Using Subject-Matter Experts to Assess Content Representation: An MDS Analysis

Stephen G. Sireci, University of Massachusetts at Amherst
Kurt F. Geisinger, State University of New York, Oswego

Demonstration of content domain representation is of central importance in test validation. An expanded version of the method of content evaluation proposed by Sireci & Geisinger (1992) was evaluated with respect to a national licensure examination and a nationally standardized social studies achievement test. Two groups of 15 subject-matter experts (SMEs) rated the similarity of all item pairs comprising a test, and then rated the relevance of the items to the content domains listed in the test blueprints. The similarity ratings were analyzed using multidimensional scaling (MDS); the item relevance ratings were analyzed using procedures proposed by Hambleton (1984) and Aiken (1980). The SMEs' perceptions of the underlying content structures of the tests emerged in the MDS solutions. All dimensions were germane to the content domains measured by the tests. Some of these dimensions were consistent with the content structure specified in the test blueprint, others were not. Correlation and regression analyses of the MDS item coordinates and item relevance ratings indicated that using both item similarity and item relevance data provided greater information of content representation than did using either approach alone. The implications of the procedure for test validity are discussed and suggestions for future research are provided. *Index terms: construct validity, content validity, cluster analysis, multidimensional scaling, subject-matter experts, test construction.*

Adequate representation of the content domain measured is a fundamental requirement of educational and psychological testing. To evaluate content domain representation, subject-matter experts (SMEs) are typically required to rate the relevance of test items to one or more of the content domains comprising a test blueprint (Crocker, Miller, & Franks, 1989; Popham, 1992). However, Sireci & Geisinger (1992) proposed a method based on gathering SME ratings of the similarity of all test items to one another [a multidimensional scaling (MDS) approach]. This study compared and evaluated these two approaches for assessing content domain representation.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 3, September 1995, pp. 241–255

0146-6216/95/030241-15$2.00

Method

Instruments

The two tests evaluated in this study were the Auditing section from the May 1990 Uniform CPA Examination (American Institute of Certified Public Accountants, 1990) and a benchmark form of the CTBS/4 Social Studies test (level 17/18) designed to measure achievement in social studies at the junior high school level (CTB McGraw-Hill, 1989). The Auditing section contained 60 multiple-choice items and four essays. Due to the difficulty in comparing essay with multiple-choice questions, the essay questions were not used in this study. Also, a 40-item subset of the 60 multiple-choice items was selected to reduce the burden on the SMEs and to ease interpretation of the MDS stimulus configurations.

The 40 Auditing items were selected by (1) eliminating nine items that were administered previously and were selected for repeated use because of their desirable content and statistical characteristics; (2) eliminating five other items that involved different item formats than the other items; and (3) removing six other items randomly so that the percentage of items representing each content area in the original 60-item test was maintained in the 40-item subset. The blueprint of the Auditing ex-

amination comprised four content areas: Professional Responsibilities, Internal Control, Evidence and Procedures, and Reporting.

The Social Studies test consisted of 40 multiple-choice items; all 40 items were included in the analysis. The blueprint of the Social Studies examination comprised seven content areas: Geography, Economics, History, Political Science, Interrelated Disciplines, Sociology/Anthropology, and Applied Social Studies.

Raters

30 SMEs provided the data for this study. 15 SMEs evaluated the Auditing examination; the other 15 evaluated the Social Studies examination. The auditing SMEs were required to be licensed certified public accountants with at least three years experience performing audits. The social studies SMEs were required to be state certified to teach social studies and to have at least three years experience teaching social studies at the junior high school level.

Procedure

Item similarity ratings. The auditing and social studies SMEs completed the same tasks. Each SME was given a booklet containing all possible pairs of the 40 test items (780 item pairings). Beneath each item pair was a 10-point Likert scale. The anchor point 1 was labeled *highly similar*, and the anchor point 10 was labeled *highly dissimilar*. The SMEs were instructed to read each item pair and make a judgment regarding the similarity of the two items in terms of the knowledge (auditing or social studies) being measured. They were instructed to circle their rating on the 10-point scale printed below the item pair. The SMEs were not provided with any further criteria on which to make their ratings. This ambiguity in instructions was used to avoid biasing their ratings in favor of the test blueprint.

Relevance ratings. Following completion of the item similarity ratings, the SMEs rated the relevance of each test item to each content area of the test blueprint. The SMEs were given a content area description sheet that described these content areas. The relevance ratings were made along a 10-point scale where 1 was *not at all relevant* and 10 was *highly relevant.*

Data Analysis

MDS analysis. The item similarity data for each group of SMEs were analyzed separately. The data for each group were analyzed using the INDSCAL (Carroll & Chang, 1970) model of the ALSCAL program of SPSSX (Young, Takane, & Lewyckyj, 1978). The INDSCAL MDS model is a generalization of the classical MDS model developed by Torgerson (1958) and 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 are computed by incorporating this weighting factor into the Euclidean distance formula used by classical MDS. The INDSCAL model defines the distance between two objects i and j as:

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

where

d_{ijk} is the Euclidean distance between points i and j as perceived by rater k,

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

r is the maximum dimensionality requested.

The results from an INDSCAL analysis include a multidimensional configuration of the attributes rated (the stimulus space), and a multidimensional configuration of the raters (the rater space). Two-dimensional through six-dimensional INDSCAL solutions were obtained for each group. These analyses were performed to discover the structure of the item similarity data and to investigate differences among the SMEs.

The fit indexes of STRESS and RSQ were used to evaluate the fit of the INDSCAL models to the data. The STRESS index measures the departure of the data from the model; thus, low values of STRESS are desired. The equation for STRESS is

$$\text{STRESS} = \left[\frac{\sum_{ij} \left(d_{ij}^* - d_{ij} \right)^2}{\sum_{ij} d_{ij}^2} \right]^{1/2}, \quad (2)$$

where the numerator represents the sum of the squared differences between the disparities (d_{ij}^*) and the fitted distances (d_{ij}), and the denominator represents the sum of the squared distances.

