| | |
|---|---|
| ED 428 121 | TM 029 592 |
| AUTHOR | Sireci, Stephen G. |
| TITLE | Evaluating Content Validity Using Multidimensional Scaling. |
| PUB DATE | 1998-04-15 |
| NOTE | 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998). |
| PUB TYPE | Reports - Evaluative (142) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Content Validity; Evaluation Methods; *Multidimensional Scaling; *Research Methodology |

ABSTRACT

Multidimensional scaling (MDS) is a versatile technique for understanding the structure of multivariate data. Recent studies have applied MDS to the problem of evaluating content validity. This paper describes the importance of evaluating test content and the logic of using MDS to analyze data gathered from subject matter experts employed in content validation studies. Some recent applications of the procedure are reviewed, and illustrations of the results are presented. Suggestions for gathering content validity data and using MDS to analyze them are presented. (Contains 3 exhibits, 7 figures, and 24 references.) (Author/SLD)

# Evaluating Content Validity Using Multidimensional Scaling[1,2]

Stephen G. Sireci
University of Massachusetts, Amherst

[1] Paper presented at the annual meeting of the American Educational Research Association as part of the symposium "The Application of Multidimensional Scaling in Educational Research," (C. Shuai-Ding, Chair), April 15, 1998, San Diego, California.

[2] Laboratory of Psychometric and Evaluative Research Report No. 326. Amherst, MA: University of Massachusetts, School of Education.

## Abstract

Multidimensional scaling (MDS) is a versatile technique for understanding the structure of multivariate data. Recent studies have applied MDS to the problem of evaluating content validity. This paper describes the importance of evaluating test content and the logic of using MDS to analyze data gathered from subject matter experts employed in content validation studies. Some recent applications of the procedure are reviewed and illustrations of the results are presented. Suggestions for gathering content validity data, and using MDS to analyze them, are provided.

## Introduction

In educational assessment, evaluating inferences derived from test scores begins with evaluating the test itself. Is the content of the test consistent with the intended purposes of the test? Does the test measure the educational objectives it is intended to measure? Is the content of the test congruent with our school district's curriculum? Are the skills measured by this test representative of the skills required for success in college? Are the test questions appropriate for all test takers? These are fundamental questions of test fairness. These are questions of content validity.

Evaluation of the content validity of a test involves inspection of all aspects of test content: directions, item contexts (passages, graphics, etc.), test questions, distractors, scoring rubrics, and response formats. Thus, the evaluation of content validity is different in kind from other aspects of validity, such as construct or predictive validity, that rely predominantly on analysis of test and item response data. This difference is important, because interpretation of statistical analysis of test and criterion data is meaningless if the content of the test is not first understood and justified.

Clearly, the affirmation of content validity is an important component of evaluating the validity of inferences derived from test scores. However, evaluating content validity is a difficult endeavor. There is no statistical index that can tell us whether a test is content valid. Rather, content validity involves the opinions of those untrustworthy souls known as humans. These humans, sometimes called subject matter experts or curriculum specialists, are needed to evaluate test content and provide their personal opinions regarding what the items measure, and how well the items measure intended objectives. Analysis of the data provided by such humans can be quite daunting to psychometricians who are only familiar with analysis of test and item response data.

The purposes of this paper are to provide some guidance for evaluating the content validity of educational tests and to illustrate how multidimensional scaling (MDS) can be used to help evaluate test content. First, some of the issues central to content validation are briefly described. Then, some reasons for using MDS to evaluate test content are presented. Next, some recent applications of MDS to the content validation problem are presented. These applied studies illustrate the important steps required in a content validity study. Finally, some guidelines for conducting content validity studies are provided.

## Understanding Content Validity

Although the basic tenets of content validity are straightforward, content validity is a controversial concept. The controversy centers primarily on nomenclature, with many arguing content validity is not a "type" or "form" of validity. This argument is drawn from the unitary conceptualization of validity, which describes validity as pertaining to inferences derived from test scores, rather than as a property of a test. In this view, construct validity is described as the general form of validity, and the concepts formerly associated with content validity are given new labels such as content representativeness or content relevance (Messick, 1989; Fitzpatrick, 1983). The unitary conceptualization of test validity is philosophically elegant and correctly emphasizes the importance of validating not the test, but inferences derived from test scores. While this conceptualization has led to increased analysis of test scores and their consequences, unfortunately, it has also resulted in less attention paid to test content (Sireci, in press). This reduced attention is unfortunate because if the content of a test is not well supported by theory, and is not considered acceptable by test evaluators, the validity of inferences derived from the "mysterious" test scores is suspect.

Thus, evaluating the validity of inferences derived from test scores begins with analysis of the test itself. *Content validity* refers to the degree to which the test measures the content domain it purports to measure. Appraising a test for content validity involves evaluating: 1) the adequacy of the test construction procedures, 2) how well the test specifications describe the domain to be tested, and 3) how well the tasks (items) on the test represent the domain of knowledge and skills to which score-based inferences are referenced. The terms *domain definition* and *content domain representation* are often used to describe the latter two components of content validity.

## Traditional Methods for Evaluating Test Content

Traditional methods for evaluating the content validity of a test focus on content domain representation. Typically, subject matter experts (SMEs) are recruited to scrutinize test items and make judgments regarding how well the items measure the knowledge and skills (objectives) they are intended to measure. These judgments are most often in the form of relevance ratings or item-objective congruence ratings. Relevance ratings require SMEs to rate the relevance of each test item to its stated objective, or to all the objectives purportedly measured by the test. A sample relevance rating sheet, from Sireci and Geisinger (1995) is presented in Exhibit 1. Item-objective congruence ratings require SMEs to match each item to one of the objectives purportedly measured by the test. A sample item-objective rating sheet, from Sireci, Rogers, Swaminathan, Meara, and Robin (1997) is presented in Exhibit 2. Relevance ratings are usually made on ordinal, Likert-type scales, whereas item-objective congruence ratings involve dichotomous ratings. Examples of these and related methods for using SMEs to gather content validity data are provided by Aiken, (1980), Crocker, Miller, and Franks (1989), and Hambleton (1984).

An inherent weakness in using relevance or item-objective congruence ratings to evaluate content validity is that the SMEs are informed of the objectives purportedly measured by the test *before* they evaluate the items. In fact, to effectively evaluate items for relevance or congruence to test objectives, the objectives must be clearly understood. This understanding is usually accomplished by providing oral and written descriptions of the measured objectives to the SMEs. Although it is clear this information is critical for completing the ratings, it is not so obvious why such information undermines the validity of the ratings. The problem is, when SMEs are informed of the specific objectives measured by the test, and of how the test developers defined these objectives, their evaluations of how well the items represent the content domain is *biased* by the test developers' conceptualization of that domain. Thus, the SMEs' ratings of the knowledge and skills measured by the items do not represent an independent appraisal of the test. Rather, their ratings are constricted by the test developers' a priori conceptualization of the domain. Furthermore, SMEs' awareness of the test objectives may sensitize them to the test developers' expectations, creating the potential for socially desirable response sets or Hawthorne effects to contaminate their ratings.

In summary, traditional content validity studies using relevance or item-objective congruence ratings have two limitations. First, they do not provide information regarding the adequacy of the test developers' *definition* of the content domain. Second, they are susceptible to systematic biases, such as Hawthorne effects, that will tend to implicitly support the test developers' conceptualization of the content domain tested.

To redress these limitations, it is necessary to gather SMEs' perceptions of the content representation of the test without informing them of the content specifications constructed by the test developers. One method for gathering such data is to use the paired comparison procedure introduced by Thurstone (1927). As described below, using MDS to analyze paired comparison data gathered from SMEs yields valuable information for evaluating content domain definition and representativeness.

