Open access • Journal Article • DOI:10.1177/014662169201600102

Analyzing Test Content Using Cluster Analysis and Multidimensional Scaling

Stephen G. Sireci, Kurt F. Geisinger

Institutions: Fordham University

Published on: 01 Mar 1992 - Applied Psychological Measurement (SAGE Publications)

Topics: Item analysis and Multidimensional scaling

Analyzing Test Content Using Cluster Analysis and Multidimensional Scaling

Stephen G. Sireci and Kurt F. Geisinger
Fordham University

A new method for evaluating the content representation of a test is illustrated. Item similarity ratings were obtained from content domain experts in order to assess whether their ratings corresponded to item groupings specified in the test blueprint. Three expert judges rated the similarity of items on a 30-item multiple-choice test of study skills. The similarity data were analyzed using a multidimensional scaling (MDS) procedure followed by a hierarchical cluster analysis of the MDS stimulus coordinates. The results indicated a strong correspondence between the similarity data and the arrangement of items as prescribed in the test blueprint. The findings suggest that analyzing item similarity data with MDS and cluster analysis can provide substantive information pertaining to the content representation of a test. The advantages and disadvantages of using MDS and cluster analysis with item similarity data are discussed.
Index terms: cluster analysis, content validity, multidimensional scaling, similarity data, test construction.

The term "content validity" traditionally has 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). Content validity has become controversial because many current test specialists define validity in terms of inferences derived from test scores, but studies of content validation rarely employ 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).

In order to retain the importance of ensuring content domain representation and at the same time to avoid use of the term "validity," several test specialists have suggested replacing "content validity" with a more technically correct term. Messick (1975, 1980) suggested the terms "content relevance" and "content representation," Guion (1978) offered the term "content domain sampling," and Fitzpatrick (1983) recommended use of the term "content representativeness." Thus, regardless of the terminology employed, the ability of a test to represent its underlying content domain continues to be an issue of paramount importance in test construction.

The present paper presents and explores a new approach designed to evaluate the content representation of a test. Because item response data are not used in this approach, the terms "content representation" and "content representativeness" are used to encompass the psychometric concerns previously attributed to content validity.

Previous Methods of Evaluating Test Content

Evaluations of test content can be classified as either subjective or empirical. Subjective methods employ subject matter experts (SMEs) to evaluate and rate the relevance and representativeness of test items to the domain of knowledge tested. Examples of subjective methods for evaluating the content representation of a test are provided by Hambleton (1980, 1984), Lawshe (1975), and Morris and Fitz-Gibbon (1978).

Hambleton's (1980, 1984) method provides an item-objective congruence index that is appropriate for criterion-referenced tests in which each

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 17-31

0146-6216/92/010017-15$2.00

item is linked to a single objective. This index reflects an averaging of SME ratings regarding the extent to which items measure their specified objective. Lawshe's (1975) procedure results in a content validity index that represents an averaging of SME item relevance ratings for items retained on a test after the review process. The procedure developed by Morris and Fitz-Gibbon (1978) provides several indices that reflect an averaging of SME judgments regarding the objectives measured by each item, the relevance of these objectives to the purpose of the testing, the appropriateness of the item formats, and the appropriateness of the expected item difficulties.

Rather than relying on subjective evaluations of test items, some test theorists recommend empirical analyses of item response data to discover underlying content structure. Empirical studies of test content include applications of multidimensional scaling (MDS) and cluster analysis (Napior, 1972; Oltman, Sticker, & Barrows, 1990) and applications of factor analysis (Cattell, 1957). These analytic procedures result in dimensions, factors, and clusters that are presumed to be relevant to the content domains measured by the test. Davison (1985) reviewed applications of factor analysis and MDS to test intercorrelations. His review, accompanied by monte carlo comparisons of the two procedures, revealed that MDS often led to a more parsimonious representation of test structure than did factor analysis. This finding was explained by the presence of item difficulty factors that appeared in the factor analyses but did not emerge as dimensions in the MDS solutions.

The subjective methods of evaluating test content have been criticized severely due to their lack of practicality. Crocker, Miller, and Franks (1989) and Thorndike (1982) point out that these techniques are rarely used in practice. One reason for this lack of application is that subjective methods tend to support the content structure of a test implicitly, because presenting judges with the content objectives of the test may bias their judgments by imposing an external structure on their ratings. Indeed, both Crocker et al. (1989) and Osterlind (1989) recommended that the judges not be informed of the item-objective specifications of the test blueprint. Furthermore, Crocker et al. pointed out that item ratings often differ due to minor changes in the wording of the directions to the judges. Clearly, it would be beneficial to modify these procedures to avoid imposing the test blueprint on the content domain experts' judgments and to gain economy of time, money, and personnel.

The empirical methods of content assessment have also been criticized. These criticisms stem from the presence of content-irrelevant factors associated with item response data (Green, 1983). The ability of a test to represent its corresponding content domain is a quality inherent in the test independent of examinee responses to the items. Item difficulty, the ability level and variability of the examinee population, motivation, guessing, differential item functioning, and social desirability are variables that may affect the results of item response analyses but are irrelevant to assessment of content representation. Analyses employing test response data allow the performance of the tested population to determine the relationships between the test items while ignoring inherent item characteristics. Although such analyses may be relevant in evaluations of construct validity or criterion-related studies, they are not central to evaluations of test content (Messick, 1989b).

