Graphical Representation of Multidimensional Item Response Theory Analyses

Terry Ackerman

University of Illinois, Champaign-Urbana

This paper illustrates how graphical analyses can enhance the interpretation and understanding of multidimensional item response theory (IRT) analyses. Many unidimensional IRT concepts, such as item response functions and information functions, can be extended to multiple dimensions; however, as dimensionality increases, new problems and issues arise, most notably how to represent these features within a multidimensional framework. Examples are provided of several different graphical representations, including item response surfaces, information vectors, and centroid plots of conditional two-dimensional trait distributions. All graphs are intended to supplement quantitative and substantive analyses and thereby assist the test practitioner in determining more precisely such information as the construct validity of a test, the degree of measurement precision, and the consistency of interpretation of the number-correct score scale. *Index terms: dimensionality, graphical analysis, multidimensional item response theory, test analysis.*

Most psychological and educational tests measure, to different degrees, multiple traits or trait composites. As such, test practitioners must establish the construct validity of each test and subsequently provide users with an interpretation of what the test measures. If a test measures several traits simultaneously, questions that need to be addressed include: (1) what composite of traits is being measured? (2) of the traits being measured, which are primary (i.e., intended to be measured) and which are secondary (i.e., not intended to be measured)? (3) how accurately are each of the various composites being assessed? (4) what is the correct interpretation of the number-correct (or standard) score scale? (5) is this interpretation consistent throughout the entire number-correct score range, or do low scores reflect levels of one composite trait and high score levels reflect another composite trait? and (6) do the secondary traits result in differential performance between identifiable groups of examinees?

Typically, item, test, and differential item functioning (DIF) analyses are conducted after every administration of a standardized test. The purpose of this paper is to present a series of graphical representations of multidimensional analyses that can supplement these analyses. The goal of pictorially representing quantitative results is to help the practitioner gain insight into the measurement process and, thus, strengthen the relationship between the test construction process and the quantitative analysis of test results. Graphical analyses serve several functions:

1. They provide a visual perspective that can triangulate or cross-validate traditional quantitative item, test, and DIF analyses.
2. They help measurement specialists gain a better conceptual understanding of the principles of measurement as they apply to a test.
3. They strengthen the link between quantitative analyses and substantive interpretations of what a test is measuring and how well it measures.
4. They can be used to establish a "feedback loop" so that information gained from each administration can be incorporated into improving subsequent test construction procedures. The feedback loop also provides insight into the merit of potential future program considerations (e.g., adaptive testing).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 4, December 1996, pp. 311–329
© Copyright 1996 Applied Psychological Measurement Inc.
0146-6216/96/040311-19$2.20

Assessing Dimensionality and Estimating Parameters

The multidimensionality of a test is difficult to determine and always subject to interpretation. One common procedure is to construct a scree plot of the eigenvalues obtained from a principal axis factor analysis of the interitem tetrachoric correlation matrix (Reckase, 1979). The problem with this approach is deciding whether the second (and possibly third) eigenvalues are large enough to represent significant dimensions (also known as primary dimensions) or whether they characterize random noise in the measurement process. Horn (1965) and Drasgow & Lissak (1983) suggested that interpretation could be enhanced by comparing the scree plot created from real data to a scree plot created from a factor analysis of randomly generated test data containing the same number of items.

Another approach used to assess dimensionality involves evaluating the assumption of local independence using the covariance matrix for examinees within different intervals on the trait scale (McDonald, 1981; Roznowski, Tucker, & Humphreys, 1991; Stout, 1987). One outgrowth of Stout's research is a computer program called DIMTEST (Stout, 1987) that allows practitioners to use large sample theory and apply a statistical test to determine whether one cluster of items is dimensionally distinct from another. Recently, Kim & Stout (1994) developed a statistic based on conditional covariances that has been successful in identifying the correct dimensionality of generated test data with different correlational structures between the trait dimensions.

Once the number of dominant dimensions has been confirmed both statistically and substantively, practitioners can use one of several multidimensional item response theory (MIRT) programs, such as NOHARM (Fraser & McDonald, 1988) or TESTFACT (Wilson, Wood, & Gibbons, 1987) to estimate MIRT item parameters. When estimating multidimensional item or trait parameters, practitioners must carefully identify the correct (and interpretable) orientation, metric, and covariance structure of the latent trait space. NOHARM estimates the covariance structure among the latent dimensions; TESTFACT assumes that the covariance matrix is the identity matrix. Thus, when representing examinees and/or items in the estimated latent space, practitioners need to make sure that the estimated parameters have been scaled to the same metric and orientation. Davey & Oshima (1994) discussed different methods by which estimates from separate multidimensional calibration runs can be placed on the same metric. For example, two-dimensional item parameter estimates from two "parallel" forms are not directly comparable if the estimated correlation between dimensions for one group is .5 and for another is .2.

Graphical Representation of Two-Dimensional Test Data

This paper focuses on graphically representing item and test analyses that are performed on a set of test response data for which it has been confirmed that there are two dominant traits being measured. Thus, all of the analyses that are discussed here assume that the dimensionality for a given test has been thoroughly investigated and that two-dimensional item response theory (IRT) item parameters have been estimated. All plots were created using the graphics package DISSPLA (Computer Associates International, Inc., 1989). For didactic purposes, it was assumed that the "pure" dimensions that compose θ_1 and θ_2 were uncorrelated; consequently, items and examinees can be represented using a Cartesian coordinate system. However, in reality it is more likely that items actually measure composites or combinations of the pure traits and that in any given population of examinees the dominant dimensions will be correlated.

Graphically, there are three characteristics of the unidimensional three-parameter IRT model that can be depicted using item response functions (IRFs): difficulty, discrimination, and guessing. Extending these concepts to the multidimensional case becomes complicated, partly because two-dimensional IRFs are actually item response surfaces (IRSs), but also because within a multidimensional framework an additional attribute must be considered for each item—the composite of multiple traits the item measures best. Within a unidimensional framework, an item discriminates, to varying degrees, between all levels of the underlying trait, although there is a range in which the discrimination is optimal. Multidimensionally, an item has the capability of distinguishing between levels of many composites, but optimally between levels of just one composite trait. The goal is to identify this composite.

MIRT Models

Basically, there are two main types of MIRT models that are used to describe dichotomously scored item response data—the compensatory model and the noncompensatory model. The probability of a correct response to item i for the compensatory model can be expressed as

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\frac{\exp[1.7\mathbf{a}_i'\theta_j + d_i]}{1.0 + \exp[1.7\mathbf{a}_i'\theta_j + d_i]}, \qquad (1)$$

where

X_i is the score (0, 1) on item i ($i = 1, ..., n$),
$\mathbf{a}_i' = (a_{i1}, a_{i2})$ is the vector of item discrimination parameters,
d_i is the scalar difficulty parameter for item i,
c_i is the scalar guessing parameter for item i, and
$\boldsymbol{\theta}_j' = (\theta_1, \theta_2)$ is the vector of trait parameters for person j ($j = 1, ..., N$).

In this formulation, there is an item discrimination parameter for each dimension being modeled but only one overall item difficulty parameter. Because the terms in the exponent (i.e., the logit) are additive, being low on one trait can be compensated for by being high on the other trait.

A second type of multidimensional model is called the noncompensatory model. This model was developed by Sympson (1978). In the noncompensatory model, the probability of a correct response is expressed as

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\prod_{k=1}^{K}\frac{\exp[1.7\mathbf{a}_{ik}(\theta_j - b_{ik})]}{1.0 + \exp[1.7\mathbf{a}_{ik}(\theta_j - b_{ik})]}, \qquad (2)$$

where $\mathbf{b}_{ik} = (b_{i1}, b_{i2}, ..., b_{ik})$ is the vector of difficulty parameters for item i on dimension k and the remaining terms are as defined for Equation 1.