RSQ is more straightforward. RSQ is the squared multiple correlation between the disparities and the distances. It is interpreted as the proportion of variance in the disparity data accounted for by the distances; thus, the higher the RSQ value, the better the fit. Kruskal & Wish (1978) reported that, for classical (one matrix) MDS, STRESS values below .10 and RSQ values above .90 indicate well-fitting models. However, these general rules have been criticized when applied to the INDSCAL model (Arabie, Carroll, & DeSarbo, 1987).

Regression analysis. To assist in interpretation of the MDS item configurations, and to compare the configurations with the content structure specified in the test blueprints, two sets of correlation and regression analyses were conducted. The first set of analyses used "dummy variable" coding to indicate the content area classifications (blueprint classifications) of every item. For each content area, items linked to that area were coded 1; all other items were coded 0. The MDS coordinates were regressed onto these data to determine the amount of variance in the blueprint classifications accounted for by the item coordinates. The second set of analyses used the item relevance ratings averaged over the 15 SMEs in each group. The MDS coordinates were regressed onto the averaged relevance data for each content area to determine the amount of variance in the relevance ratings for each content area accounted for by the item coordinates.

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

Item-objective congruence analyses. The item relevance data were analyzed independently to compare the results of the MDS procedure with those based solely on item relevance ratings. As suggested by Hambleton (1984), the relevance data were averaged over the 15 SMEs and the mean relevance rating for each item on each content area was computed. An item was considered matched to its blueprint content area if the highest mean relevance rating for the item corresponded to its blueprint classification.

In addition, Aiken's (1980) validity index was computed for each item. This index accounts for the number of categories used to rate each item and the number of judges that responded to each category. The equation for Aiken's validity index, V, is

$$V = \frac{\sum_{i=1}^{c-1} i n_i}{N(c-1)}, \quad (3)$$

where

- c is the number of categories on the item relevance rating scale,
- i is the weight given to each category,
- n_i is the number of judges who rated the item in the ith category, and
- N is the total number of SMEs.

The lowest category is given a weight (or i-value) of 0, the next category is given a weight of 1, and so forth, and the highest category is given a weight of $c - 1$. Aiken provided a formula for evaluating the significance of the validity index when a large number of SMEs is used. This formula provides a normal deviate (z) for the index and the probability of obtaining the z is obtained from a standard normal z table. The formula for deriving the normal deviate from the Aiken index is

$$z = \frac{N(c-1)(2V-1)-1}{\left[\frac{N(c-1)(c+1)}{3} \right]^{1/2}}. \quad (4)$$

The statistical significance of the Aiken index provides a practical measure of SME congruence. When the Aiken index is large and statistically significant, there is agreement among the SMEs that the item is relevant to the specific content area. When the Aiken index is small and statistically significant, there is agreement among the SMEs that the item is not highly relevant to the specific content area. Moderate values of the Aiken index signify poor agreement among the SMEs about the relevance of the item to its prescribed content area.

Results

For both the Auditing and Social Studies examinations, an evaluation of the congruence of the SMEs was conducted by visual inspection of the rater space and by comparing the fit values obtained for each SME. Although the fit values indicated substantial differences in fit among the SMEs, these differences did not affect overall interpretation of the stimulus spaces.

Interpretation of the MDS Stimulus Configurations

Six-dimensional INDSCAL solutions were selected as the appropriate MDS model for both the auditing and social studies similarity data. The six-dimensional models were selected primarily on the basis of interpretability (i.e., all six dimensions were readily interpretable) and fit to the data. The fit values of RSQ and STRESS, averaged over the auditing and social studies SMEs, indicated that a moderate amount of variation was present in these data that was not accounted for by the model (the STRESS values were .14 and .13, and the RSQ values were .63 and .70 for the auditing and social studies data, respectively). However, these values are not surprising given the relatively large number of stimuli rated (MacCallum, 1981).

Auditing Examination

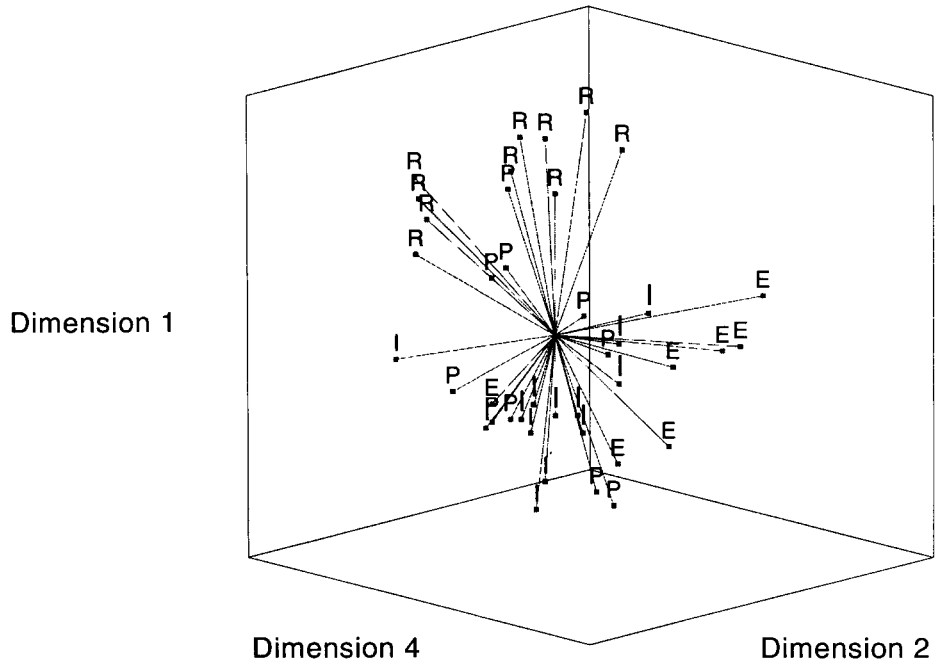
Visual interpretation of the MDS configurations. The six-dimensional auditing stimulus configuration was inspected visually by the first author and a senior auditor from the American Institute of Certified Public Accountants. All six dimensions were interpretable and appeared to separate the items according to characteristics relevant to the auditing content domain.