A pseudo-empirical study conducted by Tucker (1961) used factor analysis to evaluate test content, yet avoided the use of item response data. Tucker factor analyzed the ratings provided by the SMEs of the relevance of test items to the content domain tested. Two factors were identified: The first factor was interpreted as "a measure of general approval of the sample items" (p. 584). The second factor was interpreted as a measure of the differences between two groups of judges regarding which item types they considered more relevant (recognition items or reasoning items). Through this procedure, Tucker avoided the problems associated with factor analyzing dichotomous test data. However, his

study did not directly evaluate the content areas comprising the test. The items were not rated in terms of their relevance to the specific content areas of the test, and the arrangement of items in the test blueprint was not directly evaluated.

The present study borrows from Tucker (1961) in that subjective data obtained from SMEs were employed to evaluate test content. It deviates from Tucker's method by altering the instructions given to the SMEs and analyzing the data using MDS and cluster analysis, rather than factor analysis. MDS and cluster analysis were considered more appropriate than factor analysis in this case because similarity ratings were gathered, and because previous research (e.g., Napior, 1972; Oltman et al., 1990) demonstrated the ability of these techniques to uncover content-relevant test structure. The obtained item clusters also were compared directly with the test's blueprint specifications.

Method

Description of the Test

The content analysis was performed on a test of study skills (SST; Sireci, 1988) constructed to test the knowledge acquired by students at the end of a five-session Study Skills course. This test is a 30-item, four-option multiple-choice exam keyed to the concepts and skills that were taught in the course. The blueprint of the test specified six content areas derived directly from the course syllabus (see Table 1).

The Judges

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

Table 1
Study Skills Test Blueprint

Content Area	Item Symbols	Total Items
Study Habits	9, I, N, P	4
Time Management	2, 5, M, Q	4
Classroom Learning	4, B, C, E, G, O	6
Textbook Learning	1, 3, 8, A, F, R	6
Preparing for Exams	6, 7, H, S, U	5
Taking Exams	D, J, K, L, T	5
Total		30

Procedure

The SST was distributed to each of the three judges independently. The original order of the items on the test was randomly scrambled using a random sorting program. This procedure was used to control for any order effects that might 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 on which to rate the similarity of the test items. This ambiguity in instructions was used to avoid biasing their ratings in favor of supporting the test blueprint. The judges rated the similarity of every item pair and entered their ratings into a matrix. Because reciprocal comparisons were not requested, each judge provided a 30 × 30 lower-triangular matrix.

Analyses

A series of weighted MDS (INDSCAL) analyses were performed on the three matrices of item similarity ratings. The INDSCAL analyses were followed by hierarchical cluster analyses of the MDS stimulus coordinates.

INDSCAL analyses. The three similarity matrices were analyzed using the INDSCAL model with the ALSCAL MDS program of SPSSX (Takane, Young, & de Leeuw, 1977; Young, Takane, & Lewyckyj, 1978). INDSCAL is an individual differences MDS model that allows for differences among the raters in their relative weighting of the dimensions. The INDSCAL model was originally

formulated by Carroll and Chang (1970) and is a generalization of the classical MDS model introduced by Torgerson (1958) and later expanded by Shepard (1962) and Kruskal (1964). In the INDSCAL model, each rater's dissimilarity matrix is multiplied by a vector of weights (w) consisting of elements w_{ka} that represent the relative emphasis rater 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 distances between two objects i and j as:

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

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 of the solution.

Young et al. (1978) describe ALSCAL as an "alternating least-squares approach" to MDS that transforms the observed similarity data into distances that are subsequently configured in the multidimensional space. This alternating least-squares approach specifies a loss function (S-STRESS) that is minimized during the data transformation process. The original similarity data were by necessity transformed to dissimilarity data as a preliminary step in the MDS analysis. With dissimilarities, a larger number indicates greater dissimilarity, rather than greater similarity. For this dataset, values of 5 were converted to 1, values of 4 were converted to 2, and so forth. The data for all the ALSCAL analyses were treated as ordinal data, and ties in the data were untied using the primary approach to ties (Kruskal, 1964).

Cluster analyses. After the appropriate multidimensional solutions were identified, the item coordinates from these solutions were analyzed using the hierarchical cluster analysis program in SPSSX. The cluster analyses were performed on the item coordinates to identify homogeneous item subsets within the multidimensional space. An inspection of these clusters facilitated interpretation of the item configurations and provided a more direct comparison of the item groupings specified in the test blueprint.

The method of average linkage between groups (Johnson, 1967; Sokal & Michener, 1958) was used as the basis for the clustering. This method maximizes the average distance between all pairs of items belonging to different clusters. Items are joined to a cluster when their average similarity value is most similar to the average similarity values of the other members of the cluster.

Table 2
RSQ and STRESS Values for INDSCAL Analysis

Dimension and Index	Average	Judge 1	Judge 2	Judge 3
Dimension 2				
RSQ	.777	.760	.711	.861
STRESS	.206	.213	.236	.162
Dimension 3				
RSQ	.807	.756	.753	.912
STRESS	.164	.185	.185	.110
Dimension 4				
RSQ	.844	.803	.797	.931
STRESS	.123	.136	.140	.088
Dimension 5				
RSQ	.859	.828	.825	.924
STRESS	.101	.106	.111	.082

Results

MDS Analyses

Although one- through six-dimensional representations of the data were attempted, the six-dimensional solution could not be computed. This was due to the presence of a singular matrix, generated internally by ALSCAL, that could not be inverted. Thus, only one- through five-dimensional representations of the data were obtained.