In the noncompensatory model (Equation 2), there is a discrimination and difficulty parameter for each dimension. Because the terms in the model are multiplicative, the overall probability of correct response is bounded by the smaller probability for any one dimension. Having high trait level on one dimension does not compensate for having low trait level on another dimension. Currently, there is no software to estimate parameters for the noncompensatory model; consequently, representations in this paper are based on the compensatory two-dimensional MIRT model. Debate as to whether items should be modeled using one model versus the other is important and needs to be pursued. For a more in-depth comparison of the two models and problems encountered in parameter estimation for these models, see Ackerman (1989), Lim (1993), and Hsu (1995). Note that the characteristics being modeled should always be examined in conjunction with the substantive and contextual features of the individual items. In fact, this is the only way that the meaning of the dimensional structure can be correctly interpreted.

Different graphical representations designed to supplement traditional two-dimensional item and test analyses are presented below. Although individually each analysis may answer a different question about what the test measures, cross-validation between analyses also needs to addressed.

Data

Using NOHARM, a two-dimensional estimation of item parameters for two American College Testing

(ACT) Mathematics Usage Tests—Forms B and C—was conducted (American College Testing Program, 1989). Because NOHARM (Fraser & McDonald, 1988) estimates parameters for the two-dimensional normal ogive model, the parameter estimates were rescaled (see Hambleton & Swaminathan, 1985, p. 37) to the logistic model (Equation 1) with uncorrelated dimensions.

The graphical representations are divided into five broad categories: trait estimation representation, item representation, information representation, conditional analyses, and expected score distribution. Each analysis is prompted by a question about the measurement process.

Representation of Trait Estimates

One logical use of two-dimensional item parameter estimates is to obtain trait estimates for examinees or identified groups of examinees in order to compare their distributions. For example, how disparate are the trait distributions for black and white examinees? Whether the difference leads to DIF could be answered by a comparison of their respective trait distributions. Certain examinee response vectors may yield anomalous two-dimensional trait estimates. If item parameters are known, traits can be estimated using a Newton-Raphson iterative procedure to find the maximum value of the log-likelihood surface. The maximum likelihood estimation (MLE) procedure for trait estimation is graphically represented in Figure 1.

The log likelihood surface, the corresponding contour of the surface, and the location of the intermediate trait estimates for each iterative step of the MLE procedure for a specified set of item parameters and given response vector are represented in Figure 1. The location of the first (1) and final (3) trait estimates are indicated above the surface and on the contour. The final θ estimates were $\hat{\theta}_{j1} = 1.84$ and $\hat{\theta}_{j2} = -.66$ (these are also provided in Figure 1). In this particular example, convergence (change in either the θ_1, θ_2 estimate was less than .0001) was achieved in five iterations. $\hat{\theta}$s that do not reach convergence or that have a large standard error of measurement could be investigated from a graphical perspective to determine whether the estimation process is hindered by a "rough" likelihood surface that could yield several local maxima.

Representation of Items

Graphically, the probability of a correct response given two traits can be depicted using the IRS. A single item can be represented by constructing the IRS, constructing a contour of the IRS, or using vector notation (Reckase & McKinley, 1991).

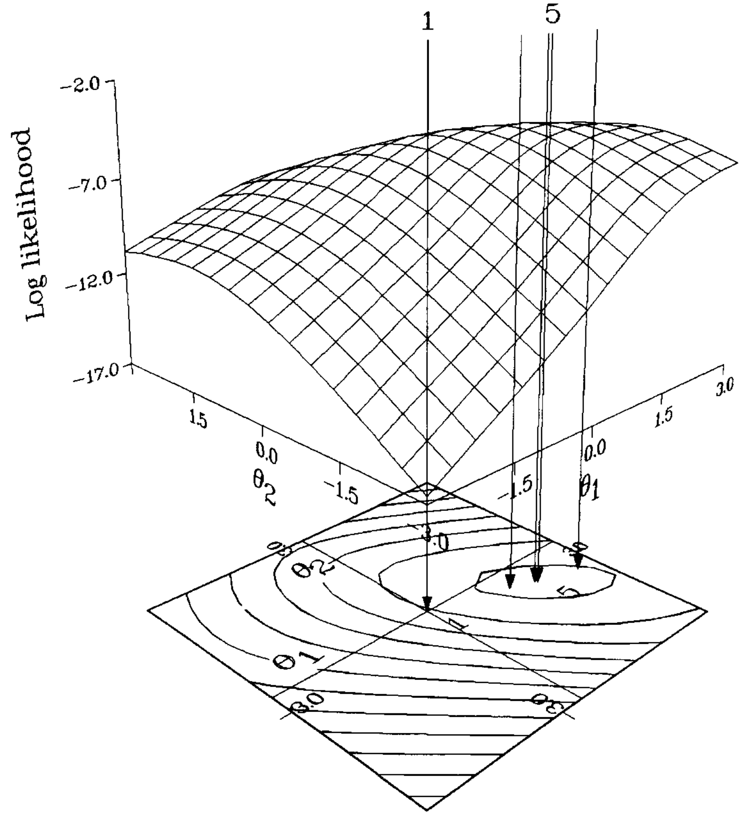
An IRS for a single item is shown in Figure 2. Figure 3 shows the contour of the IRS for the same item. The contour lines are equiprobability contours. For the particular multidimensional model described in Equation 1, these contours will always be parallel. Examinees whose θ_1, θ_2 places them on the same contour all have the same probability of responding correctly to the item. Contour plots are more informative than IRS plots because the following features of the item can be noted:

1. The trait composite the item is best measuring (i.e., the composite direction orthogonal to the equiprobability contours as indicated by the arrow in Figure 3);
2. Where in the trait plane the item is most effective in distinguishing between trait levels (i.e., where it is most discriminating). The greater the slope of the surface, the more discriminating the item and hence, the closer together the equiprobability contours;
3. The difficulty of the item. Assuming a particular underlying bivariate trait distribution, the proportion of examinees that would be expected to correctly answer the item can be estimated.

A drawback of contour plots is that, as with surface plots, only one item can be represented at a time.

A directional vector plot is another way to represent an item (see Figure 4). Thus, the final way to represent items in a two-dimensional trait plane is to use the vector notation developed by Reckase & McKinley (1991). When represented as a vector, the overall discrimination of the item is denoted by the value of MDISC:

Figure 1
Illustration of Newton-Raphson MLE Trait Estimation in a Two-Dimensional Trait Space



$$\mathrm{MDISC} = \left(a_1^2 + a_2^2\right)^{1/2}, \tag{3}$$

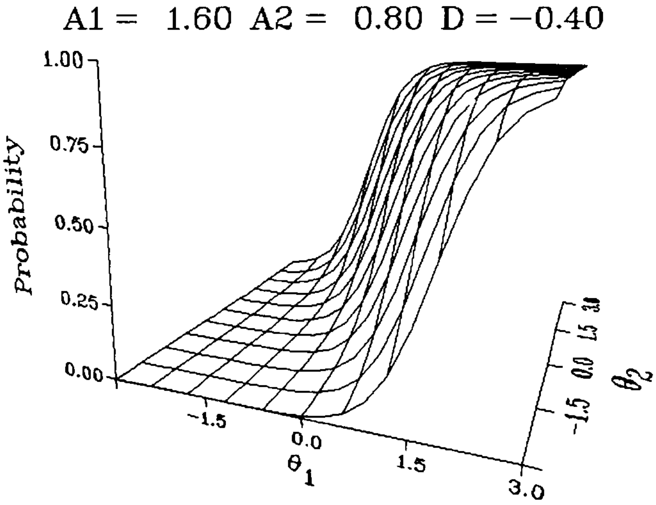
where a_1 and a_2 are the discrimination parameters for Dimensions 1 and 2, respectively.

The amount of discrimination is indicated by the length of the vector. All vectors, if extended, would pass through the origin of the latent trait plane. Because the as are constrained to be positive, vectors are located in only the first and third quadrants. The angular direction of the vector from the θ_1 axis represents the composite θ_1, θ_2 that is being most accurately measured. That is, the vector representing an item will always lie orthogonal to the equiprobability contours or in the direction in which the slope of the IRS is steepest. The difficulty of the item is denoted by the location of the vector in the space. That is, the tail of the vector lies on the $p = .5$ equiprobability contour, where p is the probability of correct response as given in Equation 1. Thus, easy items will lie in the third quadrant (such as Item 1 in Figure 4), and difficult items will lie in the first quadrant (such as Items 2 and 3 in Figure 4).