## Using MDS to Analyze Content Validity Data

### Gathering Item Similarity Data Using Paired Comparisons

One approach introduced by Sireci and Geisinger (1992) to evaluate content validity is to require SMEs to rate the *similarities* among all pairs of test items with respect to the content measured by each item. To gather these data, all possible pairings of test items are presented to the SMEs, and their task is to provide a similarity rating for each pair along a Likert-type similarity rating scale. Thurstone (1927) introduced the paired comparisons procedure to solve the problem of ordering attitude statements along a unidimensional continuum. In Thurstone's original formulation, experts were asked to inspect each pair of attitude statements and identify the statement that, if endorsed, reflected a greater degree of the attitude measured. The paired comparison method has subsequently been used to gather direct similarity data used to uncover individual's perceptions of the structure of a set of stimuli.

The paired comparisons procedure is a valuable method for discovering individuals' perceptions of the objects under investigation without informing them of what is being studied. Thus, the method is an elegant and simple manner for controlling unwanted sources of bias in ratings such as social desirability and Hawthorne effects. People's perceptions of stimuli are much richer than their ability to articulate these perceptions. Therefore, the less constrained the data gathering task, the more likely the nuances of individual perceptions will emerge. The paired comparison similarity rating task is intentionally ambiguous. The directions do not impose or suggest strict criteria for conducting the ratings. Rather than instructing respondents to "rate the items with respect to characteristic *x*," the characteristics used by the respondents are *discovered* from analysis of the data. With respect to the study of content validity, the procedure allows for discovery of the *perceived* content structure of a test, independent of any one else's a priori conceptualization of this structure.

An example of an item similarity rating sheet, taken from Sireci, et al. (1997) is presented in Exhibit 3. In this study, item similarity rating booklets were constructed for each SME, with a pair of items presented on each page. The items evaluated were part of the 1996 Grade 8 National Assessment of Educational Progress (NAEP) Science Assessment. The SMEs were required to review the item pairs and circle a rating on each page to indicate their perception of the similarity among the items in terms of the science knowledge and skills measured. The results from this study are described briefly below.

## Analyzing Similarity Ratings Using MDS

The logic motivating use of item similarity rating data to evaluate test content is straightforward: items constructed to measure similar aspects of the content domain should be perceived as more similar than items constructed to measure different aspects of the domain. This logic invokes a spatial conceptualization of the content structure of the items comprising a test. For example, a mathematics proficiency test comprising algebra and geometry items could be envisioned in two-dimensional space. One dimension could account for the degree to which the items measured algebra and the other dimension could account for the degree to which the items measured geometry. Such dimensions would be consistent with the "dimensions" (content areas) delineated in the test content specifications. The problem in discovering SMEs' perceptions of content structure is how to best reflect the relationships they perceive among the test items. MDS is an attractive option for analyzing SMEs' similarity rating data because it is designed to portray data structure visually. Thus, MDS representation of SMEs' similarity ratings can be directly compared to the relationships among the items expected from the test content specifications.

### A description of MDS

The goal of most MDS analyses is visual portrayal of latent data structure. The data analyzed are called *proximities*, which can be gathered directly, as in the case of similarity ratings, or can be derived from multivariate data, such as when correlation coefficients are used to represent similarities among variables. MDS analysis provides a visualization of data structure by computing a set of coordinates for the objects (stimuli) to be scaled along one or more dimensions. These coordinates define distances among the stimuli in unidimensional or multidimensional space. The coordinates are computed iteratively, using an optimization criterion that minimizes differences between the MDS distances and a transformation of the original proximity data (e.g., Kruskal, 1964). In non-metric MDS models (the most popular models), the original proximity data are considered to be ordinal, and so the coordinates are computed to best fit the distances to a monotonic transformation of the original proximities.

There are different distance formulae that may be used in a MDS analysis; however, the most popular MDS model uses Euclidean distance. The classical MDS model computes the distances among two stimuli as:

$$d_{ij}=\sqrt{\sum_{a=1}^{r}(x_{ia}-x_{ja})^2} \tag{1}$$

where: $d_{ij}$=the distance between stimuli $i$ and $j$, $X_{ia}$=the coordinate of point $i$ on dimension $a$, and $r$=the dimensionality of the model. When more than one matrix of proximity data are available, a generalization of this model is made to account for individual variation among the matrices. The most common individual differences model is the INDSCAL model developed by Carroll and Chang (1970). This model, also called a weighted MDS model, derives a common set of stimulus coordinates for the group of matrices and a vector of dimension weights for each matrix. These dimension weights can be used to compute "individual" stimulus coordinates to derive a "personal space" for each matrix. Incorporating these weights into equation 1 yields the INDSCAL model:

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

where: $d_{ijk}$=the Euclidean distance between points $i$ and $j$ for matrix $k$, and $w_{ka}$ is the weight for matrix $k$ on dimension $a$. The personal distances for each matrix are related to the common (group) space by:

$$x_{kia} = \sqrt{w_{ka}}\,x_{ia} \tag{3}$$

where $x_{kia}$ represent the coordinate for stimulus $i$ on dimension $a$ in the personal space for matrix $k$, $w_{ka}$ represents the weight of matrix $k$ on dimension $a$, and $x_{ia}$ represents the coordinate of stimulus $i$ on dimension $a$ in the group space.

The weighted MDS model is sometimes called an individual differences MDS model because the $k$ matrices may represent different individuals. In the context of gathering item similarity ratings for a content validity study, the separate matrices of item similarity ratings for each SME could be fit simultaneously using the INDSCAL model. This model would allow for investigation of similarities and differences among the SMEs, in addition to investigation of the similarities among the test items. If individual differences among the SMEs are not of interest, the SME matrices could be averaged to derive a single matrix, or could be fit using a replicated MDS model (Young and Harris, 1993).

## Examples of the Use of MDS to Evaluate Content Validity

This section describes some recent studies that used MDS to evaluate test content. Although studies using both direct and derived proximity data have been applied to this problem, studies utilizing direct proximity data in the form of SMEs' item similarity ratings are emphasized.

### Evaluating the Content Validity of a NAEP Science Assessment

Sireci, et. al. (1997) employed ten science teachers to scrutinize a carefully selected sample of 45 items from the 1996 grade 8 NAEP science assessment. All ten SMEs came from different states, and were selected for participation based on their involvement with science curriculum and/or assessment in their state. All had extensive experience teaching middle school science. The SMEs rated all possible pairings of the 45 items over a two-day period, and received a modest honorarium for their participation.

There were a total of 190 items on the 1996 grade 8 NAEP Science assessment. The 45 items rated by the SMEs were selected to represent the test specifications in terms of the content and cognitive dimensions, as well as item format (multiple-choice, short constructed-response, extended constructed-response). Twelve of the 45 items were from one of the "hands-on" science tasks included on the assessment. The "content framework" for this test specified four dimensions: 1) a "field of science" content dimension comprising earth science, life science and physical science; 2) a cognitive dimension described as "ways of knowing and doing science.," which comprised conceptual understanding, practical reasoning, and scientific investigation; 3) a "themes of science" dimension; and 4) a "nature of science" dimension (NAGB, 1996).

After completing the item similarity ratings, the SMEs were informed of the content frameworks and were subsequently asked to complete item-objective congruence ratings. These ratings were used to help interpret the MDS solutions and to further evaluate the content validity of the items and the framework dimensions.

An INDSCAL model was fit to the ten SME proximity matrices and a five-dimensional solution was accepted. The first two dimensions tended to configure the items according to cognitive level and item format, respectively. This two-dimensional subspace is presented in Figure 1. The horizontal dimension tended to separate the lower cognitive level "conceptual understanding" items from the higher-level "scientific investigation" items. The three conceptual understanding items with negative coordinates on this dimension tended to be rated as measuring higher-level cognitive areas by the SMEs in the follow-up item-objective ratings. All other conceptual understanding items had positive coordinates on this dimension. The correlation between the conceptual understanding item-objective congruence ratings and the coordinates on this dimension was .80. The vertical dimension tended to separate the practical reasoning and scientific investigation cognitive areas. However, closer inspection of the items suggested that this dimension more directly separated the multiple-choice items from the constructed-response items. After dummy-coding the items for format, the correlation between the dichotomous format variable and coordinates on dimension 2 was -.76. Figure 2 presents a three-dimensional subspace of the first three dimensions, which were related to cognitive area. Although some cognitive area overlap is evident, clusters of items from the same cognitive area are evident The visual and correlational analyses suggest strong correspondence between the cognitive knowledge and skills measured by the items and their coordinates in this three-dimensional subspace.