Dimensions 1 and 4 separated items comprising the different content areas. Dimension 1 distinguished Reporting (R) items and Internal Control (I) items; Dimension 4 distinguished Internal Control and Evidence and Procedures (E) items. No dimension clearly distinguished the Professional Responsibilities (P) items from the other content areas, but Dimension 2 did account for knowledge and application of professional standards and did pull most of these items in the same direction. Because these three dimensions were most relevant to the content structure of the test, they are plotted together in Figure 1.

Dimensions 3, 5, and 6 reflected other content characteristics of the items that were extraneous to the content areas listed in the test blueprint. Dimension 3 distinguished items measuring commonly-performed auditing procedures from those measuring extraordinary auditing procedures, Dimension 5 distinguished items related to the execution of the audit from those involved with planning or concluding the audit, and Dimension 6 distinguished items that measured higher-level auditing procedures from the more elementary auditing procedures. Dimensions 1 and 4 polarized specific content areas designated in the test blueprint; the other four dimensions separated the items according to other aspects of auditing not specified in the test blueprint.

The labels ascribed to the auditing MDS dimensions using visual interpretation were: reporting versus internal control (Dimension 1), application of auditing standards versus knowledge of auditing standards (Dimension 2), common versus extraordinary auditing procedures (Dimension 3), internal control versus evidence and procedures (Dimension 4), supplementary work versus field work (Dimension 5), and senior-level versus entry-level procedures (Dimension 6). Only Dimensions 1, 2, and 4 polarized content areas designated in the test blueprint. These three dimensions are plotted together in Figure 1. The separation of item groupings comprising the content areas I, E, and R is evident in Figure 1. These content area groupings are seen more clearly

Figure 1
Three-Dimensional Subspace of the Auditing Item Configuration (P = Professional Responsibilities, I = Internal Control, E = Evidence and Procedures, and R = Reporting)



in the plot resulting from cluster-analyzing these MDS coordinates (Figure 2). Items comprising the P content area did not form a distinct cluster; these items were grouped with items comprising other content areas.

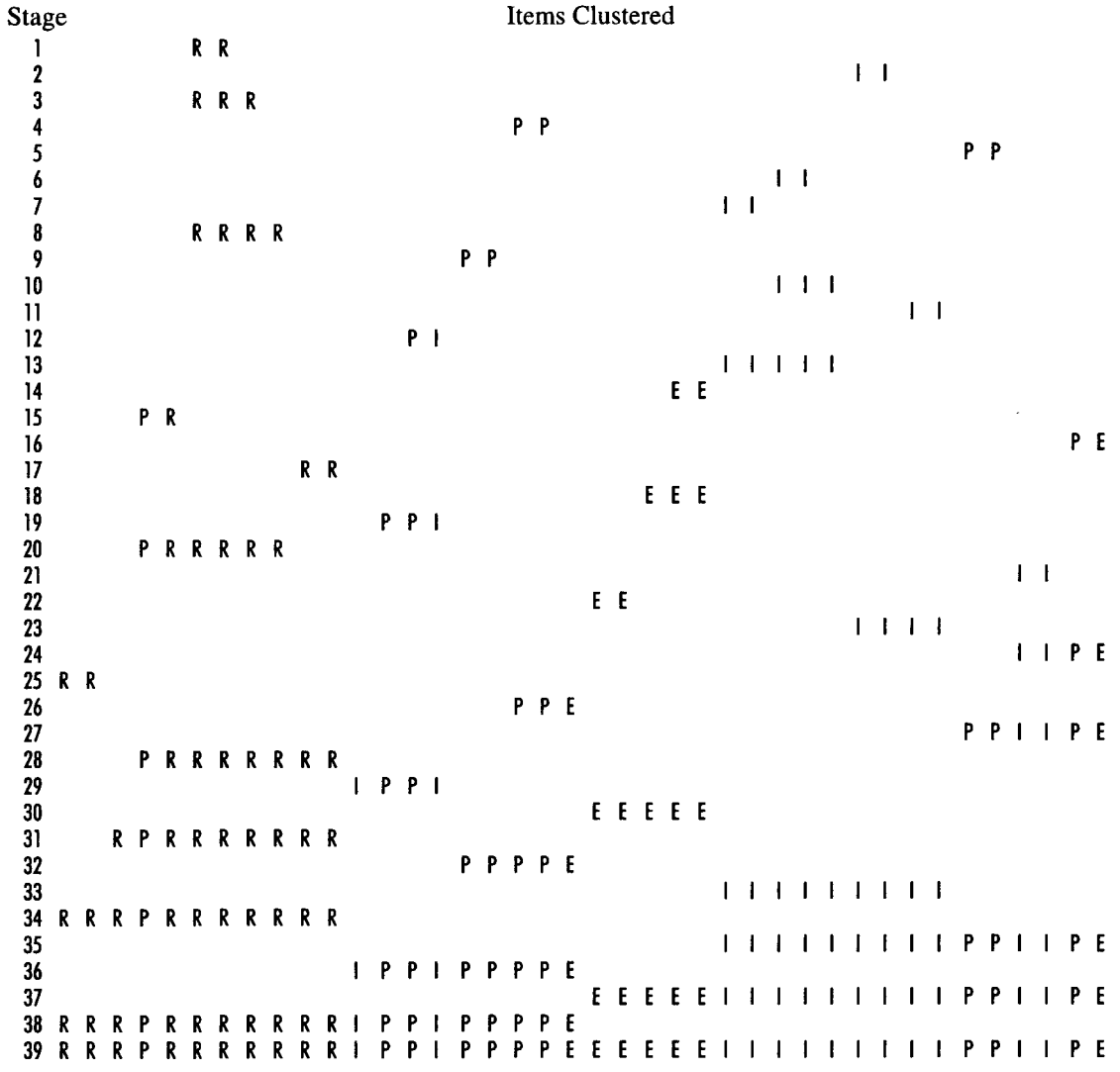
Comparison of the auditing configuration, blueprint, and relevance ratings. The results of the correlation and regression analyses for the auditing data are presented in Table 1. The correlations reported in Table 1 are Pearson correlations between the item coordinates and the blueprint classifications or relevance ratings. The R^2 values reflect the percentage of variance accounted for by regressing the item coordinates for all six dimensions across the blueprint or relevance data.