One additional feature that has been added to Reckase & McKinley's (1991) notation is to indicate the size of c. Items with a c estimate ($\hat{c}$) less than .1 are represented by an open-tipped arrowhead (e.g., Item 3); items with $\hat{c}$ between .1 and .2 are represented by a closed arrowhead (Item 1); and items with a large $\hat{c}$ (i.e., greater than .2) are represented by a solid arrowhead (Item 2).

Vectors for three items are shown in Figure 4. Item 1 (a_1 =.4, a_2 = .6, d = 2.4, c = .15) is an easy item, is

Figure 2
An Item Response Surface



moderately discriminating, and measures best in approximately a 56° direction (i.e., the item measures primarily θ_2). Item 2 ($a_1 = 1.8$, $a_2 = .2$, $d = -1.5$, $c = .25$) is a more difficult item, is highly discriminating, has a relatively large guessing parameter, and measures best in a 6° direction (i.e., the item measures predominantly θ_1). The vector for Item 3 corresponds to the item represented in Figures 2 and 3.

Thus, several questions can quickly be answered using the vector plot, including:

Figure 3
A Contour Plot of an IRS

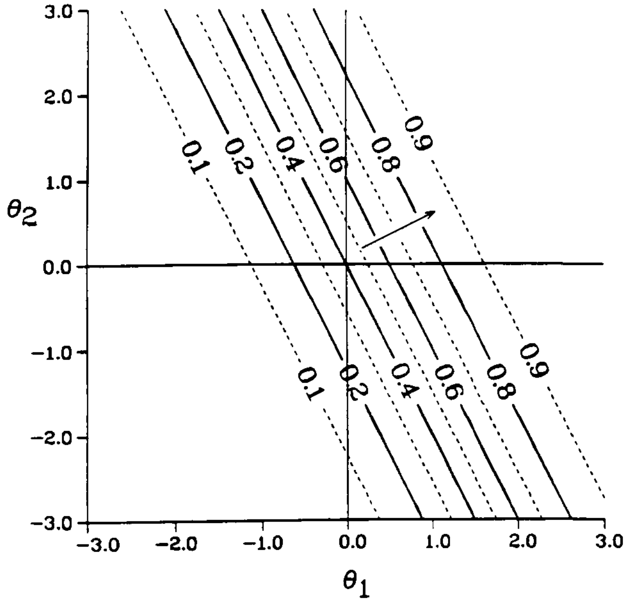
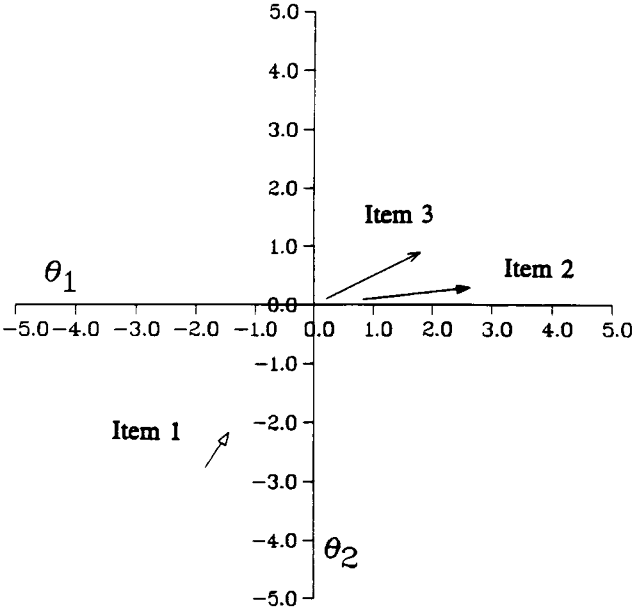


Figure 4
Vector Representation of Two-Dimensional Items



1. Do certain composites or certain content areas have more discriminating items?
2. Is item difficulty related to item content or the trait composite being assessed?
3. Is guessing a function of difficulty, or content, or a composite?
4. Do the individual content areas map onto unique or overlapping sectors of the latent space?
5. How similar are the item vector plot profiles for two or more parallel forms?

Vector plots also can be used to supplement DIF analyses. It has been argued (Roussos & Stout, 1996) that one of the underlying factors that contributes to DIF is the multidimensional nature of a test. Items that appear to be measuring composites quite distinct from those items deemed to be most valid are prime candidates to test for DIF. Ackerman (1992) discussed this situation in terms of a validity sector.

Representation of Information

In IRT, the accuracy of the measurement process is discussed in terms of "information," which is inversely related to the standard error of measurement. Thus, the smaller the standard error of measurement, the greater the information and, hence, the greater the measurement precision of the test.

Ackerman (1992) developed procedures to compute the multidimensional information function for the MIRT model given in Equation 1, taking into account the lack of local independence once a direction in the latent space is specified. Ackerman (1992) determined that the amount of information provided by an item i at a specified θ_1, θ_2 in the direction α can be computed as

$$I_i(\theta_1,\theta_2) = (\cos\alpha)^2 \mathrm{var}\left(\hat{\theta}_1 \middle| \theta_1,\theta_2\right) + (\sin\alpha)\mathrm{var}\left(\hat{\theta}_2 \middle| \theta_1,\theta_2\right) + 2(\sin 2\alpha)\mathrm{cov}\left(\hat{\theta}_1,\hat{\theta}_2 \middle| \theta_1,\theta_2\right), \quad (4)$$

where var denotes variance and cov denotes covariance.

Because items are capable of distinguishing between levels of traits for different θ_1, θ_2 composites, several questions arise:

1. What composite of traits is being best measured by each subtest as well as by the total test?
2. Is the same composite of traits being most accurately measured throughout the two-dimensional trait plane?
3. Does the number-correct (NC) score scale have the same meaning or interpretation throughout the observable range of scores?
4. Which subtests provide the greatest measurement precision and for which trait composites?
5. Is the same degree of measurement precision for various composite traits displayed across forms?

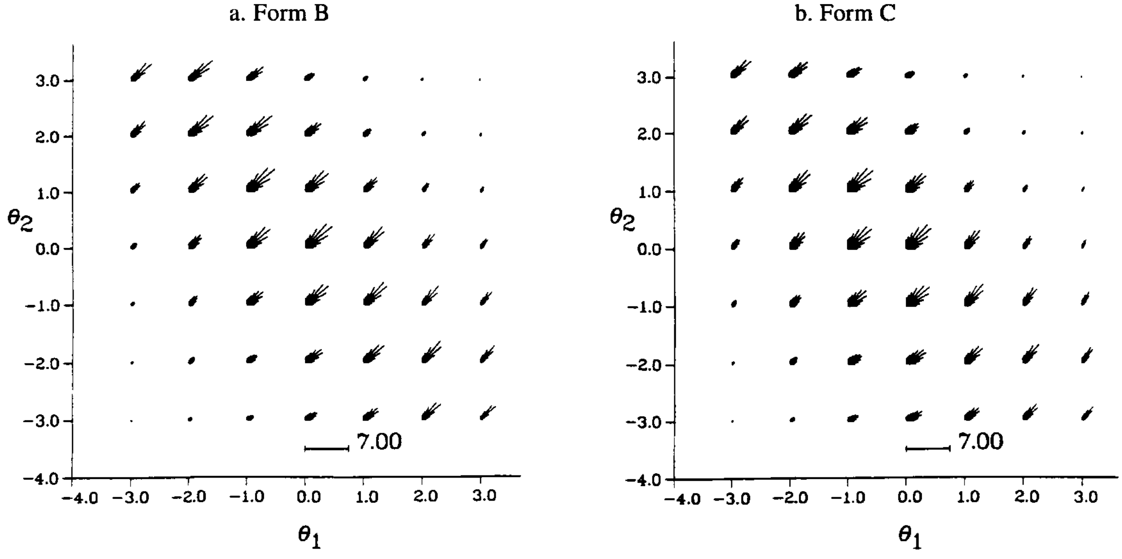
Clamshell plots. An interesting way to graphically display information was developed by Reckase & McKinley (1991). Referred to as clamshell plots because of their shape, these plots denote the amount of information at selected points in several different directions. Specifically, a clamshell plot is created by computing the amount of information (using Equation 4) at selected θ_1, θ_2 points. The amount of information that the test provides for 10 different composites or measurement directions (from 0° to 90° in 10° increments) is computed for each point and represented by a vector originating from the specified point. The result of the 10 vectors at each two-dimensional trait point resembles a "clamshell."