Figure 3 presents a second two-dimensional subspace that tended to account for the content area distinctions among the items. All but one of the life science items had negative coordinates on the horizontal dimension. This item was classified as an earth science item by three of the SMEs. The item-objective congruence ratings for the earth and life science areas correlated .61 and -.65 with the coordinates for the items on this dimension. As for the vertical dimension, only one physical science item exhibited a large positive coordinate. This item was classified as an earth science item by eight of the ten SMEs. The physical science item-objective congruence ratings correlated -.75 with the item coordinates on this dimension. The visual and correlational analyses suggest a strong relationship between the content designations of the items and their coordinates in this two-dimensional subspace. Although some overlap among the content areas is evident, in general, the items comprising any one of the three fields of science content areas tend to be configured more closely to one another than they are to items from other content areas.

In summary, analysis of the item similarities data using MDS uncovered cognitive- and content-related dimensions that were congruent with those dimensions specified in the NAGB frameworks. Items that did not tend to group together with the other items in their content or cognitive area tended to be the same items that were identified as problem items from analysis of the item-classification congruence ratings. The other two dimensions specified in the NAGB frameworks (nature of science and themes of science) did not emerge in the MDS solution.

### Analysis of a Licensure Exam and a Social Studies Achievement Test

Sireci and Geisinger (1995) analyzed item similarity ratings for two very different tests: the auditing section of the Uniform CPA Examination, and a nationally-standardized middle school social studies achievement test. Two separate groups of 15 SMES (i.e., 15 CPAs specializing in auditing, and 15 middle school social studies teachers) provided the item similarity ratings. As in Sireci et al. (1997), an INDSCAL model was fit to the data for each SME group. The CPA data illustrated strong congruence between the test specifications and the MDS solution. The social studies data illustrated less congruence with the test specifications (see below). Analysis of the subject weights for the social studies SMEs also revealed some interesting differences among the SMEs.

Six-dimensional MDS solutions were selected for both the auditing and social studies data. Figure 4 presents a two-dimensional subspace of the auditing data. The auditing exam comprised four content areas. The content area designations for each item are indicated in the legend. The horizontal dimension separated items measuring the reporting content area from the other items.

The vertical dimension did not directly reflect a content distinction specified in the test blueprint. Rather, this dimension separated items measuring knowledge of auditing standards from items measuring application of these standards. The ellipses encircling the items in Figure 4 are based on groupings obtained from a hierarchical cluster analysis of the item coordinates from the complete six-dimensional solution. The clustering results are strongly related to the general content structure of the exam. Figure 4 illustrates the usefulness of cluster analysis for discovering subsets of items close to one another in high-dimensional space, and portraying them in two-dimensional space.

In addition to making item similarity ratings, the auditing and social studies SMEs were also required to rate the *relevance* of each item to each of the content areas measured on the test. These data were used to help interpret the MDS dimensions, and uncover differences between traditional content validity data and item similarity data. The relevance data were regressed across the coordinates from each six-dimensional solution. The multiple regression analysis can be embedded into the MDS space by projecting an attribute vector, corresponding to the dependent variable, into the MDS space. The direction of the vector in the space corresponds to increasing amount of the dependent variable. Furthermore, when the multiple correlation is high (e.g., above .80 and $p < .01$) and the regression weight for a dimension is large, the angle between the attribute vector and the dimension will be small, indicating congruence between the attribute and the dimension. Figure 5 illustrates an attribute vector drawn into another two-dimensional subspace from the six-dimensional auditing solution. The attribute vector is the result of a multiple regression analysis of the relevance ratings for the "reporting" content area across the coordinates from a six-dimensional solution ($R^2$=.93). The angle between the reporting attribute vector and Dimension 1 is about 37 degrees. This angle is calculated by taking the inverse cosine of the normalized regression weight for the dimension. The close correspondence between this attribute vector and the dimension illustrates that, in making their similarity ratings, the SMEs' strongly considered the relevance of the items to the "reporting" aspects of a professional audit.

A two-dimensional subspace of the social studies solution is presented in Figure 6. The horizontal dimension tends to separate the six items measuring geography from the other items, and the vertical dimension tends to separate the six economics items from the others. Thus, the MDS solution suggests that the teachers perceived the geography and economic content characteristics of these items when making their similarity judgments. Interestingly, the sixth dimension of the solution (not shown) separated items measuring American history from items measuring world history. This content distinction was not part of the content specifications of the test. Thus, the "definition" of the content domain derived from the SMEs differed from that articulated by the test developers.

Figure 7 presents the subject space for the two dimensions displayed in Figure 6. The vectors portrayed in this figure represent the relative differences among the teachers in their use of the "Geography" and "Economics" dimensions. The tip of each vector is a point whose coordinates equal the teacher's weights along the two dimensions (i.e., the estimates of the weights $w_{ka}$ in Eq. 2). The closeness of teacher "13" to the horizontal dimension indicates this teacher had a substantial weight on the Geography dimension and a near zero weight on the Economics dimension. Thus, this teacher attended heavily to the geography characteristics of the items and virtually ignored the economic characteristics of the items. Teacher "8," on the other hand, provides an example of someone who essentially used both of the dimensions equally. The two-dimensional subset of weights for this subject (.40 and.36) portray her vector along a 45° angle between these two dimensions. Teacher "5" has a near zero weight on both dimensions and did not seem to emphasize either dimension in making her similarity ratings. Follow-up analysis revealed that this teacher primarily rated the item similarities based on cognitive, not content, considerations[3].

Summary. The results of the NAEP, CPA Exam, and social studies analyses indicate how MDS can be used to evaluate the content structure of a test. By using SMEs to provide both item similarity ratings and traditional relevance or congruence ratings, the independent structure perceived by the SMEs can be compared to the hypothesized structure intended by the test developers. In cases where structure is revealed that is not captured in the test specifications, the content domain definition provided by the test developers may be inadequate. The American versus world history distinction noted for the social studies data provides one example. These studies also reveal that MDS is useful for studying differences among SMEs with respect to their similarity ratings.

## Evaluating Test Content Using MDS Analysis of Derived Proximity Data

The content structure of a test can also be evaluated by MDS analysis of test or item response data (e.g., Guttman, et al., 1990; Napior, 1971; Oltman, Stricker, & Barrows, 1990). An exemplary study using MDS analysis of both direct and derived proximity data was provided by Deville (1996). This study used SMEs' item similarity ratings (direct proximity data) and squared Euclidean distances among the items (proximity data derived from the person-by-item matrix) to evaluate the structure among the items. The test data analyzed comprised 32 items from a "can-do" language self-assessment. Using cannonical correlation, Deville found that the dimensional structure derived from the SMEs' ratings was highly related to the structure derived from the item response data. Given the different types of data analyzed, he concluded the procedure provides evidence of both content- and construct-related validity. Regardless of debates surrounding validity nomenclature, Deville's study illustrates the utility of MDS for evaluating content structure using both direct and derived proximity data. When these very different types of data lead to similar conclusions regarding content structure, greater understanding of the content

[3] The distance of the points plotted in Figure 7 to the origin reflect the proportion of variance of their (transformed) proximity data accounted for by the stimulus coordinates of their personal space. Distances between the endpoints of the vectors in the weight space cannot be interpreted as distances between points in the space. It is the *direction* and *length* of the vectors in the space that describes the variation among matrices. The variance of the transformed proximities accounted for by the personal space for a matrix is given by the square root of the sum of the squared weights (Davison, 1992; Young & Harris, 1993). The difference in the length of the weight vectors for teachers 8 and 5 indicate the large difference in the percentage of variance of these two teachers' disparity data accounted for by the two-dimensional subspace (30% versus 3%).

domain measured is achieved.