As indicated in Table 1, several of the correlations between the coordinates for a given dimension and the blueprint classifications for a given content area were statistically significant, indicating that the dimensions were relevant to the content areas. Regressing the item coordinates on all six dimensions across each content area accounted for a minimum of 43% (Professional Responsibilities) and a maximum of 80% (Internal Control and Reporting) of the variance of the item blueprint classifications.

Results of the correlation and regression analyses using the MDS stimulus coordinates and item relevance ratings, also shown in Table 1, showed several statistically significant correlations indicating a strong relationship between the MDS coordinates and the relevance ratings. Using the coordinates to predict content area relevance accounted for between 67% (Evidence and Procedures) and 93% (Reporting) of the relevance rating variance for the four blueprint content areas.

Comparison of the auditing relevance ratings and test blueprint. The auditing relevance data, averaged over the 15 SMEs, indicated that the SMEs

Figure 2
Sequential Item Groupings From the Auditing Cluster Analysis (P = Professional Responsibilities, I = Internal Control, E = Evidence and Procedures, and R = Reporting)



matched 33 out of the 40 items to their blueprint content area. The largest mean relevance ratings for the other seven items were not congruent with their blueprint classifications; however, in all seven cases the second largest mean relevance rating corresponded to the prescribed content area. These results are summarized in Table 2 where "Number Rated First" indicates the number of items within the content area that had higher mean relevance ratings for that area than for any other. The correlations between the auditing relevance data and the item blueprint classifications were high, ranging from .66 (Evidence and Procedures) to .92 (Reporting).

The Aiken (1980) validity indexes (see Table 3) identified seven items rated relatively lower to their blueprint content areas than were the other items

Table 1
Correlation and Regression Results for the Auditing Examination

Type of Data and Content Area	Number of Items	Dimension 1	2	3	4	5	6	R^2
Blueprint Classification Data								
Professional Responsibilities	10	−.04	−.42*	−.36*	−.31*	−.13	.10	.43
Internal Control	13	−.59*	.42*	.05	.57*	−.19	.11	.80
Evidence and Procedures	7	−.16	.30	−.03	−.56*	.18	−.21	.48
Reporting	10	.82*	−.30	.32*	.18	.18	−.03	.80
Relevance Rating Data								
Professional Responsibilities	10	.16	−.48*	−.31*	−.57*	−.15	.23	.73
Internal Control	13	−.71*	.51*	0.00	.41*	−.33*	.19	.91
Evidence and Procedures	7	−.50*	.36*	.21	−.42*	.16	−.39*	.67
Reporting	10	.87*	−.39*	.34*	.24	.13	.06	.93

*$p \leq .05$.

(i.e., indexes below .70). Six of these items (3, 13, 19, 24, 35, and 36) were the same items identified through analysis of the averaged data. Most of the Aiken indexes were statistically significant, suggesting that their values were not due to chance.

Social Studies Examination

Visual interpretation of the MDS configurations. Visual interpretation of the social studies stimulus configuration was conducted by the investigators and an independent social studies SME. Similar to the auditing configuration, all six dimensions were interpretable. Five dimensions distinguished the items according to their content characteristics, and one distinguished items according to the cognitive levels measured.

The labels ascribed to the social studies dimensions using visual interpretation were: geography versus other (Dimension 1), other versus economics (Dimension 2), lower-order versus higher-order thinking skills (Dimension 3), other versus history (Dimension 4), cultural versus other (Dimension 5), and international versus national (Dimension 6). Substantial overlap between items comprising the seven content areas was observed in the MDS stimulus configurations. The two-dimensional scatterplot that best separated blueprint content areas is presented in Figure 3. Figure 3 illustrates a moderate degree of separation of items comprising the G, E, and P content areas, and substantial overlap between items comprising the other content areas. A three-dimensional configuration is not presented because content area groupings were not as evident in the three-dimensional subspaces. P and E item groupings also emerged in the cluster analysis of the coordinates (Figure 4), but the substantial overlap between items comprising the other content areas was conspicuous.

Comparison of the social studies configuration, blueprint, and relevance ratings. Several statistically significant correlations between the coordinates on a given dimension and the blueprint classifications and relevance ratings for a particular con-

Table 2
Averaged Auditing Content Area Relevance Ratings and the Pearson Correlation Between the Relevance Ratings and the Blueprint Classifications (r)

Content Area	Number of Items	Number Rated First	Percent First	r
Professional Responsibilities	10	7	70%	.74
Internal Control	13	11	85%	.91
Evidence and Procedures	7	6	86%	.66
Reporting	10	9	90%	.92

Table 3
Aiken Validity Indexes for the Auditing Items

Content Area and Median Value	Item	Value
Professional Responsibilities		
	31	.96*
	32	.95*
	33	.81*
	34	.87*
	35	.64*
	36	.60
	37	.81*
	38	.96*
	39	.72*
	40	.93*
	Median	.84
Internal Control		
	18	.88*
	19	.59
	20	.93*
	21	.93*
	22	.96*
	23	.83*
	24	.58
	25	.88*
	26	.88*
	27	.79*
	28	.81*
	29	.69*
	30	.85*
	Median	.85
Evidence and Procedures		
	1	.79*
	2	.74*
	3	.64*
	4	.73*
	5	.93*
	6	.93*
	7	.93*
	Median	.79
Reporting		
	8	.98*
	9	.99*
	10	.97*
	11	.97*
	12	.99*
	13	.69*
	14	.94*
	15	.98*
	16	.90*
	17	.94
	Median	.97

*$p \leq .05$.

tent area were observed (see Table 4). The regression of the MDS coordinates onto the blueprint classifications accounted for between 5% (Interrelated Disciplines) and 65% (Economics) of the variance of the social studies blueprint classifications. Regressing the coordinates across the averaged content area relevance ratings accounted for between 60% (Applied Social Studies, History) and 93% (Geography) of the variance of the relevance ratings for each content area (Table 4).