The clamshell plots for each of the two ACT Mathematics Usage forms are displayed in Figures 5a and 5b based on estimated item parameters using a 7 × 7 grid. The two clamshell plots are very similar; thus, all composites were measured approximately equally well. The 40° to 50° direction was measured most accurately for both forms. Relatively little information was provided for examinees who were high on both traits or who were low on both traits.

Such plots would be interesting to use in conjunction with cutscores and expected NC score (ENC) surface contours (shown in Figure 8 below). The precision of a test should be greatest in regions around such decision points. Clamshell plots not only indicate the relative precision about such cutscores, but also indicate which composite of traits is being assessed (Ackerman, 1994a, 1994b).

Information directional plots. Although clamshell plots help to ascertain the amount of information and the composite that is being best measured, it may also be of interest to know what direction is being best measured overall at each of the 49 points. This direction can be found by examining all composites from 0° to

Figure 5
Clamshell Plots With Test Information Vectors for 10 Different Trait Composites
From 0° to 90° in 10° Increments at 49 Selected θ_1,θ_2 Points



90° in 1° increments and noting the direction with the largest value. This process would be repeated at each of the selected points. One way to display these results is shown in Figure 6. At each selected point is the number denoting the angular direction of the θ_1, θ_2 composite that is being measured most precisely. The font size of the number indicates the relative amount of information in the specified direction: the larger the font, the greater the measurement precision. The degree of similarity between Forms B (Figure 6a) and C (Figure 6b) suggests that they are measuring similar trait composites. Samejima (1977) referred to tests that have similar information profiles as "weakly parallel" tests. Note also that the numbers coincide with the plots in Figure 5.

Directional composite correlations. For ease of interpretation, practitioners usually report test results in terms of NC scores or some transformation of NC scores such as standard scores. Thus, there is a need to determine the relationship between the NC score scale and the two-dimensional latent trait plane. One way to examine this relationship is to compute the linear weighting of θ_1 and θ_2 that provides the greatest relationship or correlation with the NC scale.

To investigate this relationship, the following procedure was performed. First, a group of examinees was randomly generated from a bivariate normal distribution. The latent plane was then divided into eight "wedges" or octants, numbered counterclockwise beginning from the positive θ_1 axis. Using the generated response data for the examinees in each octant, the angular direction or θ_1, θ_2 composite that had the greatest correlation with the observed score was calculated. The results are displayed in Figures 7a and 7b. The number indicating the angular direction of the maximum correlation is given in each octant. The font size of the number indicates the magnitude of the linear relationship (r^2): the larger the number, the stronger the relationship or correlation.

Two important features should be noted in Figure 7. First, it is important that the observed score scale has a consistent interpretation throughout the observable range. Noticeable differences occur between the correlations for the first and second octants. Another way in which to extract meaning from Figures 7a and 7b is to ask the following question: If examinees were to attempt to raise their observed score, which

Figure 6
Angular Composites of Maximum Information at 49 Selected θ_1, θ_2 Points

a. Form B

b. Form C

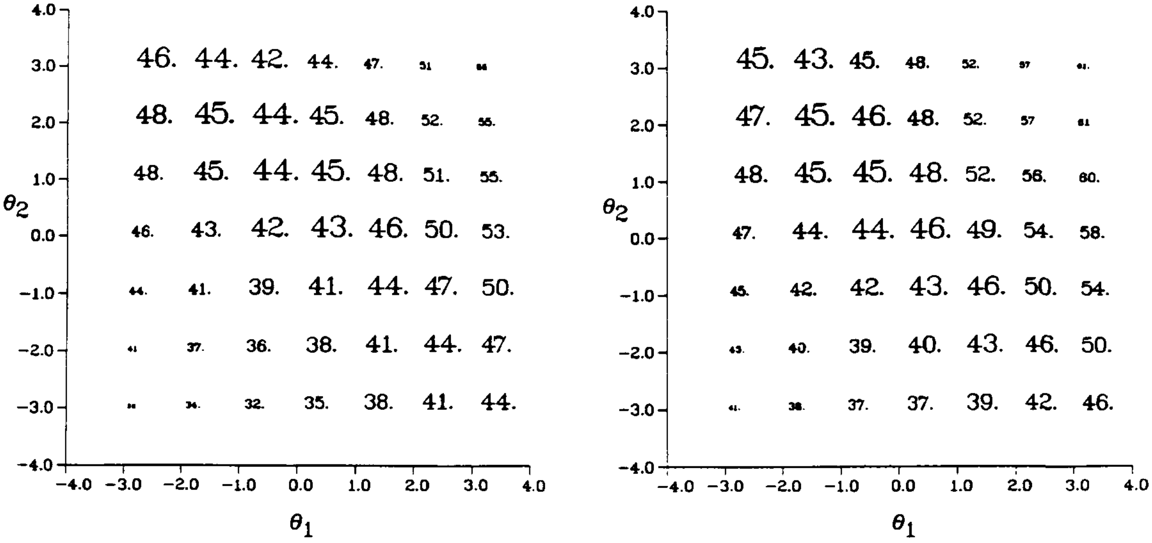
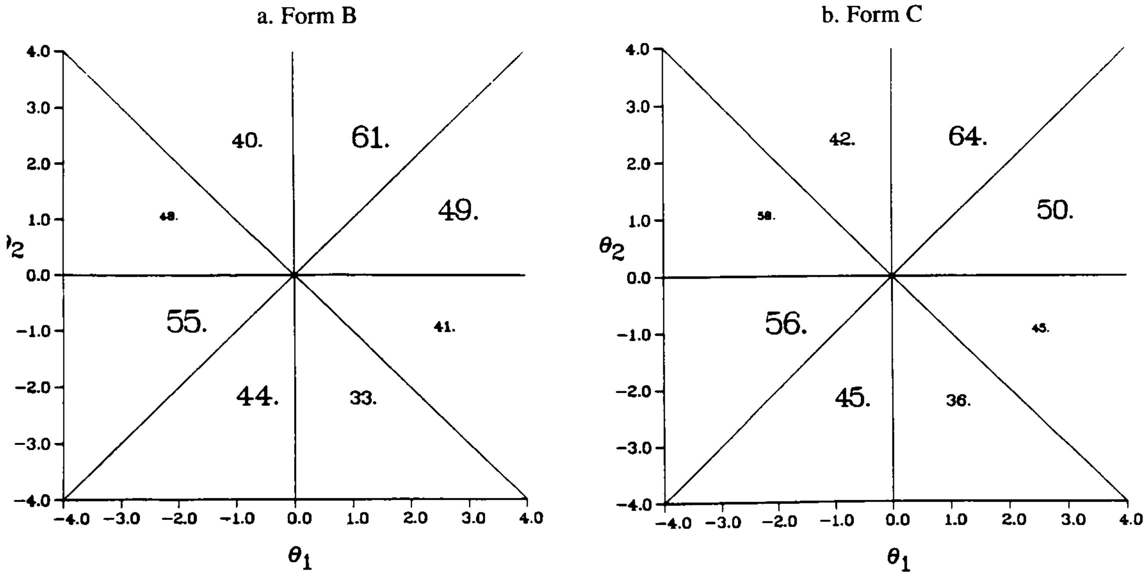


Figure 7
Angular Composite Values Indicating the Largest Correlation With NC Score for Generated Examinees in Eight Regions of the Two-Dimensional Trait Plane


a. Form B

b. Form C

composite of traits would they have to improve? Ideally, the composite should be the same throughout the observable score range and throughout the latent trait space.