Interjudge agreement. Table 2 presents the fit indices of STRESS (departure of data from the model) and RSQ (proportion of variance accounted for by the model) for the INDSCAL solutions. The STRESS goodness-of-fit index generated by ALSCAL is Kruskal's STRESS formula 1 (see Schiffman, Reynolds, & Young, 1981, p. 175) and should be distinguished from S-STRESS, the loss function minimized in computation of the coordinates. Table 3 provides the dimension weights for each of the three judges. Inspection of the individual dimension weights and the individual fit indices for each matrix (judge) revealed slight differences among the judges.

The individual STRESS and RSQ values listed in Table 2 indicate that there is a moderate degree of error variance 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 3 shows that in the four- and five-dimensional solutions, Judge 3 has relatively smaller weights on the highest dimension in comparison to the other judges. In contrasting these weights with the rater weights obtained in the three-dimensional solution, it appears that the judges are most similar in the lower-dimensional space (two and three dimensions), whereas the differences among the judges are magnified in the higher-dimensional solutions (four and five dimensions).

The addition of a fourth or fifth dimension appears to be contributing information regarding individual differences among the judges. How-

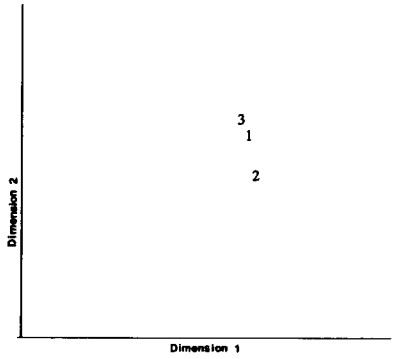
Table 3
Weights for the Judges from the INDSCAL Analysis

	Dimension				
Source	1	2	3	4	5
5-Dimensional Solution					
Judge 1	.5359	.3785	.3315	.3354	.4180
Judge 2	.4493	.4935	.4473	.3733	.2012
Judge 3	.5991	.6365	.2601	.2887	.0938
Overall*	.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			

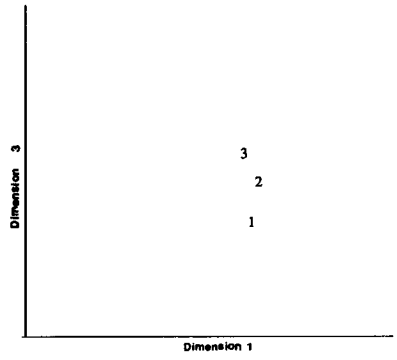
*Proportion of variance among the judges accounted for by the dimension. The sum of the weights across dimensions equals RSQ.

Figure 1
Rater Space From the Three-Dimensional INDSCAL Solution (1a through 1c), and a Subset of the 4-Dimensional INDSCAL Rater Space (1d)

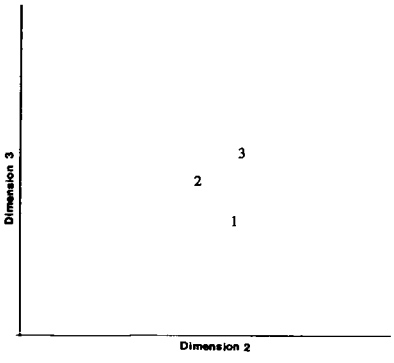
a. Dimension 1 Versus Dimension 2



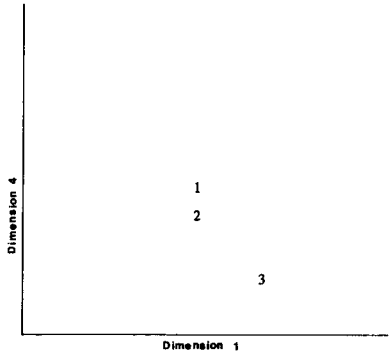
b. Dimension 1 Versus Dimension 3



c. Dimension 2 Versus Dimension 3



d. Dimension 1 Versus Dimension 4



ever, because no data were gathered on the differential characteristics of the judges and because the primary objective of the analysis was to discover information about the stimuli rather than the raters, an empirical investigation of these differences was not conducted.

Figure 1a through 1c display the rater space from the three-dimensional INDSCAL solution, and Figure 1d displays the first dimension plotted against the fourth dimension from the four-dimensional solution. In comparing these configurations, it can be seen that the judges are relatively similar in three-dimensional space, whereas the fourth dimension separates Judge 3 from the other two. Thus, in two- or three-dimensional space, the judges can be perceived as similar in their item ratings; in the higher-dimensional space, the judges appear less similar.

The average STRESS and RSQ values (see Table 2) indicate that there is a relatively large amount of error in these data. This finding may be due to the fact that the 5-point scale used to rate the stimuli was too restrictive and, therefore, more ties were present in the data than would be ideal. Davison (1983) stated that a scale containing 6 through 9 points "usually works quite well" (p. 42). However, many researchers using the method of paired comparisons employ larger scales. For example, Messick (1958) employed an 11-point scale; Wainer, Hurt, and Aiken (1976) and Wainer and Kaye (1978) employed a 15-point scale.

Selection of dimensionality. Two criteria often used to determine the appropriate dimensional representation of a dataset are fit and inter-

pretability (Davison, 1983). MDS solutions that fit the data well and are readily interpretable are desired. However, a paradox exists between fit and interpretability. Higher-dimensional solutions usually exhibit better fit but are usually more difficult to interpret.

