible qualified judges.

These analyses also may aid the test construction process in the identification of ambiguous items. Items which are not perceived as similar to the remaining items would emerge as outliers in the multidimensional configurations. Such items should be inspected and/or modified. Another advantage of this method is that the judges are not influenced by the content areas that are predicted to emerge. Instead of trying to match test items with content areas, they rate the similarity of the items to each other. Since the judges are experts in the specific domain of knowledge it is presumed they will rate the similarity of the items with regard to their relation to the domain of knowledge tested. The method proposed here is less subjective than previous methods since the judges focus on the test items only, rather than on predetermined content specifications.

The proposed method may also be useful in detecting item bias toward specific minority or other concerned groups. Judges who are members of such groups could be employed to judge the similarity of the items. Their spatial configurations of test items could be compared to the spatial configuration of items derived from an original group of judges to determine if there are any discrepancies regarding individual items. Items that cluster differently between the two groups of judges may be flagged for potential bias. An INDSCAL analysis of the two groups of judges may also provide useful information. This potential application should be investigated in future research.

The method presented here may also be useful in providing information regarding the appropriate number of items to include in a given content area. If content areas overlap it may be due to an insufficient number of items in the various content areas. For example, if more test items were added to the Time Management subdomain, perhaps it would not cluster with the Study Habits subdomain. In this way the MDS/CA analyses can provide information regarding how well the content areas are defined.

It would be beneficial to compare the results from the present analysis with results from item analyses employing test response data. Minimally, it would be interesting to identify the item-to-total score correlations for those items which are outliers in the MDS solutions. If these correlations are relatively small, it may support the removal of those items from the test. Correlations of items within a content area to the total score of that content area would also be informative. Napior's (1972) method of multidimensional item analysis would provide results that could be directly comparable to the data collected via the present method. Davison (1985), describes advantages of using MDS on the intercorrelations of test items. Such analyses are likely to supplement the analyses of item similarity data.

Limitations of the Procedure:

There are some disadvantages to the proposed method which should also be noted. Rating the similarity of the items becomes increasingly more complex as the test size increases. A thirty-item test yields a 30 X 30 lower triangular matrix and 435 pairs of test items. The item similarity matrix increases exponentially as the test size increases. The larger the number of comparisons to be made, the greater the demand on the judges. This problem may be alleviated through a reduction in the number of necessary stimulus comparisons as Spence (1982, 198[illegible]) recommends, or by increasing the time interval required for the judges to make their comparisons. With a test of great length, the judges could complete their task over a period spanning a few days. Pairs of test items could be presented in subsets each requiring about an hour to complete. The judges could spend one to three hours a day until all subsets were completed. This would result in a complete data matrix while minimizing fatigue and frustration.

A major limitation of the present study was that data on item-domain relevance were not gathered. The addition of another step in the method presented would provide evidence pertaining both to domain definition and item-domain relevance. This step could be easily accomplished if after the judges have made their item similarity ratings, they would be handed descriptions of the content areas and asked how strongly each test item corresponded to each of the specified domains on a five-point scale. These "item-relevance" results would be utilized in a multiple regression procedure where the relevance data is regressed on the dimensions. If the items are relevant to their specified content areas, and the MDS solution supports the test blueprint, then the content areas should help in interpretation of the dimensions. For example, if six items were highly rated as corresponding to content area X, then all of the items within content area X should have similar loadings on the dimensions and the six items should cluster together in multidimensional space. Adding this multiple regression analysis to the present method would result in the following four-step procedure recommended for the analysis of test content:

1. Have a small number of content domain experts rate the similarity of the test items to each other and perform MDS analyses. Determine the similarity among the judges and the appropriate number of dimensions that best describes the data.

2. Next, provide the judges with descriptions of the content areas of the test, and have them express the degree to which each item corresponds to a particular content area.

3. Regress the item relevance data upon the dimensions obtained in the MDS solution. This analysis will determine if the dimensions obtained are consistent with the content areas defined in the test blueprint and will provide a measure of the relevance of each item to its particular domain.

4. Employ the dimension coordinates obtained from the MDS solution in a cluster analysis to confirm the MDS results in evaluating the fit of the obtained clusters to the original test blueprint.

Steps 1 and 4 would provide evidence pertaining to the adequacy of the test blueprint (domain definition) and steps 3 and 4 would provide evidence of item-domain relevance.

Summary:

Although the analysis of item similarity data has some limitations, the present results indicate its value to the evaluation of test content. However, it is important to keep in mind, that this technique merely confirms or disqualifies the test constructor's hypothesis of the test blueprint. It does not determine whether the test adequately measures the defined construct, or whether the test will have relevance in determining future criterion performance. These are issues of construct and criterion-referenced validity respectively, which (as is prominently pointed out by Guion, Fitzpatrick, Messick, and others), cannot be assessed independently of test scores. Thus, this proposed method is offered as a test construction tool, and as a method of gathering evidence to support the content relevance of a test. The information gathered through these analyses is limited when performed in isolation, but should prove valuable when viewed with results from item analyses and studies of construct and criterion-related investigations.

References

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.

Cronbach, L.J. (1971). Test Validation. In R.L. Thorndike (Ed.) *Educational measurement* (2nd ed). Washington, D.C. American Council on Education.

Davison, M.L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin*, *97*, 94-105.

Fitzpatrick, A.R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*, 3-13.

Guion, R.M. (1977). Content validity: the source of my discontent. *Applied Psychological Measurement*, *1*, 1-10.

Guion, R.M. (1978). Scoring of content domain samples: the problem of fairness. *Journal of Applied Psychology*, *63*, 499-506.

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.

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

Kruskal, J.B., and Wish, M. (1978). Multidimensional scaling. Newbury Park, CA: Sage.

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

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. Educational Researcher, 18, 5-11.

Messick, S. (1989b). Validity. in R. Linn (Ed.), Educational measurement, (3rd edt.). 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.

Ramsay, J.O. (1986). MULTISCALE II manual. Montreal, Quebec: McGill University.

Schiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). Multidimensional scaling: theory, methods, and applications. New York: Academic Press.

Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika, 27, 125-140.

Sireci, S.G. (1988) The SST: a test of study skills. Unpublished test, Fordham University: Bronx, NY.

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.

Takane, Y., Young, F.W., and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42, 7-67.

Tenopyr, M.L. (1977). Content-construct confusion. Personnel Psychology, 30, 47-54.

Thorndike, R.L. (1982). Applied psychometrics. Boston: Houghton Mifflin.

Young, F.W. (1981). Quantitative analysis of qualitative data. Psychometrika, 46, 357-388.

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.