A second important feature is to compare results across forms. If different forms are truly parallel, the relationship between the underlying traits and the NC score will be the same. That is, the meaning imparted to the NC score scale should be constant within and between forms. For the two test forms used here, there appears to be a "consistent" pattern between θ_1, θ_2 and NC, with the largest angular composite difference within forms occurring in the second and seventh octants for each form. For Form B the correlations were .61 and .33, respectively.

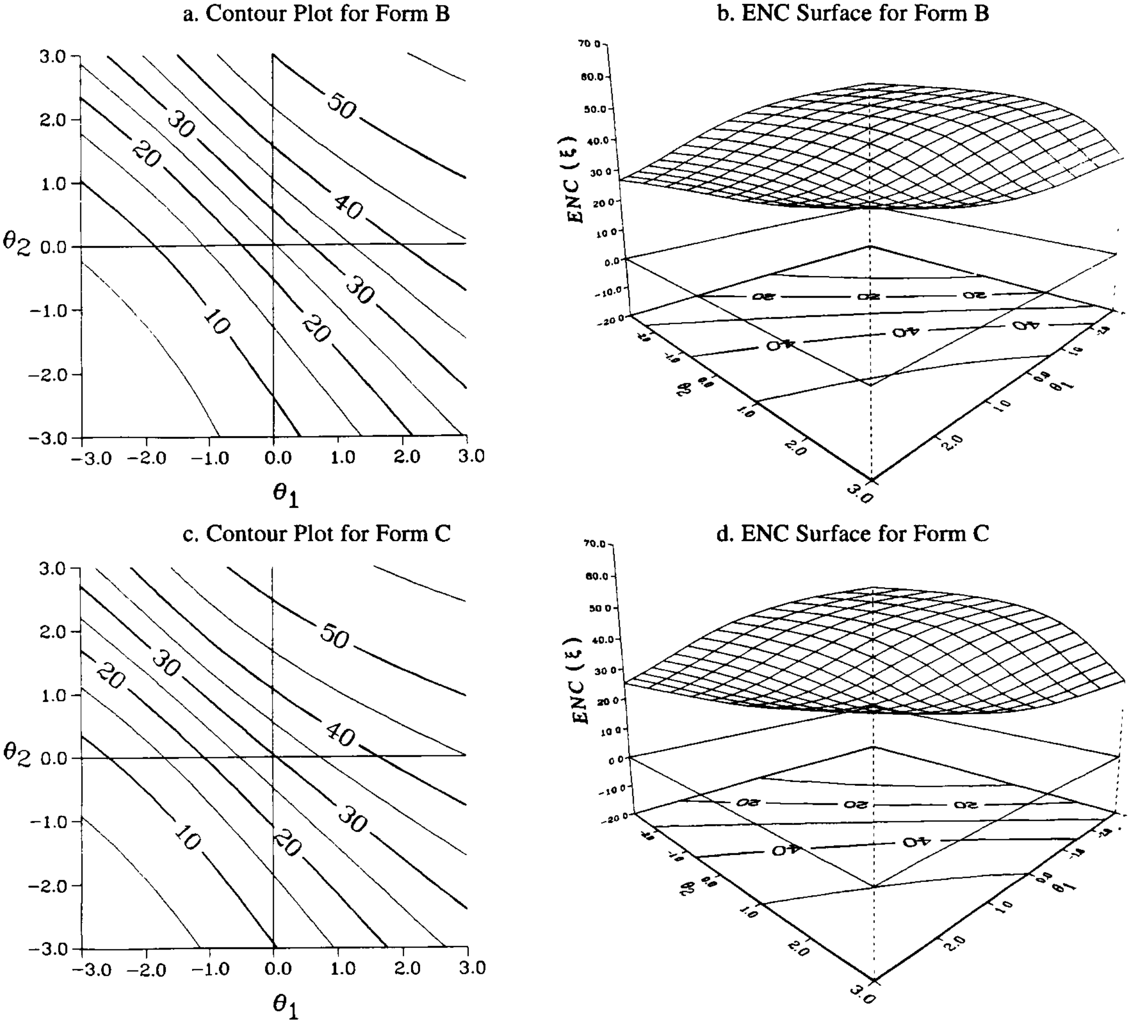
ENC surface plots. Another question that relates to measurement precision is: What is the ENC for each examinee? This can be graphically determined by creating a plot of the ENC surface plot. The ENC, ξ, for any examinee is simply the sum of the probabilities of a correct response, as given in Equation 1, for each of the n items in the test,

$$\xi = \sum_{i=1}^{n} P_i(\theta_1, \theta_2). \qquad (5)$$

To generate the ENC surface, ξ is computed for examinees at selected θ_1, θ_2 points (a 31 × 31 grid is generally used) over the trait plane. Surface and corresponding contour plots are then created. The contours provide insight into how the latent trait plane is mapped onto the ENC scale. Note that the contour lines do not have to be parallel and that some curvilinearity may occur if there is a great deal of heterogeneity in the composites that the test items are measuring. The ENC surfaces for the two ACT forms are displayed in Figures 8b and 8d, and their corresponding contours are displayed in Figures 8a and 8c.

The contour of the ENC surface plot can be directly related to the clamshell plots; that is, the longest vectors in the clamshell plot are orthogonal to the ENC contour lines. Likewise, regions that contain the longest vectors in the clamshell plot (i.e., the regions in which the measurement precision is the greatest)

Figure 8
Two-Dimensional ENC Surface and Corresponding Contour Plots



a. Contour Plot for Form B

b. ENC Surface for Form B

c. Contour Plot for Form C

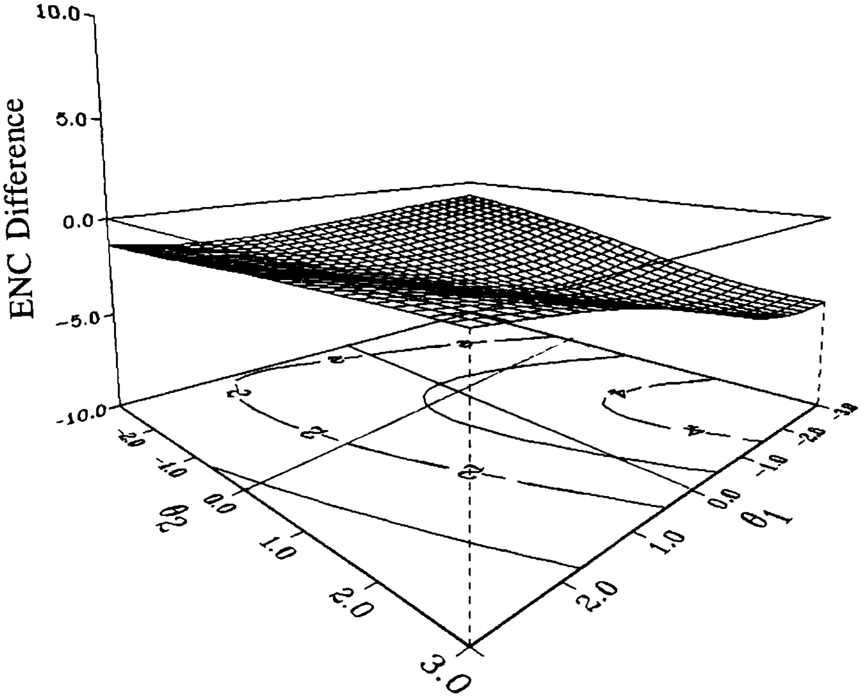
d. ENC Surface for Form C

should be the same regions in which the ENC contours are closest together (indicating regions in which the test is doing a better job at distinguishing between levels of ENC).