## Guidelines for Conducting Content Validity Studies Using MDS

The previous sections illustrated the utility of MDS for evaluating the content structure of a test. Studies using derived proximity data are relatively straightforward from an experimental perspective. Item response data from a test administration are available and there are essentially only two critical choices to be made: 1) type of inter-item proximity matrix to derive (e.g., tetrachoric correlations if the items are score dichotomously), and 2) the type of MDS model to fit to the data (e.g., weighted or unweighted). Studies involving direct proximity data involve consideration of many more issues, most pertaining to how the data are gathered from the SMEs. This section provides guidelines for gathering item similarity data and traditional ratings of item relevance and congruence.

Studies gathering direct proximity data must be carefully designed. Critical issues threatening the internal validity of the data are: SMEs' lack of comprehension of the rating tasks, systematic response biases in ratings due to improper ordering of item pairs, SME fatigue effects, and inaccurate or incomplete interpretation of the MDS solution. Critical issues threatening the external validity of the data are non- representativeness of the SMEs and lack of reliability of the similarity ratings. The 15 guidelines presented below are based on experience in gathering content validity data (e.g., Sireci & Geisinger, 1992, 1995; Sireci et al., 1997) and should help improve the internal and external validity of the results from content validity studies.

1) Select competent and representative SMEs: As in any study involving the use of expert judges, the qualifications of the experts is critically important. The SMEs used in content validity studies must be familiar with the content tested and with the knowledge and skill levels of the tested population. The panel of SMEs should also be representative of the pool of potential SMEs. Important demographic variables, such as geographic, racial, and ethnic diversity in the population, should be represented in the sample. Variability with respect to specializations within the domain of content tested should also be represented. For example, the SMEs used to evaluate the auditing items in the Sireci and Geisinger (1995) study included both professional auditors working in the private sector and those working in not-for-profit organizations. Acquiring a competent, diverse, and representative sample of SMEs is difficult; especially considering that the number of SMEs is usually small (i.e., typically 15 or fewer). The selection of SMEs for content validity studies is similar to the selection of panelists used in standard setting studies. Jaeger (1991) provided some useful suggestions for selecting standard setting panelists that are applicable to selecting SMEs for content validity studies.

2) Select representative samples of items: When the number of items on a test is large, say 50 or more, it will be difficult to analyze the content structure of the entire set. One limiting factor is the number of elements in the paired comparisons matrix. For *n* items there are *n*(*n*-1)/2 comparisons (e.g., 900 comparisons for 45 items). A second limiting factor is interpretation of the MDS solutions (or factor analytic solutions) when there are many stimuli to be displayed. A

modest solution to this problem is to select a representative sample of items to be studied from the entire pool of items. This strategy is particularly applicable for evaluating the content validity of pools of test items, such as those used in computerized-adaptive tests.

3) Use rating scales of sufficient length: Another important consideration is the number of scale points on the item similarity rating scale. There is no standard number of points to use, but there are at least two factors to consider. First, shorter scales, such as four- or five-point scales, may result in more undesirable "ties" in the data. That is, pairings of stimuli that truly differ with respect to their similarities may be given the same similarity value. Thus, scales with more points are desirable. On the other hand, longer scales may be overly burdensome for the respondents. Sireci and Geisinger (1992) used a five-point scale and found an excessive number of ties in the data; however, Sireci and Geisinger (1995) used a ten-point scale and found that many SMEs did not use the full scale. Sireci et al. (1997) used an eight-point scale, which seemed to work well with the SMEs in their study. Davison (1992) suggests using scales containing between six to nine response categories (p. 42). This advice is consistent with our experiences. Additionally, we prefer even-numbered scales to prevent SMEs from excessive use of the neutral point.

4) Familiarize SMEs with stimuli and rating tasks: Paired comparison similarity ratings may seem strange at first to SMEs. It is important for SMEs to become familiar with the set of test items, and with the rating task, before making their ratings. One strategy for familiarizing SMEs with the items is to have them take the test under conservative time constraints (i.e., standardized time limit or slightly less time). To familiarize SMEs with the rating task, a few sample pairs could be rated privately, and then discussed as a group. In such discussions, it is important to let the SMEs know that differing perceptions are appropriate, and that they are on-target with respect to rating the similarities. Occasionally, a SME who is using irrelevant criteria (e.g., items close to one another in the test booklet are rated as more similar) can be corrected. The point of training is to avoid inconsistencies in the similarity ratings for an SME due to them having to learn the item characteristics and similarity rating task as they go along.

5) Make rating task easy for the SMEs: When designing a method for gathering similarity ratings, two strategies can be chosen: minimize the work the investigator needs to do in developing materials, or minimize the burden on the SMEs in reviewing the items and recording their ratings. Although it takes more preparation time, the second method should be used. Preparing individual booklets for each SME, with a rating scale below each item pair, seems to work best (see Exhibit 3). The SMEs review the pair of items on each page and enter their ratings directly into the booklet. Before handing in their booklets, the SMEs should check the booklets for inadvertently omitted ratings.

6) Order item pairings systematically or use multiple random orderings: The specific ordering of item pairs presented to SMEs can affect their similarity ratings. To avoid these problems, the item pairs should be ordered in random or systematic fashion so that the rate at which an item appears is consistent across items. Each item should also appear as the first item in a pair about the same number of times as it appears as the second item. These two

recommendations aim towards counterbalancing the ordering of items so that order effects (i.e., space and time effects, Davison 1992) do not contaminate respondents' perceived similarity ratings. Ross (1934) provided an algorithm, computerized by Cohen and Davison (1973), for counterbalancing stimuli to be used in paired comparison rating tasks. Another important issue is whether a common ordering should be used across all respondents. Some designs provide a unique random ordering of item pairs for each SME.

7) Provide frequent breaks: Many researchers argue that paired comparison ratings are impracticable because of the large number of ratings to be made. For example, all possible pairings among the 45 NAEP items studied by Sireci et al. (1997) involved 900 similarity ratings. Although this may seem like an unreasonable task for the SMEs, it need not be. If SMEs are given adequate time to complete their ratings, are allowed to take frequent breaks, and are invested in the importance of the study, large numbers of similarity ratings are not problematic. For example, the item ratings gathered in the NAEP study were collected over a two-day period.

8) Provide incentives: Another important mechanism for keeping SMEs motivated and on-task is to provide incentives for their participation. Monetary compensation is a popular incentive. Knowledge that they are providing an important service to the profession is another. Content validity data are invaluable for supporting test validity. SMEs should be paid for providing this important information whenever possible. Monetary compensation may also facilitate acquisition of more competent SMEs.

9) Consider incomplete MDS designs: In some cases, time may not allow for the administration of all possible item pairings to all SMEs. In such cases, an incomplete paired comparison design may be necessary. There are two general types of incomplete designs. The first type limits the number of inter-stimulus ratings, resulting in an incomplete matrix of inter-stimulus similarities (Spence, 1982, 1983). The other method is to require that all inter-stimulus comparisons be made by a subset of respondents, rather than by all respondents. This strategy was used by Sireci et al. (1997). Across the ten SMEs, seven rated each stimulus pairing. This strategy reduced the number of similarity ratings required from each SME by 200 (i.e., 700 similarity ratings rather than 900), but still provided multiple ratings for all possible stimulus pairs. Thus, the similarity matrix for each respondent was incomplete, but a complete inter-item proximity matrix could be derived across the respondents. Sorting procedures also have been suggested for reducing the burden on SMEs; however, much information is lost when sorting procedures are substituted for paired comparisons.

10) Gather data on SMEs' comprehension of rating tasks: It is important to ensure that the SMEs understood the rating task. This information can be obtained using an exit survey.

11) Gather data on criteria used by SMEs: An exit survey is also useful for discovering criteria used by the SMEs in making their similarity judgements. Although the SMEs may not be able to articulate all criteria used, the criteria they do list should be helpful for evaluating the MDS solution. This information should be collected after they complete their similarity ratings.

12) Include replicated item pairs: One way to evaluate the reliability of SMEs' ratings is to repeat some item pairings. The replicated item pairs should also reverse the order in which the items were presented. If the absolute difference in ratings across these pairs are small, evidence is provided that the SMEs ratings are reliable.