The STRESS and RSQ values provided by ALSCAL indicated that the five-dimensional solution exhibited the best fit. To investigate further the best-fitting solution, these data were reanalyzed using the MULTISCALE-II computer program (Ramsay, 1981, 1986). Because MULTISCALE computes MDS distances using maximum likelihood, the log likelihood of each solution was used in a χ^2 difference test to determine whether the addition of another dimension provided significant improvement in fit (Ramsay, 1980, 1986). The results of this analysis also indicated that the five-dimensional solution exhibited the best fit. [However, not all researchers, e.g., Arabie et al. (1987), agree that this test is appropriate for the INDSCAL model.]

Although the five-dimensional solution exhibited the best fit, Kruskal and Wish (1978) pointed out that higher-dimensional solutions may provide better fit to the data because the spatial configurations adapt to random error. Therefore, goodness of fit is not a sufficient criterion for the selection of appropriate dimensionality.

Davison (1983) and Kruskal and Wish (1978) asserted that readily interpretable solutions are preferable over solutions containing uninterpretable dimensions, even if the higher-dimensional solutions exhibit superior fit. Certain MDS solutions are more intuitively appealing to investigators and, therefore, the interpretability criterion often carries the most weight in determining dimensionality (Arabie et al., 1987, p. 36). However, accepting solutions based on interpretability is controversial. For example, Schiffman et al. (1981) asserted that "Dimensions which can not be interpreted probably do not exist" (p. 12), although Kruskal and Wish (1978) pointed out that "the fact that a particular investigator can not interpret a dimension does not mean that the dimension has no interpretation" (p. 57). Because the focus of this analysis was to obtain information regarding the relationships among the test items, acceptance of the higher-dimensional configurations will be contingent on their interpretability. In this study, interpretation of the solution space was enhanced by a priori knowledge of the items—namely, their content area specifications designated in the test blueprint. Therefore, interpretation of the stimulus space focused on item groupings that reflected common content attributes.

Figures 2 through 5 present selected stimulus configurations from the two- through four-dimensional INDSCAL solutions. The item content area specifications and the item symbols necessary to interpret the configurations are presented in Table 1. In the INDSCAL model, the orientation of the stimulus space is unique. Therefore, rotation of the dimensions is not necessary and the dimensions are directly interpretable. This feature obviates the need to search for other meaningful directions in the stimulus

Figure 2
Stimulus Configuration From the Two-Dimensional INDSCAL Solution

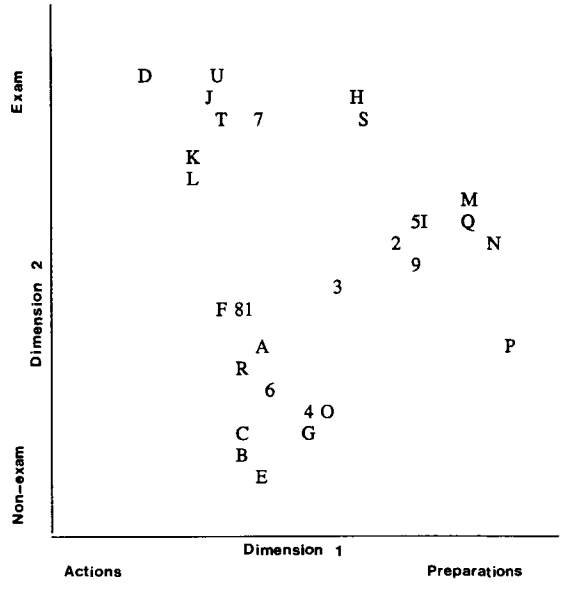
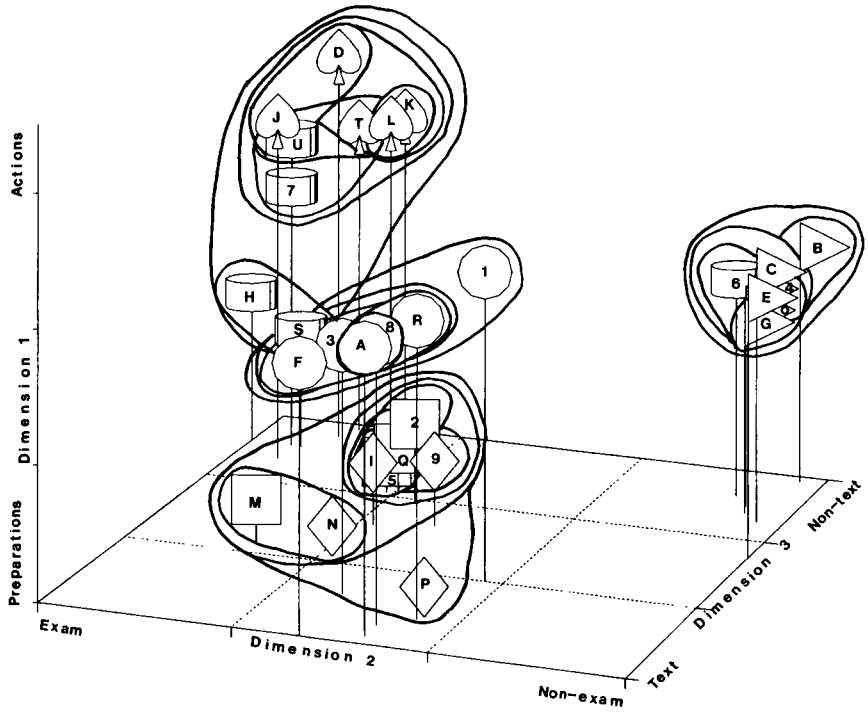