Plot of the difference between two ENC surfaces. After computing the ENC surface for each form, the next logical step is to compare them. If two test forms are truly parallel, any examinee would be expected to have the same ENC no matter which form he or she was administered. This implies that if two ACT forms are parallel, their ENC surfaces should be very similar. One way to analyze the degree of parallelism is to graph the surface representing the difference between the two ENC surfaces. An example of this type of plot—a surface illustrating the ENC values for Form B minus the ENC values for Form C—is shown in Figure 9.

The *zero plane* is outlined on the base of the cube in Figure 9. If there were no difference between the ENC surfaces (i.e., the forms were strictly parallel), the difference surface would lie in this zero plane. Regions in which the difference surface lies above the zero plane indicate regions in which examinees

Figure 9
Surface Representing the Difference Between ENCs for Form B Minus ENCs for Form C



would have a higher ENC on Form B; if the difference surface dipped below the zero plane, this would indicate regions in which examinees would have a higher ENC on Form C. Where the majority of examinees would lie (i.e., in the first and third quadrants), the maximum difference indicated by the contour is approximately two to three ENC points. This is a typical difference of 1 standard deviation (American College Testing Program, 1989, p. 47) for the two 60-item tests before equating.

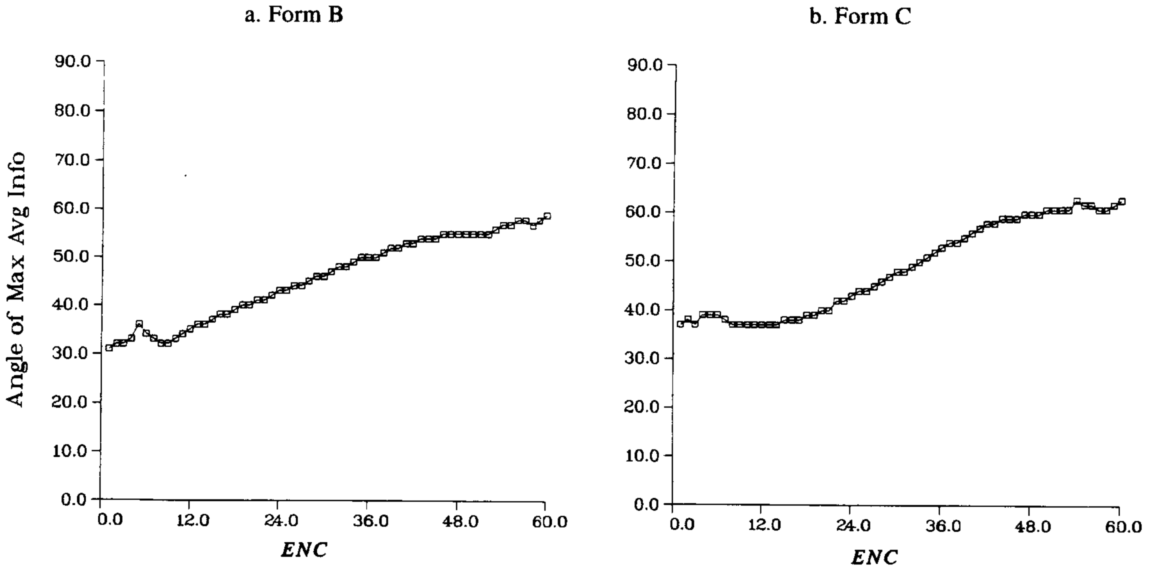
If pretest sample sizes permitted accurate MIRT calibration of the item parameters (e.g., $N = 2{,}500$), this analysis could be conducted as part of the test construction phase—before any forms were administered—to insure that the test forms were multidimensionally parallel.

ENC composite directions. Another question that remains concerns the EC surface. Any curvilinearity in the ENC contours implies that differences between levels of ENC may not have the same meaning (in terms of the θ_1, θ_2 composites) throughout all regions of the trait plane. What composite of θ_1, θ_2 is being best measured for the examinees at each ENC level?

To investigate this issue further, 30 different θ_1, θ_2 combinations, each having the same ENC, were systematically located for each possible ENC. At each of the 30 trait points, the direction of maximum information was computed. An average direction, weighted by the density of the examinees at each point, was then calculated. This process was repeated for each possible ENC. The results for each of the two ACT forms are displayed in Figures 10a and 10b.

Notice that there is a slight confounding of difficulty and dimensionality. That is, the direction corresponding to the maximum average information changes as ENC increases. For low ENC (e.g., 0–15) the most informative composite was slightly under 40°. However, for high ENC (e.g., 50–60) the most informative composite was near 60°. This pattern was evident for both forms.

Figure 10
Angle of Average Maximum Information for Each Possible ENC



Conditional Analyses

Another set of analyses focuses on measurement questions such as: (1) what is the mean θ_1 and mean θ_2 (or centroid) of the examinees who would be expected to achieve a particular score? and (2) how variable (in terms of their θ_1 values and θ_2 values) is the distribution of examinees who would be expected to achieve a particular score? These questions are also central to the issue of score scale consistency and have strong implications for issues related to equating and DIF analyses.

To answer the first question, the conditional bivariate distribution

$$h(\theta_1,\theta_2|X=x), \quad 0 \le x \le n \tag{6}$$

for each possible NC score, x, is computed. Assuming a bivariate normal distribution of traits, $f(\theta_1,\theta_2)$, and using the estimated IRF $P_i(\theta_1,\theta_2)$ given in Equation 1, the conditional trait distributions, $h(\theta_1, \theta_2|x)$ can be estimated by the Bayesian formula

$$h(\theta_1,\theta_2|x) = \frac{P(\boldsymbol{\Theta}=\boldsymbol{\theta}\text{ and }X=x)}{P(X=x)} = \frac{P(X=x|\boldsymbol{\Theta}=\boldsymbol{\theta})f(\theta_1,\theta_2)}{\iint P(X=x|\boldsymbol{\Theta}=\boldsymbol{\theta})f(\theta_1,\theta_2)d(\theta_1)d(\theta_2)}, \tag{7}$$

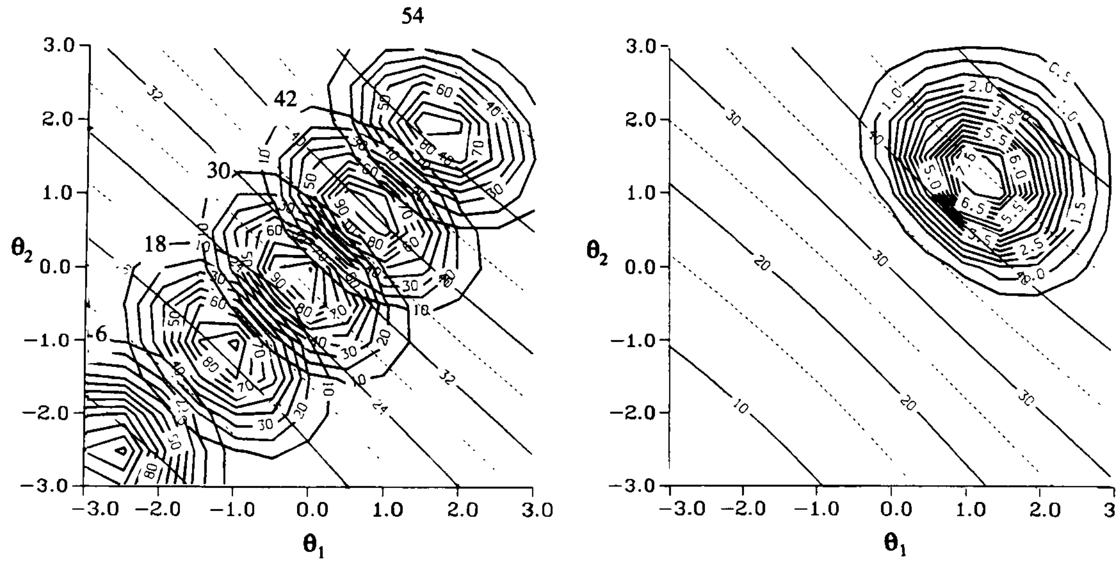
where $\boldsymbol{\Theta}$ represents the random variable trait vector and d indicates the derivative. Using a recursive formula developed by Stocking & Lord (1983), the $P(X=x|\boldsymbol{\Theta}=\boldsymbol{\theta})$ were computed for a grid of (θ_1,θ_2) values, and $h(\theta_1, \theta_2|x)$ was estimated using Equation 7.