13) Determine whether a weighted or unweighted MDS model should be used: Weighted MDS models are valuable for discovering differences among the SMEs. However, fitting a MDS model to multiple matrices (three-way data) will result in poorer data-model fit in comparison to an analysis that averages the similarity ratings across SMEs and analyzes the single, averaged proximity matrix. Obviously, averaging over SMEs results in loss of information, but it can improve interpretation of the stimulus space. Thus, the importance of discovering differences among the SMEs should be considered in deciding how to analyze the data.

14) Gather external data on items: A quality MDS study includes both proximity data and other data regarding characteristics of the stimuli to be scaled. In studies of content validity, ratings of item-to-content area relevance, and item-objective congruence ratings, are important sources of external information. Examples were provided earlier of how such data can be used to facilitate interpretation of MDS solutions. Ratings of item relevance and congruence can be gathered from the same SMEs who conducted the similarity ratings, or from an independent group. If the same group of SMEs is used, it is critical that these data be gathered after they complete the item similarity ratings. ANOVA, multiple regression, and canonical correlation are useful procedures for relating these external data to the MDS coordinates.

15) Use cluster analysis to interpret high-dimensional MDS solutions: MDS solutions of item similarity ratings tend to be high-dimensional solutions. It is difficult to visually interpret solutions in greater than two or three dimensions. Cluster analysis is a useful technique for helping evaluate higher-dimensional solutions. As illustrated in Figure 4, item groupings within six-dimensional space can be discovered using cluster analysis, and then portrayed in a two-dimensional subspace.

## Conclusion

This paper briefly described content validity theory and the logic of using MDS analysis of item similarity data to evaluate test content. Several studies using this procedure provided illuminating information regarding data structure. In some cases the results supported the content structure claimed by the test developers; in other cases the results suggested a different content structure. MDS analysis of item similarity data provides unique information beyond that gathered in traditional content validity studies. However, an investigation of content validity should include both item similarity ratings and more traditional ratings of item relevance or congruence. Further investigation of test structure using MDS and cluster analysis should shed new light on the content structure of educational tests, and should provide further guidance for conducting these studies.

## References

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

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

Cohen, H. S., & Davison, M. L. (1973). Jiffy-scale: A Fortran IV program for generating Ross-ordered paired comparisons. Behavioral Science, 18, 76.

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

Davison, M.L. (1992). Multidimensional scaling. Malabar, FL: Krieger.

Deville, C. W. (1996). An empirical link of content and construct validity evidence. Applied Psychological Measurement, 20, 127-139.

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

Guttman, R., Epstein, E. E. Amir, M., & Guttman, L. (1990). A structural theory of spatial abilities. Applied Psychological Measurement, 14, 217-236.

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

Jaeger, R. M. (1991). Selection of judges for standard setting.. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1-27.

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

Napior, D. (1972) Nonmetric multidimensional techniques for summated ratings. In Shepard, R. N.; Romney, A.K.; and Nerlove S.B. (Eds.), Multidimensional scaling: Volume 1: Theory (pp. 157-178). New York: Seminar Press.

National Assessment Governing Board (1996). Science framework for the 1996 National Assessment of Educational Progress. Washington, DC: Author.

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

Ross, R. T. (1934). Optimum orders for presentations of pairs in paired comparisons. Journal of Educational Psychology, 25, 375-382.

Sireci, S.G. (in press). The construct of content validity. Social Indicators Research.

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

Sireci, S. G., & Geisinger, K. F. (1995). Using subject matter experts to assess content representation: A MDS analysis. Applied Psychological Measurement, 19, 241-255.

Sireci, S.G., Rogers, H. J., Swaminathan, H., Meara, K., & Robin, F. (1997). Evaluating the content representation and dimensionality of the 1996 Grade 8 NAEP Science Assessment. Commissioned paper by the National Academy of Sciences/National Research Council's Committee on the Evaluation of National and State Assessments of Educational Progress, Washington, DC: National Research Council.

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

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

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286.

Young, F. W., & Harris, D. F. (1993). Multidimensional scaling. In M.J. Noursis (Ed.). SPSS for windows: Professional statistics (computer manual, version 6.0) (pp. 155-222). Chicago, IL: SPSS, Inc.

## Exhibit 1

Please use the following scale to rate the relevance of the test items to the content areas specified in the test blueprint. Provide four relevance ratings for each item (i.e., rate each item to all four content areas).

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| **Not at all relevant** | | | | | | | | | **Highly Relevant** |

Professional Responsibilities    Internal Control    Evidence & Procedures    Reporting

1. The scope and nature of an auditor's contractual obligation to a client ordinarily is set forth in the

a. management letter.
b. scope paragraph of the auditor's report.
c. engagement letter.
d. introductory paragraph of the auditor's report    PR ______ IC ______ EP ______ RP ______

2. Before issuing a report on the compilation of financial statements of a nonpublic entity, the accountant should

a. apply analytic procedures to selected financial data to discover any material misstatements.
b. corroborate at least a sample of the assertions management has embodied in the financial statements
c. inquire of the client's personnel whether the financial statements omit substantially all disclosures.
d. read the financial statements to consider whether the statements are free from obvious material errors.    PR ______ IC ______ EP ______ RP ______

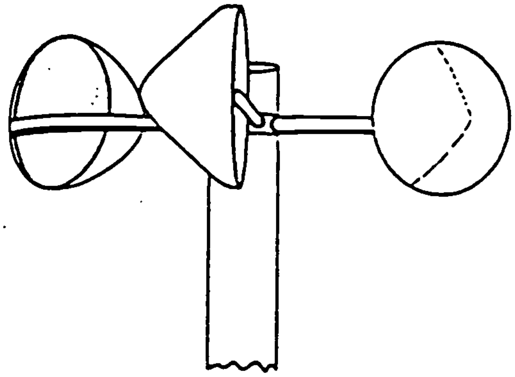
**Exhibit 2**

NAEP Science Assessment - Item Rating Form

| | Field of Science (choose one) | | | Knowing and Doing Science (choose one) | | | Themes (choose one) | | | | Nature of Science (choose one) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item # | Life Science | Physical Science | Earth Science | Conceptual Understanding | Scientific Investigation | Practical Reasoning | Patterns of Change | Models | Systems | None | Yes | No |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |

## Exhibit 3

Sample Item Similarity Rating Sheet



2. The instrument shown above is used to measure

A wind direction

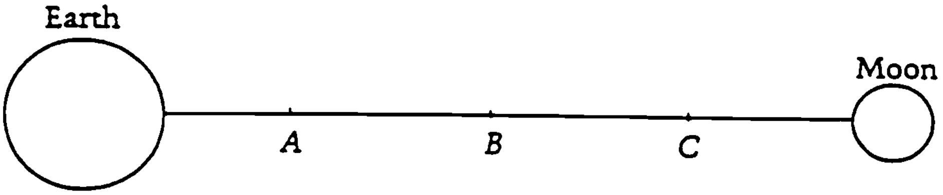
B wind speed

C air pressure

D relative humidity

ILC01078

15 A space station is to be located between the Earth and the Moon at the place where the Earth's gravitational pull is equal to the Moon's gravitational pull. On the diagram below, circle the letter indicating where the space station should be located.



Explain your answer.