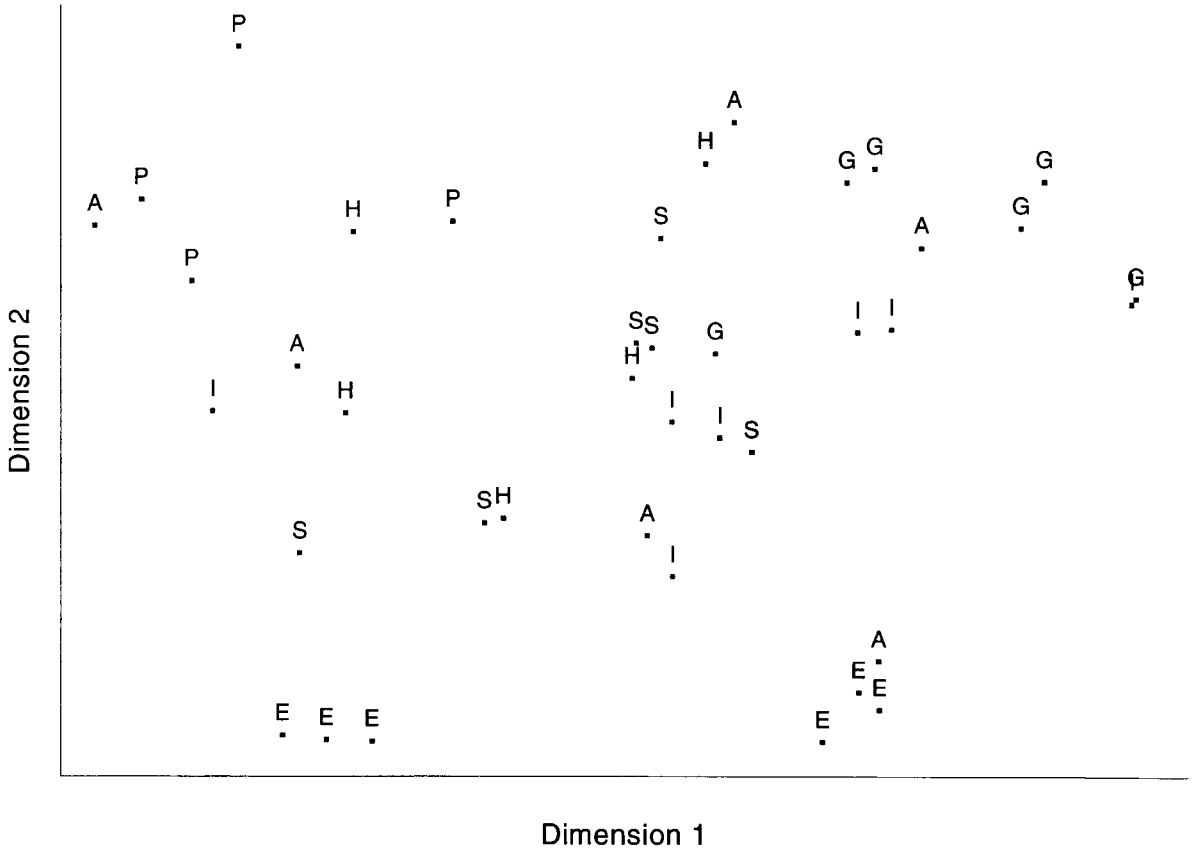
Comparison of the social studies relevance ratings and blueprint. Analysis of the relevance data averaged over the 15 social studies SMEs indicated that the items comprising the Geography, Economics, History, Political Science, and Interrelated Disciplines content areas were rated highly congruent to their blueprint specifications (Table 5). At least 80% of the items comprising each of these content areas were matched to their blueprint specifications. The items comprising Sociology/Anthropology and Applied Social Studies were predominantly rated as more relevant to a different content area. Table 5 also shows that the correlations between the relevance ratings and blueprint classifications varied widely across the content areas from .02 (Applied Social Studies) to .77 (Political Science). The Aiken indexes (Table 6) yielded similar results. The items comprising the Geography, Economics, History, and Political Science content areas exhibited higher values than those comprising the other areas.

Discussion

The results demonstrate that the content structure of a test can be evaluated thoroughly by analyzing item similarity and item relevance data provided by SMEs. In comparing the "item similarity" approach with the "item relevance" approach, it is evident that the former approach represents a more comprehensive examination of content domain representation.

The item similarity approach provides both convergent and divergent information regarding the adequacy of both the test blueprint and the operational definition of the content domain. Inspection of the MDS stimulus configurations uncovered content-relevant relationships among the test items that

Figure 3
Two-Dimensional Subspace of the Social Studies Item Configuration (G = Geography, E = Economics, H = History, P = Political Science, I = Interrelated Disciplines, S = Sociology/Anthropology, and A = Applied Social Studies)



were not specified in the blueprint. For example, Dimension 6 of the social studies solution (international versus national) revealed a distinction between U.S. history and world history items. The distinction between items measuring supplementary and field auditing procedures (Dimension 5 of the auditing solution) provides another example of convergent evidence.