Figure 3
Three-Dimensional INDSCAL Stimulus Configuration Illustrating Item Content Area Specifications: Study Habits (Diamonds), Time Management (Squares), Classroom Learning (Flags), Textbook Learning (Balloons), Preparing for Exams (Cylinders), and Taking Exams (Spades); Ellipsoids Illustrate Results From Hierarchical Cluster Analysis



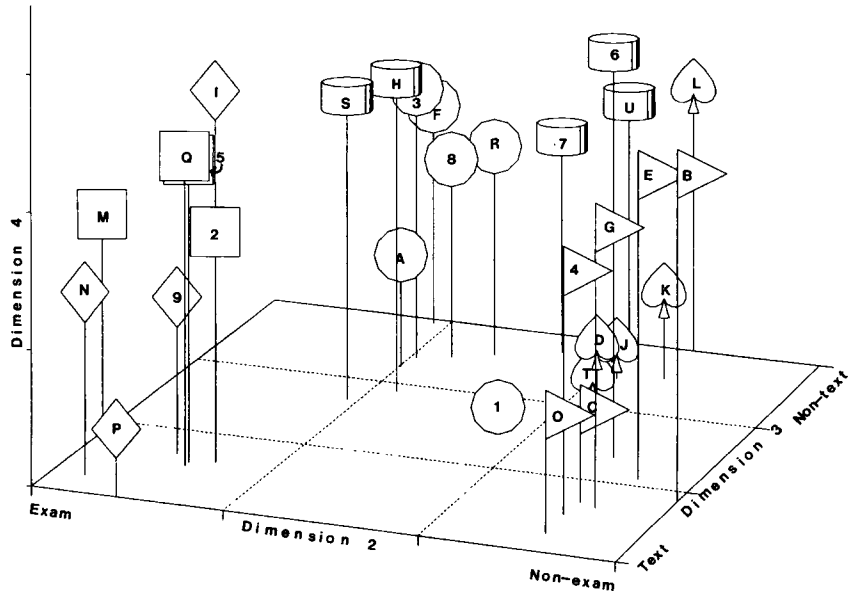
space (Arabie et al., 1987; Schiffman et al., 1981).

Figure 2 presents the stimulus configuration resulting from the two-dimensional solution. Dimension 1 (horizontal)—labeled Actions versus Preparations—appears to separate the content areas pertaining to scheduling and organizing activities (Time Management, Study Habits) from the more action-oriented activities (Taking Exams, Textbook Learning). Dimension 2 (vertical)—Exam-Related Versus Non-Exam Related—appears to separate the content areas pertaining to exam-related activities (Preparing for Exams, Taking Exams) from those less related to examinations (Classroom Learning, Textbook Learning).

The three-dimensional solution is presented in Figure 3, in which common symbols represent the content area specifications of the items. The first two dimensions reflect the dimensions obtained in the two-dimensional solution. The addition of the third dimension—Text-Related Versus Non-Text Related—further separated the content area of Textbook Learning from the other content areas. With only a few exceptions, items corresponding to the same content areas were perceived as highly similar by the SMEs. The circles surrounding the items in Figure 3 reflect the results of the cluster analysis that are described below.

The four-dimensional solution was more difficult to interpret than the lower-dimensional solutions. Dimensions 1 through 3 reproduced the three dimensions obtained in the three-dimensional solution, but the fourth dimension was not readily interpretable in terms of the content attributes of the items; it distinguished somewhat between the content areas of Taking Exams and Preparing for Exams. A three-

Figure 4
Three-Dimensional Subspace From the Four-Dimensional INDSCAL Stimulus Configuration
(Content Area Symbols are the Same as Figure 3)



dimensional subset of the four-dimensional solution that illustrates this distinction is presented in Figure 4. All possible three-dimensional subsets of the four-dimensional solution are not presented because the first three dimensions were highly similar to the three-dimensional solution.

The five-dimensional solution resulted in configurations highly similar to the lower-dimensional solutions. Three dimensions emerged corresponding to those noted in the three-dimensional solution and another dimension separated "Taking Exam" items from "Preparing for Exams" items. However, the fifth dimension could not be interpreted, even when considering characteristics of the items not relevant to content (e.g., positively worded versus negatively worded items). Thus, the five-dimensional solution was not regarded as a valid representation of the data. The results generally indicate that at least three dimensions were necessary to distinguish between the content areas of "Classroom Learning" and "Textbook Learning," and that four dimensions were necessary to distinguish between the content areas of "Taking Exams" and "Preparing for Exams."

Cluster Analysis

Tables 4 and 5 show the items clustered at each stage of the clustering solution. In Figure 3, circles have been placed around those items that formed clusters at each stage, with the exception of the last three stages where the major content areas were collapsed.

The cluster analysis performed on the three-dimensional coordinates revealed two item clusters that corresponded directly to two of the content areas specified in the test blueprint, "Classroom Learning" and "Textbook Learning" (see Table 4 and Figure 3). The three-dimensional cluster analysis merged the content areas of "Time Management" and "Study Habits" to form one cluster, and merged "Taking Exams" and "Preparation for Exams" to form another. One exception to these item groupings was Item 6, whose content area designation