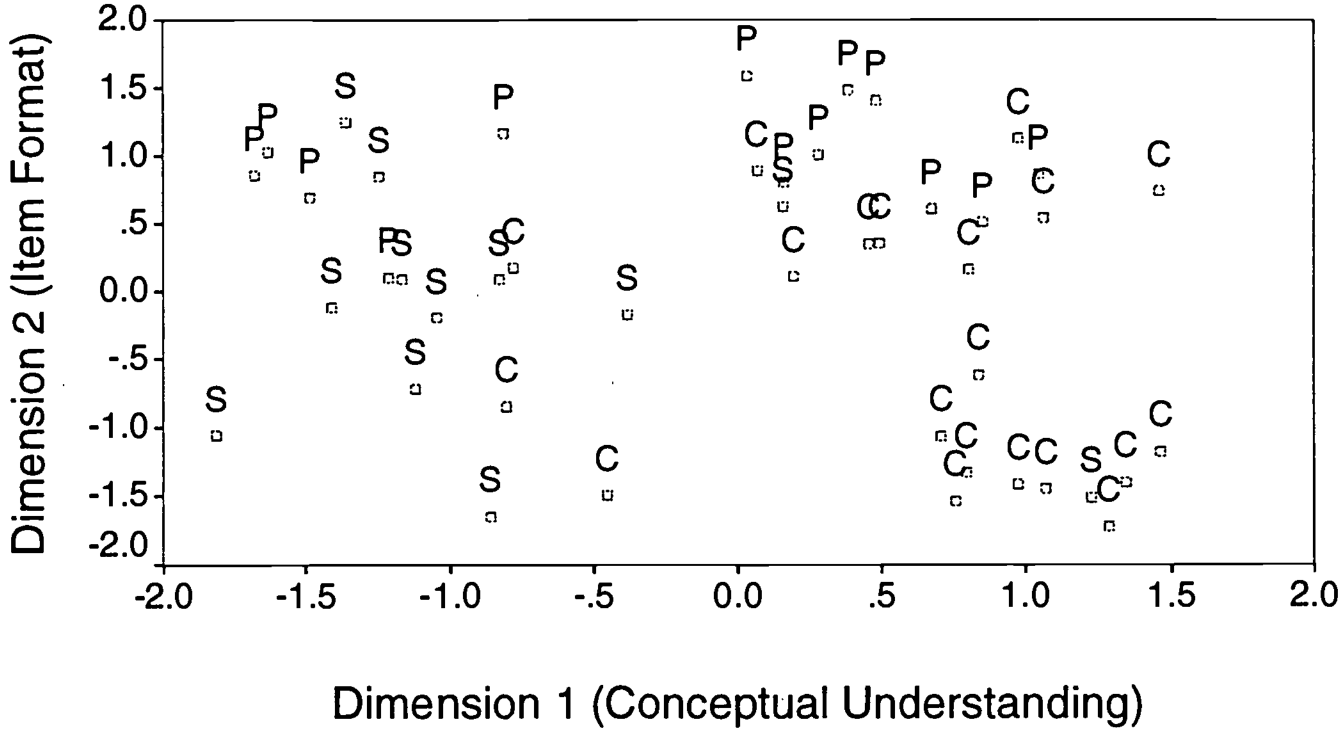
HE001703

Very Similar Very Different

1 2 3 4 5 6 7 8

# Figure 1

## Two-Dimensional NAEP Item Subspace

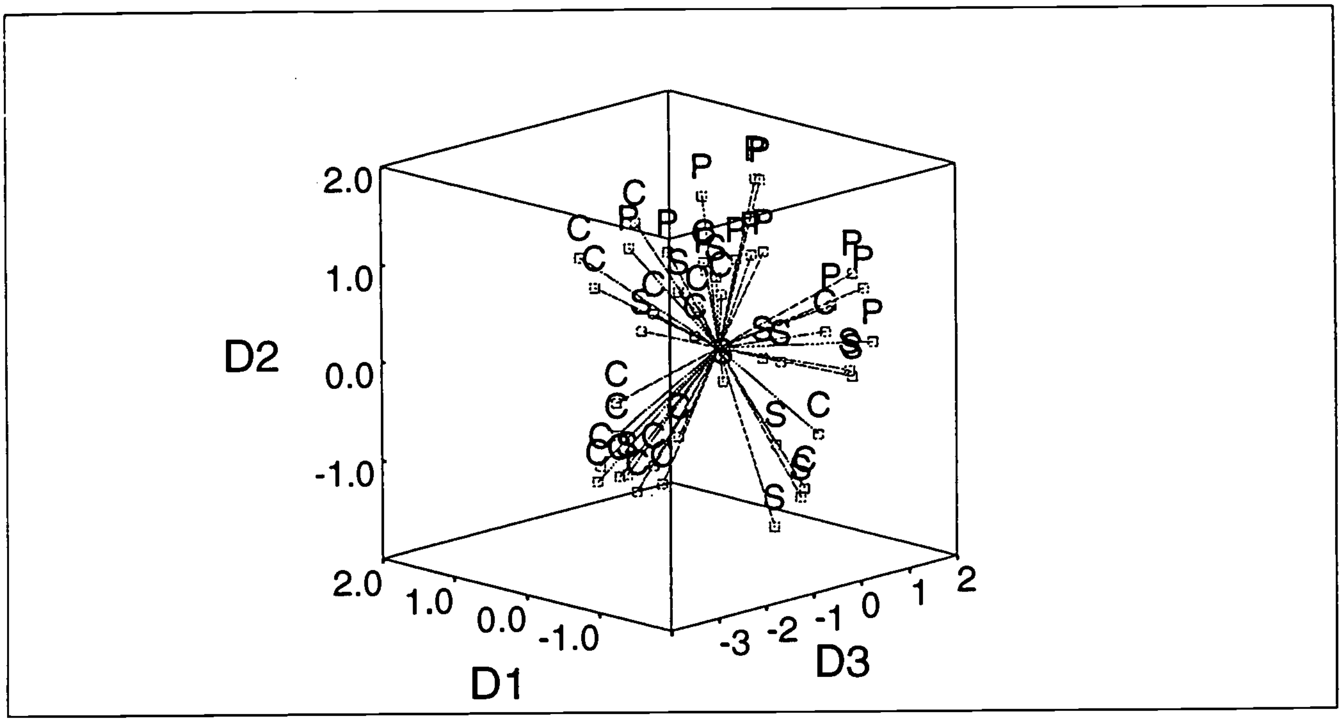


C=conceptual understanding, P=practical reasoning, S=sci. investigation

Source: Sireci, Rogers, Swaminathan, Meara, & Robin (1997)

# Figure 2

## Three-Dimensional NAEP Item Subspace

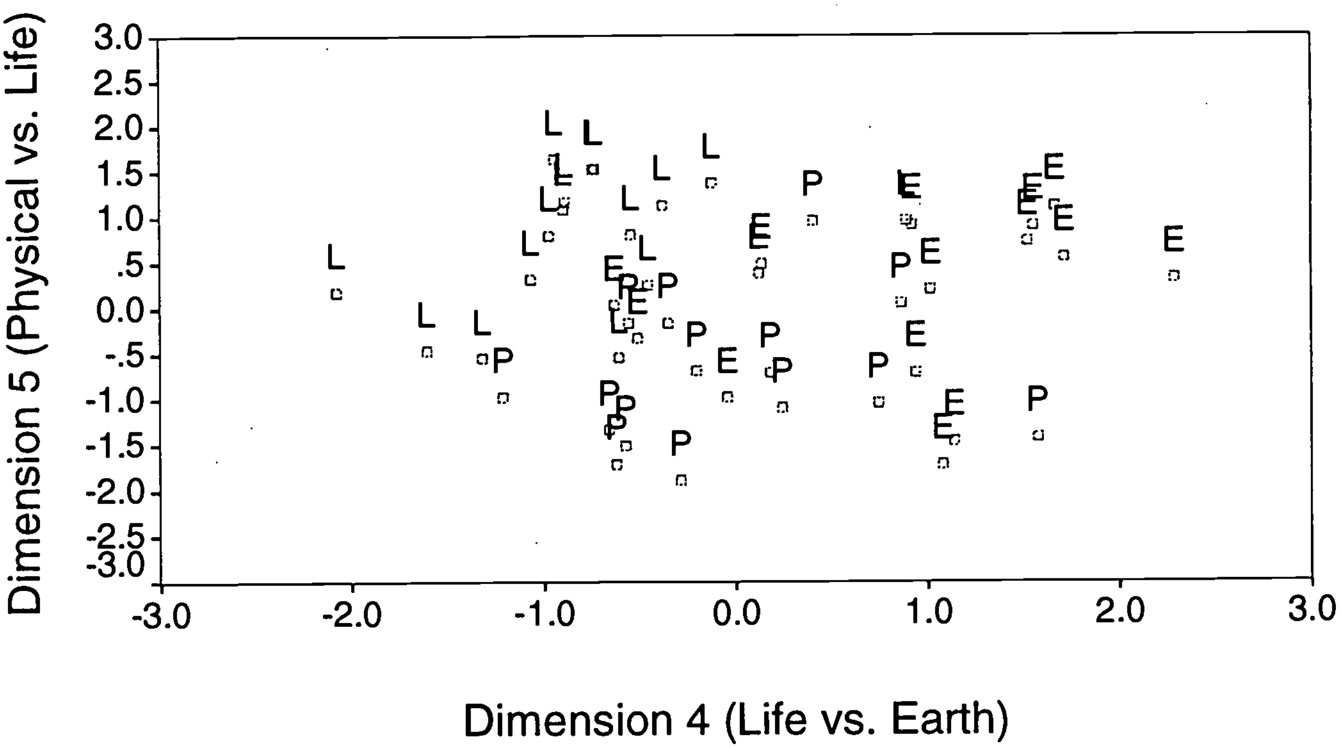


C=conceptual understanding, P=practical reasoning, S=sci. investigation

Source: Sireci et al., (1997)