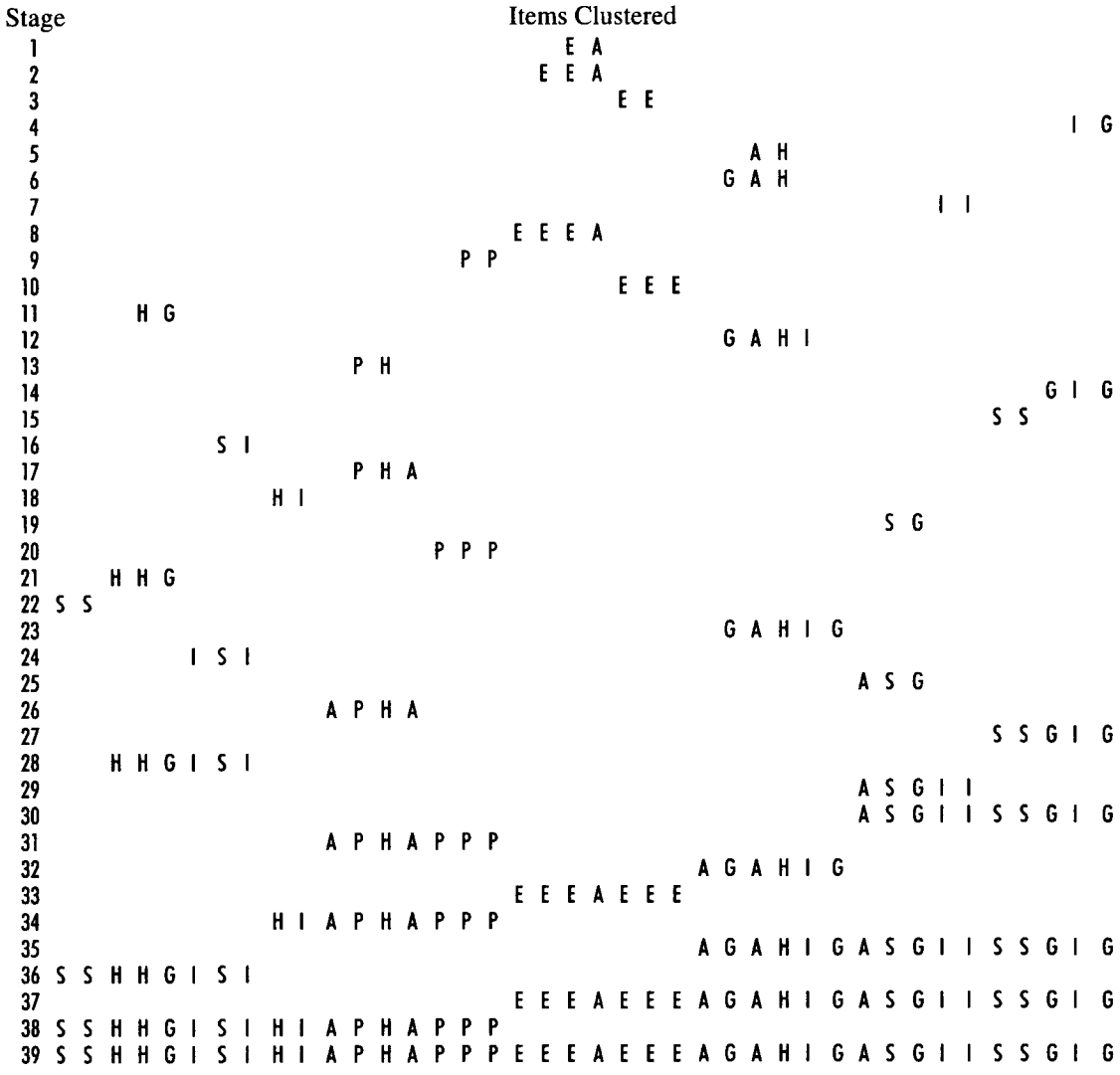
Comparison of the Two Approaches

An evaluation of the similarities and differences between the MDS (item similarity) and item relevance approaches can be made by comparing the results based on data gathered when the SMEs were unaware of the content areas comprising the test blueprint (i.e., similarity ratings) with those gathered when they were aware of these content areas (i.e., relevance ratings). For the Auditing examination, all four content areas were supported by analysis of the item relevance data: high correlations were observed between the relevance ratings and blueprint classifications, the averaged relevance ratings matched almost all of the items to their prescribed content areas, and the Aiken validity indexes were, for the most part, large and statistically significant. However, MDS analysis of the item similarity ratings accounted for less than 50% of the variance in the blueprint classifications for Professional Responsibilities and Evidence and Procedures, and the items comprising the Professional Responsibilities content area did not cluster together in the MDS space.

For the social studies data, analyses based on

Figure 4

Sequential Item Groupings From the Social Studies Cluster Analysis (G = Geography, E = Economics, H = History, P = Political Science, I = Interrelated Disciplines, S = Sociology/Anthropology, and A = Applied Social Studies)



the item relevance data illustrated poor convergence between the test developers and the SMEs for only two content areas: Applied Social Studies and Sociology/Anthropology. These two content areas exhibited relatively small correlations between the blueprint classifications and relevance ratings, few matches to the blueprint content areas using the averaged relevance ratings, and predominantly small Aiken validity indexes. Analysis of the social studies item similarity data revealed poor convergence between the test developers and SMEs for five of the seven content areas. The Applied Social Studies, Sociology/Anthropology, Geography, History, and Interrelated Disciplines content areas exhibited substantial overlap in the MDS space, and the MDS coordinates accounted for less than 50% of the variance in the blueprint classifications.