ENC distribution contours. Several different types of plots can be created from this information. One such plot, shown in Figure 11a, illustrates the different conditional expected distributions of θ_1,θ_2 for selected score categories. This plot was created using the two-dimensional item parameters for Form B with an underlying bivariate normal distribution centered at (0,0), unit variance, and a θ_1,θ_2 correlation of .5. Contours of the θ_1,θ_2 distributions for the five score categories—6, 18, 30, 42, and 54—are superimposed on the contour of the ENC surface. The values among the density contours represent the relative frequency value multiplied by 1,000. Plots like this could be used to evaluate the degree of parallelism between forms

Figure 11
Contours of the Conditional θ_1, θ_2 Distributions From Form B

a. For NC Scores of 6, 18, 30, 42, and 54

b. For Examinees Expected to Score Above 40



by enabling the test developers to examine whether the conditional distributions for the possible score categories are similar for each test form.

It is also possible to plot a contour of the density distribution of examinees who would be expected to score above a particular "cutscore" for a test. This type of plot would give practitioners an idea about where in the trait space examinees would lie who would be expected to achieve a specified passing score. For example, in Figure 11b, the conditional distribution for the group of examinees who would be expected to score 40 or higher on the 60-item Form B is plotted along with the contour of the ENC surface. These plots can be created for any underlying bivariate normal distribution.

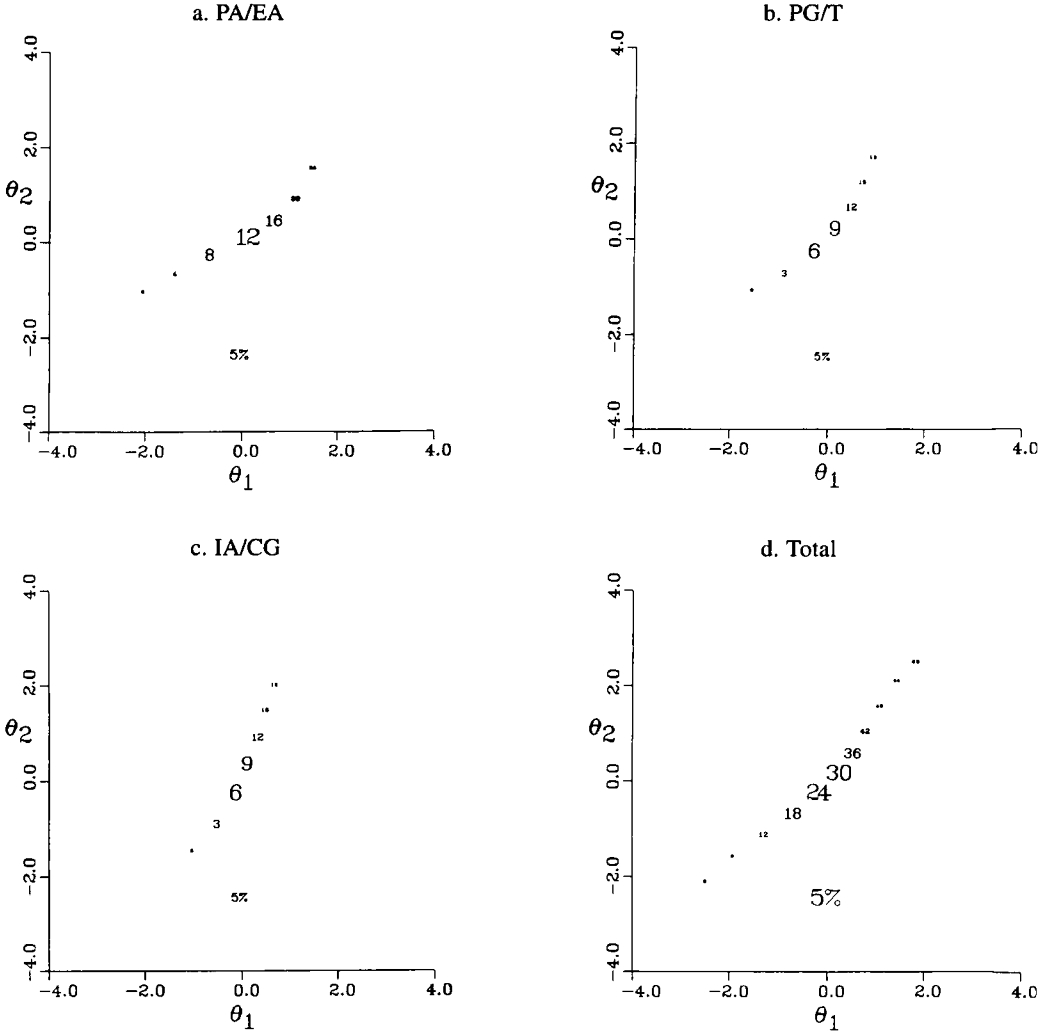
ENC centroids. A simplified version of this type of graph can be created by plotting just the centroids,

$$(\bar{\theta}_1,\bar{\theta}_2)=\mathcal{E}\left[h(\theta_1,\theta_2)|X=x\right] \tag{8}$$

of the conditional distributions. The results from each ACT form are plotted in Figures 12 and 13 for each of the three subtests—pre-algebra and elementary algebra (PA/EA), intermediate algebra and coordinate geometry (IA/CG), and plane geometry and trigonometry (PG/T)—and for the total test.

The numbers plotted in Figure 12 indicate the particular score category. The location of the number is the position of the centroid for that score category. The size of the font used corresponds to the percent of examinees that would be expected to achieve that indicated score category. The font size indicating a relative frequency of 5% is indicated at the bottom of the graph. The differences between the various subtests are quite noticeable. Such analysis helps substantively to define θ_1 as representing algebraic symbol manipulation skill and θ_2 as a text translation skill. The PA/EA items appear to be measuring more of θ_1 except at the upper score levels. That is, the difference between an ENC of 4 and 8 or 8 and 12 represents a difference in primarily the θ_1 ability. Differences in IA/CG, and to some extent the PG/T items, represent a greater increase in θ_2 than in θ_1. Note that differences throughout the ENC scale represent different levels of the θ_2 ability. In all subtests, there appears to be a slight curvature to the plotted centroids, suggesting that

Figure 12
Conditional Centroids for Selected NC Score for the Three ACT Subtests and Total Test For Form B