# Figure 3

## Two-Dimensional NAEP Item Subspace



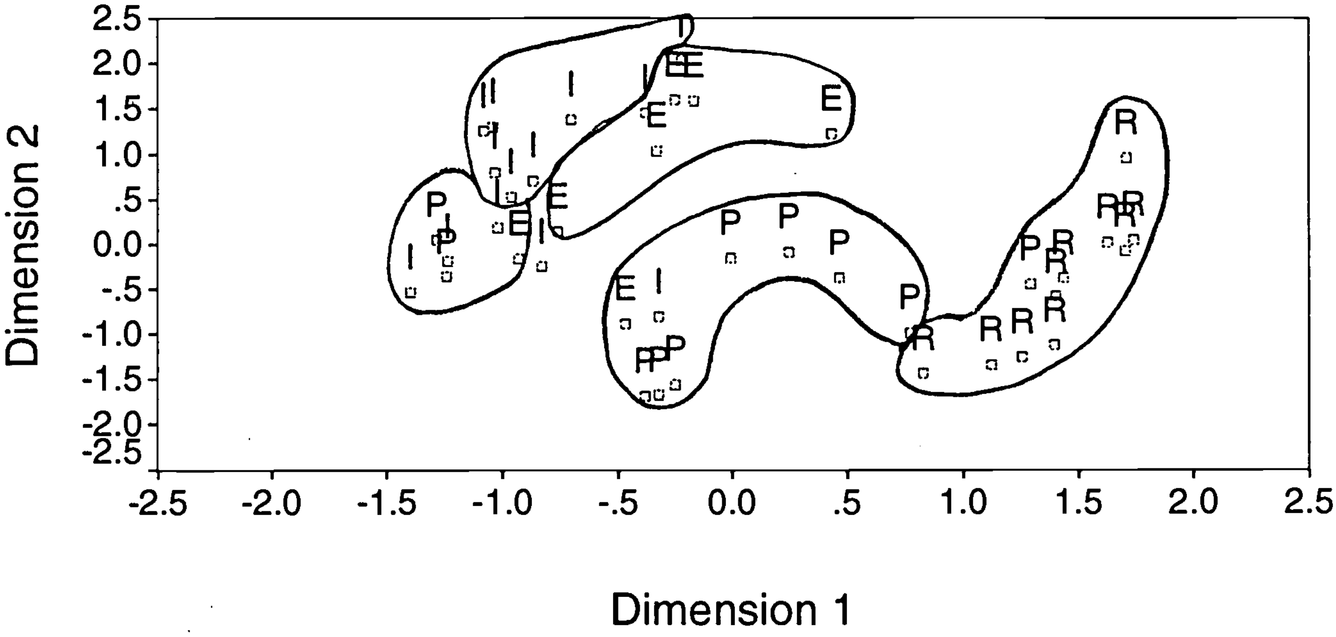
E=Earth Science, L=Life Science, P=Physical Science

Source: Sireci, et al. (1997)

# Figure 4

## Stimulus Space of 40 Auditing Items

Source: Sireci & Geisinger (1995)



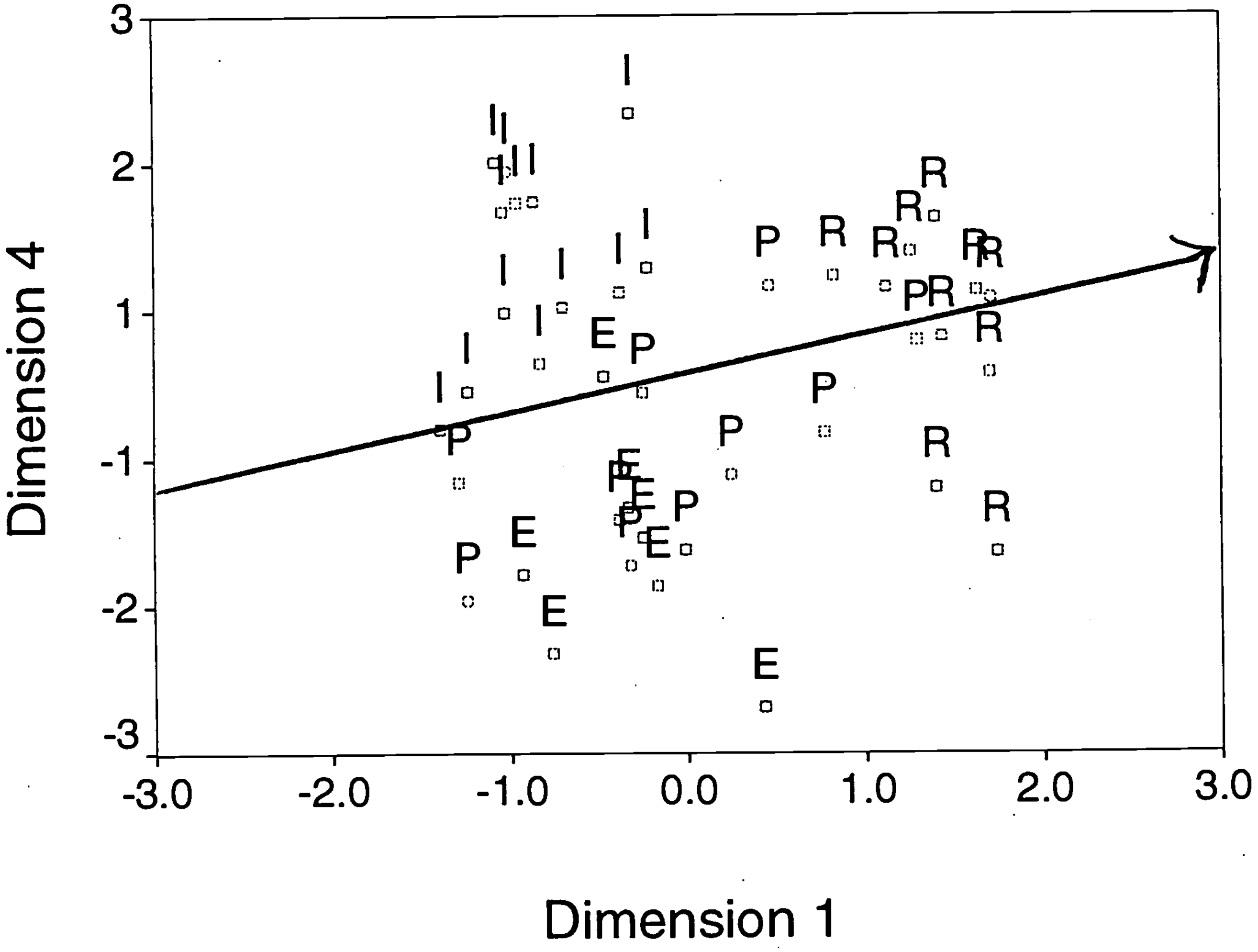
Content Areas: R=Reporting, E=Evidence & Procedures