The differences noted between the item similar-

Table 4
Correlation and Regression Results for the Social Studies Examination

Type of Data and Content Area	Number of Items	Dimension 1	2	3	4	5	6	R^2
Blueprint Classification Data								
Geography	6	.47*	.34*	.05	.03	−.16	.18	.36
Economics	6	−.05	−.73*	.04	.34*	−.13	−.21	.65
History	5	−.16	.10	.45*	−.24	.13	.11	.28
Political Science	4	−.44*	.35*	.08	.25	.45*	−.25	.61
Sociology/Anthropology	5	−.06	−.04	−.47*	−.19	.13	.30	.34
Applied Social Studies	6	−.04	.07	−.07	−.18	−.18	−.18	.13
Interrelated Disciplines	7	.18	−.02	−.05	.01	−.15	.03	.05
Relevance Rating Data								
Geography	6	.84*	.30	−.23	.12	−.49*	.20	.93
Economics	6	−.12	−.79*	−.08	.42*	−.39*	−.10	.86
History	5	−.18	.24	.29	−.70*	.08	.01	.60
Political Science	4	−.67*	.36*	.37*	.01	.40*	−.27	.79
Sociology/Anthropology	5	−.34*	.10	−.20	−.44*	.61*	.26	.66
Applied Social Studies	6	−.56*	−.05	−.27	−.33*	.34*	−.17	.60

Note. Relevance ratings were not gathered for the "Interrelated Disciplines" content area.
*$p \leq .05$.

ity and item relevance approaches suggest that when SMEs are aware of the content areas designated in the test blueprint, their judgments tend to implicitly support the test developers' conceptualization of the content domain. When SMEs' judgments of item similarity are gathered before they are informed of the specific content areas comprising the test blueprint, their unique perception of the underlying content structure of the test is revealed.

Although the item similarity approach represents a stronger evaluation of content domain representation, the results suggest that both item similarity and relevance data should be gathered when evaluating content representation. The strong relationships observed between the MDS coordinates and item relevance ratings verified that the MDS configurations were germane to the content structures of the tests. With only one exception (the Evidence and Procedures Auditing content area), the MDS coordinates accounted for more variance in the relevance ratings than in the blueprint classifications. This finding reflects the fact that the relevance ratings contained more information regarding the relationships among the items and content areas than did the blueprint classification data (i.e., the relevance of an item to a content area was rated along a 10-point scale, whereas the blueprint classifications were dichotomous). For example, the MDS coordinates accounted for 60% of the variance of the relevance ratings, but only 13% of the variance of the blueprint classifications for the Applied Social Studies content area (Table 4). Thus, the SMEs considered the content area of Applied Social Studies to be important to the content domain, but did not agree strongly with the blueprint specifications for all the items with respect to this domain.

An advantage in using item relevance ratings is that a statistical test can be conducted to determine whether the SMEs agree that an item is relevant to its specified content area. The statistical significance of the Aiken index provides a practical measure of SME congruence. The Aiken index and averaged relevance ratings provided similar information; therefore, computing both indexes is probably not necessary. The Aiken index appears preferable because it can be evaluated for statistical significance. However, guidelines for interpreting indexes based on averaged relevance ratings have also been proposed (Popham, 1992).

Although the item relevance data were useful in verifying and interpreting the MDS configurations, with the exception of a statistical index of item-objective congruence, they did not provide any unique information beyond that discovered using MDS analy-

Table 5
Averaged Social Studies Content Area Relevance Ratings and the Pearson Correlation Between the Relevance Ratings and the Blueprint Classifications (r)

Content Area	Number of Items	Number Rated First	Percent First	Number Rated Second	Percent First or Second	r
Geography	6	5	83%	1	100%	.50
Economics	6	5	83%	1	100%	.74
History	5	5	100%	0	100%	.31
Political Science	4	4	100%	0	100%	.77
Sociology/Anthropology	6	0	0%	2	33%	.41
Interrelated Disciplines	7	6	86%	—	86%	—
Applied Social Studies	6	1	17%	0	17%	.02

sis of the similarity ratings. There was strong agreement among visual inspection of the MDS space, cluster analysis of the MDS coordinates, and the averaged relevance ratings and Aiken indexes. Items with low Aiken indexes, and content areas with low averaged relevance ratings, were identifiable as incongruous with respect to the test blueprint in the MDS configurations. Thus, although both types of data are recommended for a thorough evaluation of content representation, given a choice between item similarity and item relevance ratings, item similarity ratings are preferred.

Appraising the Test Blueprints

The results provided strong support for many, but not all, of the content areas designated in the blueprints for these two tests. In addition, the results suggested ways in which these blueprints could be revised. Evidence of content domain representation was pronounced for the Auditing content areas of Internal Control, Evidence and Procedures, and Reporting, and for the Social Studies content areas of Political Science and Economics. These content areas exhibited strong clustering within the MDS space, high correlations between the respective stimulus coordinates and relevance ratings, and high item-objective congruence ratings. Other content areas, such as Geography, History, and Professional Responsibilities were less distinct and tended to overlap with items measuring other content areas.

The results suggest that portions of the social studies and auditing test blueprints deviated from the SMEs' perceptions of the content structure of the tests. In particular, it appears that the content areas comprising the social studies content domain were not easily segregated, and that the History content area encompassed two related areas: U.S. history and world history. These findings have important implications, especially if content area-specific subscores are to be derived for diagnostic purposes. In terms of the auditing content domain, it appears that the content area of Professional Responsibilities is highly related to the other content areas and may not represent a unique aspect of auditing practice.

In comparing the utility of the procedure with respect to the two different tests, some differences were noted. Based on the item groupings within the MDS space, the results for the Auditing examination were more congruent with their blueprint specifications than were the results for the Social Studies examination. Several reasons may account for this observation, including the fact that more content areas comprised the Social Studies examination blueprint, more contextual dependencies existed among the social studies items (i.e., more item sets such as items referring to a map or graphic), and the content areas of social studies are more intricately related to one another than are the auditing content areas (e.g., all geography has an associated history).