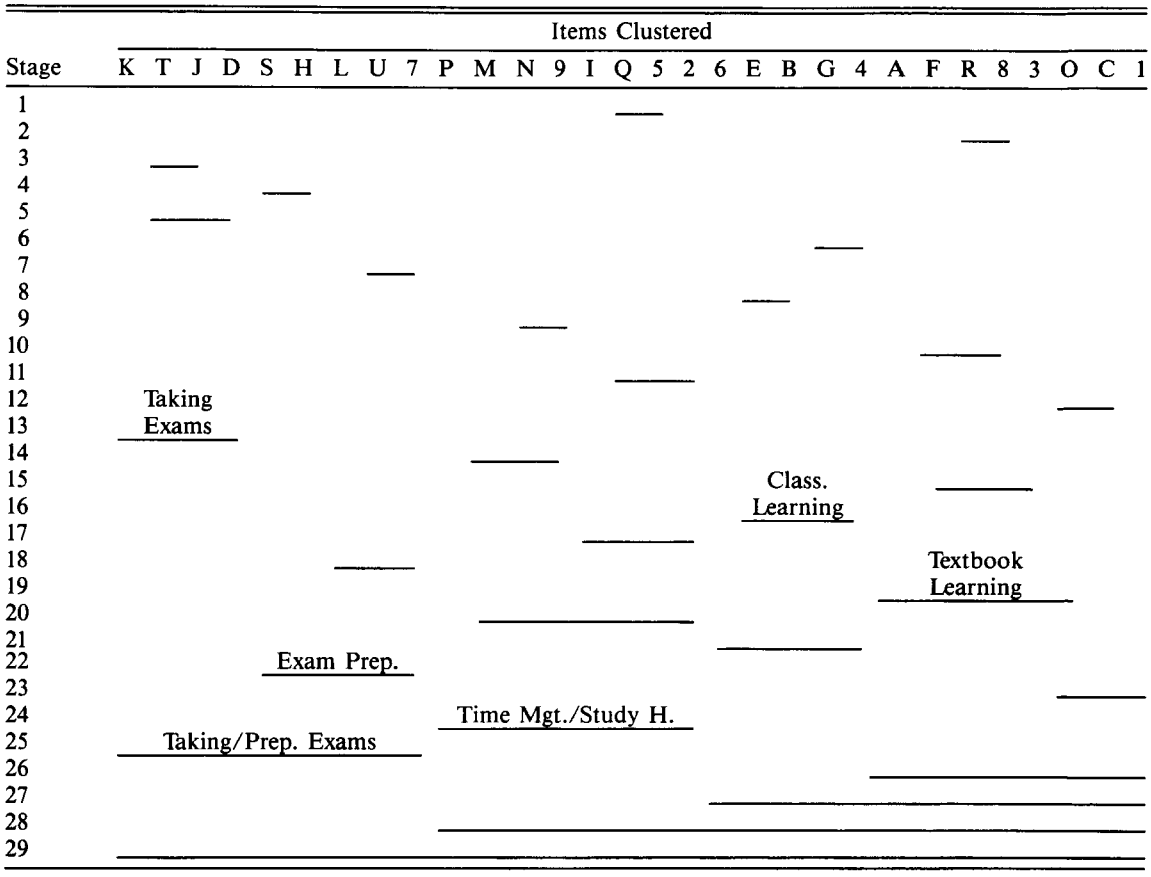
Table 4
Clustering Solution From Three-Dimensional INDSCAL Coordinates

Stage	Items Clustered																										
	S	H	L	K	J	D	U	T	7	E	O	B	C	6	G	4	P	N	M	I	9	Q	5	2	F	R	A	8	3	1
1																													
2																													
3																													
4																													
5																													
6																													
7																													
8																													
9																													
10																													
11																													
12																													
13																													
14																													
15																													
16																													
17																													
18																													
19																													
20																													
21			Taking Exams																										
22																													
23																									Textbk. Lrn.				
24											Class. Learning																		
25																	Time M./Study H.												
26		Taking/Prep. Exams																											
27																													
28																													
29																													

was "Preparing for Exams," but clustered together with the "Classroom Learning" items. Thus, with the exception of one item, of the four substantive clusters that emerged in the three-dimensional cluster analysis, two directly corresponded to content areas prescribed in the test blueprint, and two represented combinations of two highly-related content areas. Figure 5 presents a revised grouping of the items where the content areas of "Time Management" and "Study Habits" are combined (plotted as diamonds), and Item 6 is reclassified as belonging to the "Classroom Learning" content area (plotted as flags).

The cluster analysis resulting from the four-dimensional INDSCAL coordinates revealed several item clusters that were also congruous with the test blueprint (see Table 5). Two of the clusters observed in the three-dimensional analysis also emerged: The cluster consisting of "Textbook Learning" items and the cluster consisting of the "Time Management" and "Study Habits" items. The content areas of "Taking Exams" and "Preparing for Exams" clustered as specified in the blueprint with two exceptions: Items 6 and L. Item L was specified as a "Taking Exams" item, but it clustered with the "Preparing for Exams" cluster. Item 6 again clustered with items belonging to the "Classroom Learning" content domain. The four-dimensional cluster analysis did not fully support the retention of a "Classroom Learning" cluster.