I=Internal Control, P=Professional Responsibilities

# Figure 5

## "Reporting" Vector Drawn Into 2-D Subspace

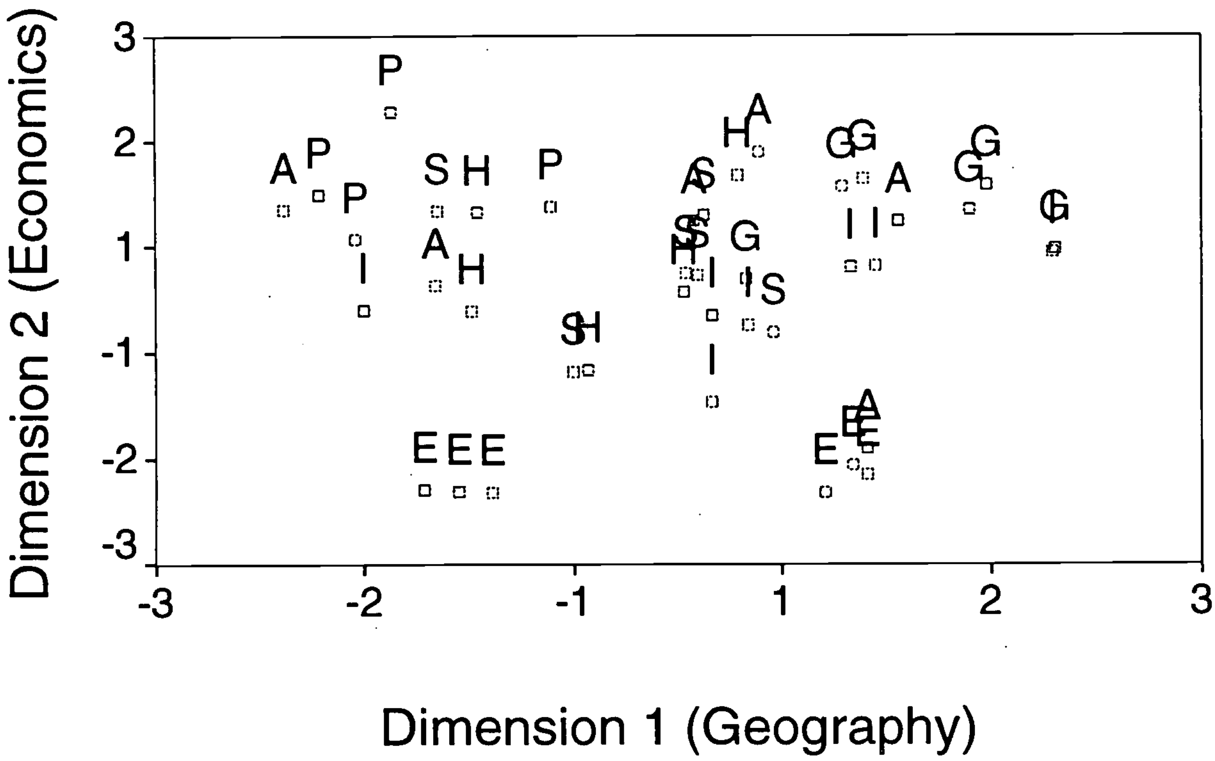
Source: Sireci & Geisinger (1995)



# Figure 6

## Stimulus Space of Social Studies Test Items

Source: Sireci & Geisinger (1995)



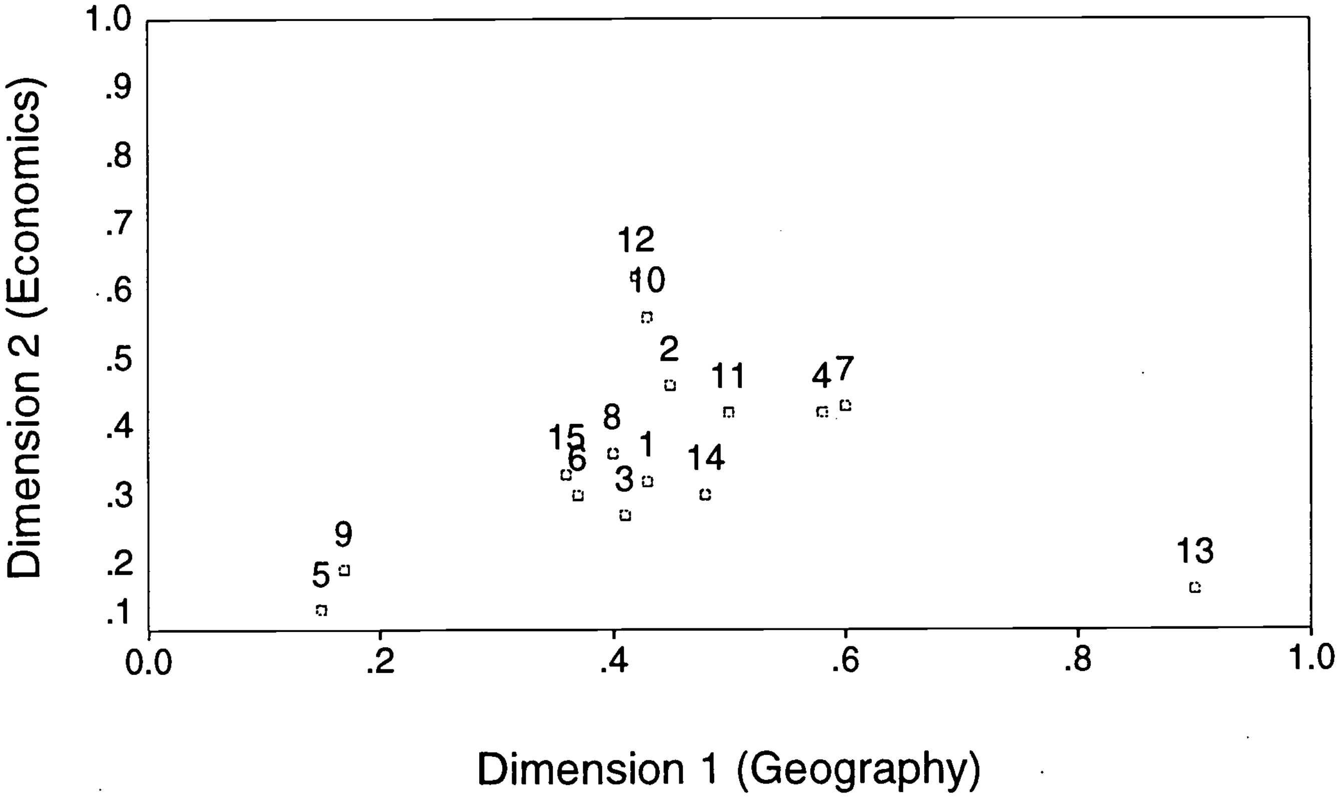
Content Areas: G=Geography, E=Econ., H=History, P=Pol. Sci.

S=Sociology/Anthro., I=Interrelated, A=Applied Soc. Studies

Figure 7

Weight Space of 15 Social Studies Teachers

Source: Sireci & Geisinger (1995)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

TM029592

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Evaluating Content Validity Using Multidimensional Scaling

Author(s): Sireci, S.G.

Corporate Source: University of Mass. - Amherst

Publication Date: 1998

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑

Check here
**For Level 1 Release:**
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) ***and*** paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

**Level 1**

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

**Level 2**

□

Check here
**For Level 2 Release:**
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

*"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."*

**Sign here→ please**

Signature:

Printed Name/Position/Title: Stephen G. Sireci

Organization/Address:
University of Massachusetts
156 Hills South
School of ED
Amherst, MA 01003-4140

Telephone: 413-545-0564

FAX: 413-545-1523

E-Mail Address: sireci@acad.umass.edu

Date: 2/5/99

(over)

## III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

## IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

## V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

**ERIC Processing and Reference Facility**
**1100 West Street, 2d Floor**
**Laurel, Maryland 20707-3598**

**Telephone: 301-497-4080**
**Toll Free: 800-799-3742**
**FAX: 301-953-0263**
**e-mail: ericfac@inet.ed.gov**
**WWW: http://ericfac.piccard.csc.com**