Utility of MDS for Analyzing Similarity Data

This study demonstrated the strengths and weaknesses of MDS. Perhaps the greatest strength of MDS is that it can capture raters' perceptions of similarity without informing them of the attributes being measured. This attractive feature of MDS is directly related to its primary deficiency. Because the similarity rating task is ambiguous, rather than

Table 6
Aiken Validity Indexes for the Social Studies Items

Content Area and Median Value	Item	Value
Geography		
	1	.86*
	7	.99*
	14	.87*
	15	.97*
	22	.79*
	36	.98*
	Median	.92
Economics		
	18	.99*
	19	.99*
	20	.99*
	30	.70*
	31	.73*
	32	.70*
	Median	.86
History		
	11	.90*
	23	.76*
	34	.96*
	38	.93*
	40	.78*
	Median	.90
Political Science		
	5	.97*
	10	.98*
	13	.91*
	28	.96*
	Median	.97
Sociology/Anthropology		
	2	.29*
	9	.65*
	17	.59
	25	.46
	26	.41
	37	.48
	Median	.47
Applied Social Studies		
	6	.10*
	16	0.00*
	27	.24*
	29	.06*
	35	.03*
	39	.04*
	Median	.05

*$p < .05$.

directive, the opportunity for rating error is increased. Thus, researchers must trade-off between keeping the directions general enough so that the raters provide an unbiased assessment of similarity, yet directive enough so that the raters rate the similarities according to the attributes of interest.

The results of this study complement those of Schaefer, Raymond, & White (1992) who demonstrated that applying MDS and cluster analysis to SME ratings of job task similarity was more informative than applying these techniques to task frequency data when developing test specifications. In the present study, and in Schaefer et al., SME similarity ratings proved more valuable for evaluating the content domain than did more traditional data. As demonstrated by Sireci & Geisinger (1992), as few as three SMEs can be used to evaluate the content structure of a test using item similarity ratings. Thus, this procedure may be especially beneficial to test developers who have access to only a few SMEs, but can use them for relatively long periods of time.

Implications for Future Research

There are two major limitations when using MDS to analyze SMEs' ratings of item similarity. First, the rating procedure places considerable burden on the SMEs, primarily when a large number of test items is involved. Second, when a test comprises a large number of content areas, it becomes increasingly difficult to evaluate the test blueprint using MDS. More dimensions are needed to identify the content structure, and high-dimensional solutions are difficult to interpret, especially when substantial differences exist among the SMEs. To reduce the demands on the SMEs, future research should explore incomplete MDS designs (e.g., Spence, 1982, 1983) and sorting procedures in which the SMEs are required to sort items into a limited number of categories according to their similarity.

If SME congruence is not a concern, future research should consider averaging the similarity data over the SMEs to provide a single matrix for classical MDS analysis. This action would reduce the dimensionality of the data by obscuring dimensions resulting from the idiosyncrasies of a small number of SMEs. However, if future studies wish to explore differences among SMEs (e.g., compare minority SMEs' perceptions to majority SMEs' perceptions) individual differences MDS models are required. Future research

should also consider gathering item response data to determine how factor analysis or MDS analysis of item, parcel, or test score data compare with the dimensions obtained from the SME similarity ratings.

Implications for Test Validity

Content domain representation is critical for demonstrating the validity of inferences derived from test scores (Sireci, 1995; Smith & Greenberg, 1993; Yalow & Popham, 1983). All inferences derived from test scores are valid only to the extent to which the test measures the constructs it purports to measure. Because test specifications and blueprints represent operational definitions of the constructs measured, test developers must demonstrate that: (1) the content specifications adequately define the content domain, (2) the test blueprint adequately represents the content specifications, and (3) the test items adequately represent the test blueprint. The procedures studied here will help test developers evaluate these fundamental attributes of content domain representation. In particular, MDS analysis of SMEs' ratings of item similarity provides both convergent and divergent information regarding the content structure of the test and the adequacy of the test developers' blueprint. Given the recent increase in test accountability, both types of information are essential for verifying the validity of a test for a particular use (Messick, 1989).

References

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

American Institute of Certified Public Accountants. (1990). *Uniform CPA Examination. May 1990: Questions and unofficial answers.* New York: Author.

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

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.

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.

CTB McGraw-Hill. (1989). *Comprehensive Test of Basic Skills* (4th ed., Benchmark Level 17/18). Monterey CA: Author.

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.

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

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

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

MacCallum, R. (1981). Evaluating goodness of fit in nonmetric multidimensional scaling by ALSCAL. *Applied Psychological Measurement, 5,* 377–382.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13–103). Washington D.C.: American Council on Education.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5,* 285–301.

Schaefer, L., Raymond, M., & White, A. S. (1992). A comparison of two methods for structuring performance domains. *Applied Measurement in Education, 5,* 321–335.

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

Sireci, S. G. (1995, April). *The central role of content representation in test validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16,* 17–31.

Smith, I. L., & Greenberg, S. (1993). Content validity procedures. *CLEAR Exam Review, 4,* 19–22.

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.

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

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12,* 10–14.

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

Acknowledgments

The authors thank Michael Green, Richard Lukaschek, Susan Menelaides, Patrick Shelley, and the Research in Social Studies Education SIG of AERA for their help in recruiting the subject-matter experts; the AICPA and CTB McGraw-Hill for permission to use the tests studied; Thanos Patelis for writing a data input program; and Lynn Shelley-Sireci for her editorial assistance. This research was part of the first author's dissertation while at Fordham University, and was supported in part by a Dissertation Research Award from the American Psychological Association. A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, April, 1993. The quality of this report was improved by the first author's dissertation committee (Kevin Moreland, Warren Tryon, and John Walsh) and by suggestions from William Koch and two anonymous reviewers.

Author's Address

Send requests for reprints or further information to Stephen G. Sireci, School of Education, University of Massachusetts, 156 Hills South, Amherst MA 01003, U.S.A. Internet: sireci@acad.umass.edu.