Table 5
Clustering Solution From Four-Dimensional INDSCAL Coordinates



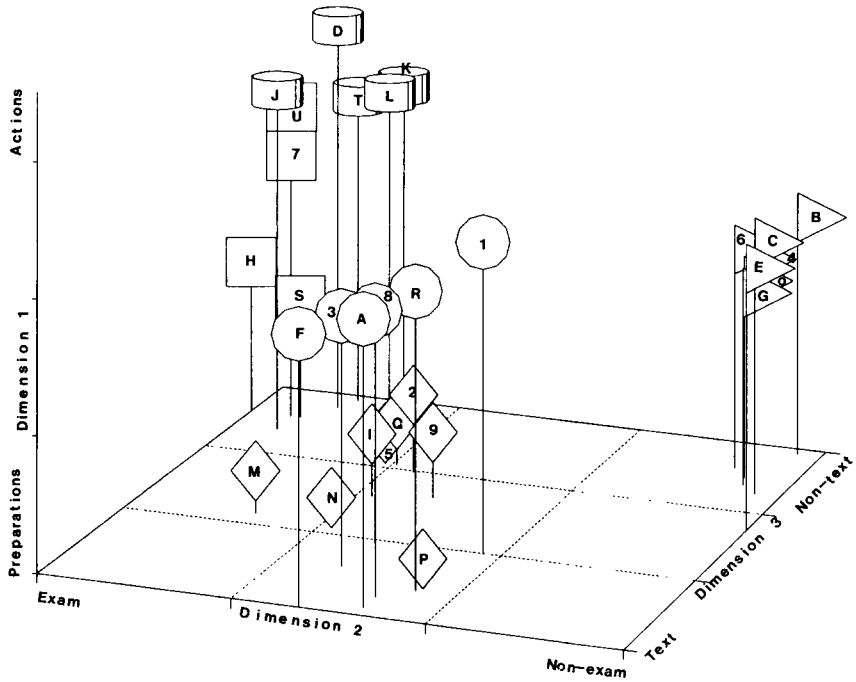
Four of the items in this domain did cluster together (Items 4, B, E, and F); however, two of the items (C and O) clustered together with the "Textbook Learning" cluster. Thus, the cluster analysis stemming from the four-dimensional INDSCAL solution made finer distinctions between items corresponding to the same content area (e.g., "Classroom Learning" items) and items corresponding to different content areas ("Taking Exams" and "Preparing for Exams" items).

Discussion

The results indicate that the SMEs relied primarily on content characteristics of test items when making their similarity judgments. This finding is interesting considering that the SMEs were not informed of the content areas of the test and that they were not instructed to rate the items according to content characteristics. The task presented to the SMEs in this study might be effective in circumventing an expectancy bias that may occur when SMEs are informed of the content areas of the test blueprint or of item-content area specifications.

The results also illustrated the utility of MDS and cluster analysis to identify groups of items perceived to be similar by the SMEs. The MDS configurations allow for visual inspection of item groupings, and the cluster analyses facilitated this visual inspection. It should be noted that, in this study, the purpose of the hierarchical cluster

Figure 5
Stimulus Configuration From the Three-Dimensional Solution Illustrating "Revised" Content Area Specifications: Time Management/Study Habits (Diamonds), Classroom Learning (Flags), Textbook Learning (Balloons), Preparing for Exams (Squares), and Taking Exams (Cylinders)



analysis was to facilitate visual inspection of the MDS configurations, not to provide an alternative inspection of the data. Because the clustering was performed on the MDS item coordinates, it is not surprising that the items within the resulting clusters were proximal in the MDS space. Had the cluster analyses been performed on the original similarity data (using the INDCLUS program designed for three-way data; see Carroll & Arabie, 1983) discrepancies between the MDS and clustering solutions could have been investigated.

It should also be noted that because the clustering was performed on the unweighted item coordinates, an assumption was made that the dimensions were equally important to the SMEs. Future analyses should consider weighting the dimensions according to their proportion of variance accounted for, or performing separate cluster analyses for each SME.

Inspection of a series of MDS and cluster analyses allows an investigator to determine the level of scrutiny to be imposed on the items. If the investigator wishes to have only highly homogeneous content areas, items that cluster as prescribed in the lower-dimensional space but do not cluster as prescribed in the higher-dimensional space may be rejected. An investigator can decide whether to flag these items for inspection (as the four-dimensional solution would suggest), consider them homogeneous (as the three-dimensional solution would suggest), or integrate both positions within a hierarchical model. Thus, the eclectic information obtained through joint MDS and cluster analyses illustrates Napior's (1972) contention that joint MDS/cluster analyses can detect the "more subtle and complex relations" among stimuli that would not be detected by cluster analysis alone (p. 165).

Benefits of the Procedure for Test Construction

In addition to portraying global content structure, the method presented here may also prove useful for item selection purposes. Items that emerge as outliers or do not cluster with other items in their prescribed content area can be flagged for removal or modification. The proposed method may also be useful in item sensitivity review. Judges who are members of concerned (e.g., minority) groups could be employed to judge the similarity of the items. The stimulus configurations of test items from these judges could be compared to the stimulus 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 further sensitivity review. INDSCAL analyses of the two or more groups of judges may also identify sensitive items and/or content areas.

The present method also may 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 content areas, rather than truly overlapping content domains. In this way, the MDS and cluster analyses can provide information regarding the number of items necessary to adequately represent a content domain.

Limitations of the Procedure

Although the proposed procedure shows potential as a test construction tool, there are problems and limitations. One problem is that rating the similarity of test items becomes increasingly complex as test size increases. A 30-item test requires 435 item comparisons. The item similarity matrix will increase 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 stimulus comparisons (Spence, 1982, 1983), by dividing the item comparisons among groups of judges, or by increasing the time interval required for the judges to make their comparisons.

Another limitation of the current study was that only one sample of three SMEs was employed. Future research should employ larger groups of SMEs to determine if the dimensions and clusters are consistent across samples comprising different numbers of judges, and to cross-validate the results obtained in the different samples. Osterlind (1989) recommended that a minimum of four or five judges is necessary to evaluate a test of moderate size.

The quality of the SMEs employed is of crucial importance. For this method, or any method of evaluating test content, to be successful, it is imperative that the judges be knowledgeable of the specified content domain and that they are representative of the domain of all possible qualified judges.

A major limitation of the present study was that data on item-domain relevance were not gathered. Thus, although the present study assessed the content structure of the blueprint, it did not directly assess the relevance of the items to the content areas defined by the test blueprint. This limitation could be remedied by gathering both item similarity data and item relevance data as described below.

Implications for Future Research

Although the results of the present study proved valuable in understanding the content structure of the test, it could be supplemented by other methods to provide additional information pertaining to the test's content representation. For example, the results from the present method could be compared with results from item analyses employing test response data. It would be interesting to identify the item-to-total score correlations or item-to-content area score correlations for those items that do not cluster as predicted. If these correlations are relatively small, it may support the removal of those items from the test. Napior's (1972) method of multidimensional item analysis would provide results

that could be directly comparable to the data collected in the present study. Davison (1985) describes advantages of using MDS to analyze test item intercorrelations. Such analyses are likely to supplement analyses of item similarity ratings.

The data gathered in the present study could also be subjected to a confirmatory MDS analysis in which the distances between the items are constrained according to their blueprint specifications. Borg and Lingoes (1980) describe a procedure in which a "pseudo-matrix" is constructed that represents the hypothesized item relations specified in the test blueprint. The distances resulting from the data could be constrained by the distances resulting from a MDS of the pseudo-data matrix. The degree to which the data fit the pseudo-data would then be taken as an index of the test's fit to its blueprint. This procedure would also allow for items to be deleted until a satisfactory fit of test to blueprint was obtained. Heiser and Meulman (1983) describe other methods of constrained MDS that could also be used to impose blueprint constraints on the item configurations.

Future applications of this procedure should also gather item-domain ratings from the SMEs to evaluate the relevance of the items to their perceived content areas. This could be accomplished by having the SMEs provide the similarity ratings, and then asking them to rate how strongly each item corresponded to each of the content areas specified in the test blueprint. These "item-relevance" ratings could be used in a multiple regression procedure where the relevance data were regressed on the dimensions. If the items were relevant to their specified content areas and the MDS solution supported the test blueprint, then the content areas would help in interpretation of the dimensions.

References

Arabie, P., Carroll, J. D., & DeSarbo, W. J. (1987). *Three-way scaling and clustering.* Newbury Park CA: Sage.

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

Carroll, J. D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 48,* 157–169.

Carroll, J. D., & 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.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* New York: Harcourt, Brace, and World.

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

Davison, M. L. (1983). *Multidimensional scaling.* New York: Wiley.

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.

Green, S. B. (1983). Identifiability of spurious factors with linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139–147.

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* (pp. 80–123). 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* (pp. 199–230). Baltimore: Johns Hopkins University Press.

Heiser, W. J., & Meulman, J. (1983). Constrained multidimensional scaling, including confirmation. *Applied Psychological Measurement, 7,* 381–404.

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

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115–129.

Kruskal, J. B., & 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. (1958). The perception of social attitudes. *Journal of Abnormal and Social Psychology, 52,*

57-66.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955-966.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.
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 ed.), (pp. 13-103). Washington DC: American Council on Education.
Morris, L. L., & Fitz-Gibbon, C. T. (1978). *How to measure achievement.* Beverly Hills: Sage.
Napior, D. (1972) Nonmetric multidimensional techniques for summated ratings. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences. Volume 1: Theory* (pp. 157-178). New York: Seminar Press.
Oltman, P. K., Stricker, L. J., & 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. (1980). Some small sample results for maximum likelihood estimation in multidimensional scaling. *Psychometrika, 45,* 139-144.
Ramsay, J. O. (1981). How to use MULTISCALE. In S. S. Schiffman, M. L. Reynolds, & F. W. Young (Eds.), *Multidimensional scaling: Theory, methods, and applications* (pp. 211-235). New York: Academic Press.
Ramsay, J. O. (1986). *MULTISCALE-II manual* [Computer program 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.
Sokal, R., & 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* (pp. 29-46). Minneapolis: University of Minnesota Press.
Spence, I. (1983). Monte carlo simulation studies. *Applied Psychological Measurement, 7,* 405-426.
Takane, Y., Young, F. W., & 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.
Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.
Tucker, L. T. (1961). *Factor analysis of relevance judgements: An approach to content validity.* Paper presented at the Invitational Conference on Testing Problems, Princeton NJ. [Reprinted in A. Anastasi (Ed.), *Testing problems in perspective,* (1966), (pp. 577-586) Washington D.C.: American Council on Education.]
Wainer, H., Hurt, S., & Aiken, L. (1976). Rorschach revisited: A new look at an old test. *Journal of Consulting and Clinical Psychology, 44,* 390-399.
Wainer, H., & Kaye, K. (1978). Multidimensional scaling of concept learning in an introductory course. *Journal of Educational Psychology, 66,* 591-598.
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.

Acknowledgments

The authors are indebted to the helpful comments and suggestions provided by two anonymous reviewers of an earlier version of this paper. The authors thank the following for their comments on the work as it progressed: Bruce Biskin, Herman Friedman, Barbara Helms, Janet F. Carlson, Soonmook Lee, Samuel Messick, Thanos Patelis, and Howard Wainer. Gratitude is also extended to James Ramsay for his helpful comments and for providing the MULTISCALE-II program. An earlier version of this paper was presented at the 1990 Annual Meeting of the Northeastern Educational Research Association, Ellenville NY.

Author's Address

Send requests for reprints or further information to Kurt F. Geisinger, 601 Culkin Hall, SUNY College, Oswego NY 13126, U.S.A.