a. PA/EA

b. PG/T

c. IA/CG

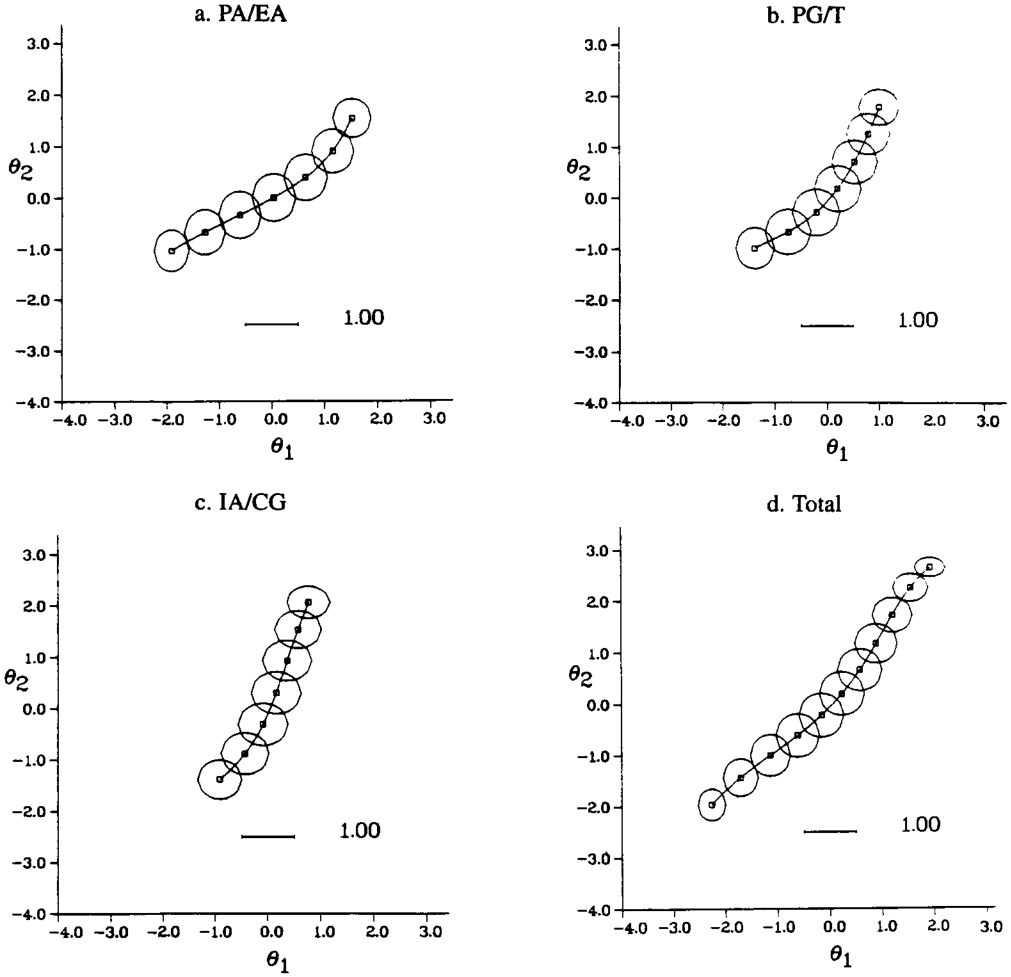
d. Total

differences between the low scores represent θ_1 differences, and that differences at the high score levels represent greater increases in θ_2 than in θ_1. The curvature is less pronounced for the total score categories, which seem to represent differences in equal composites of θ_1 and θ_2.

Such plots enable test practitioners to determine not only the degree of parallelism among forms, but also whether differences along the score scale can be interpreted as differences in the same trait composite within each form. Score scale consistency should be directly related to the intended uses of the test. It may not always be the case that the composite being best measured need be the same. For example, it could be argued that the composite skills needed in the score range required for college admissions should be different than the composite of skills needed to achieve a score in the range for a college scholarship.

Related to the plot of the centroids is another graph (not shown here) highlighting the second moments of the conditional distributions discussed with respect to Figure 12. For the NC score scale to have a consis-

Figure 13
Conditional Variance Ellipses for Selected NC Score Values of the Three ACT Subtests and Total Test For Form C



tent interpretation in terms of θ_1 and θ_2, not only would the centroids have to be linear but the conditional θ_1 and θ_2 for each expected score would have to be similar. For a given score category, the trait with the smaller variance represents the dimension being measured more accurately.

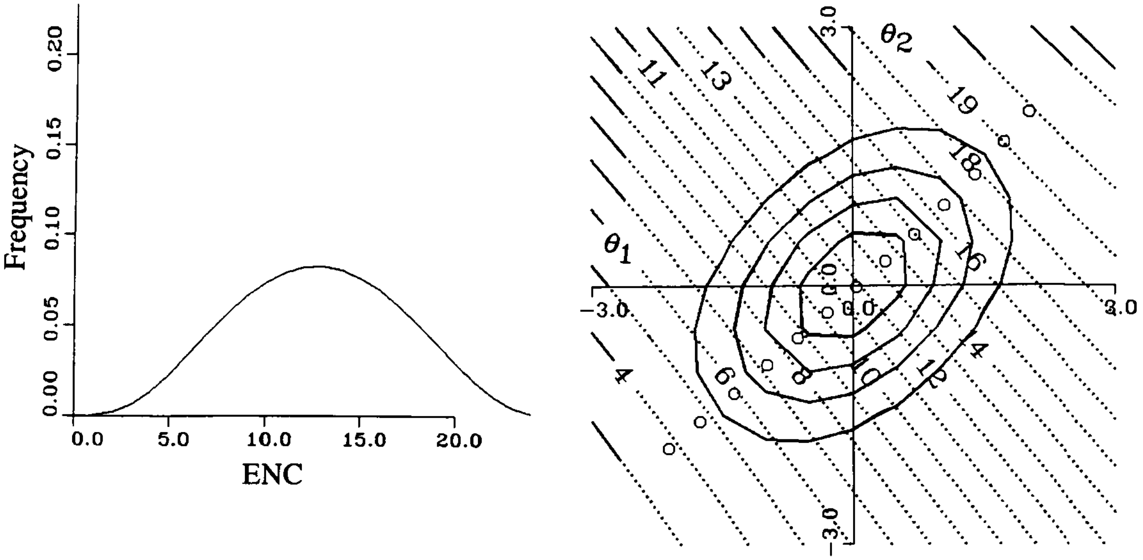
In the process of computing the conditional centroids, the conditional variances were also calculated. The size of these variances is represented as an ellipse. The length of the horizontal axis of the ellipse denotes the size of the θ_1 variance. The height of the vertical axis indicates the θ_2 variance. Thus, a circle would indicate that both traits are measured equally well for that expected score category. Note that each ellipse is centered about its corresponding centroid. For the sake of clarity, ellipses are drawn for only selected score categories. The number indicating the particular score category can be located at the centroid inside the ellipse. The conditional variance ellipses for the Mathematics Usage subtests and total test are illustrated in Figure 13 for Form B.

The vertically elongated ellipses in Figure 13a imply that these items measure θ_1 best. (This could be confirmed by constructing a plot like that shown in Figure 6 for each subtest individually.) A somewhat "opposite" case occurs with the PG/T and IA/CG items (Figures 13b and 13c) for which θ_2 is measured more accurately. When all three subtest types are combined (Figure 13d), it appears that low scores reflect a more accurate measure of θ_1, whereas θ_2 is being assessed more accurately for the high scores.

ENC Distribution

The final graphical analysis relates to the expected score distribution given a set of two-dimensional item parameters. This information, which is computed in conjunction with the centroid analysis, is illustrated in Figure 14. Figure 14a shows a relative frequency function of the ENC distribution and Figure 14b shows the contour of the underlying trait distribution and corresponding centroid. This information is important in order to evaluate the degree of parallelism between test forms before they are administered. In Figure 14a, the expected score distribution for the 24-item PA/EA subtest for Form B is illustrated. This type of graph would be useful to determine how parallel forms of pre-equated test items actually are and if they would have the same ENC score distribution. Figure 14b can be interpreted similarly to Figure 12. Ideally, the centroids for the ENC range would indicate that the same composite traits are being assessed across all forms. For example, the difference between an ENC of 5 and an ENC of 10 represents the same composite trait increase for each form.

Figure 14
ENC and Contour Plot of Specified Underlying Two-Dimensional Trait Distribution for the Form B PA/EA Subtest



a. ENC Distribution

b. Contour Plot

Discussion

There should always be a close tie between the test construction phase and the post-administration analyses. The bridge between statistical analyses and item writing is always an important one, but even more so here because of the issue of multidimensionality. The work of an item writer should not stop after a test has been administered. The work of psychometricians should not stop after the item analysis and equating have been completed. Together, both groups need to perform a post hoc analysis to consider such questions as: (1) what makes an item more discriminating or more difficult than others? and (2) what features of the item contribute to its potential to measure different trait composites? Once hypotheses about these issues have been developed, item writers should be challenged to create items and subsequently predict certain multidimensional characteristics. Through an iterative and cooperative effort, the goal should be to find the ideal magnification power provided by the multidimensional analytical lens in order to accurately detect the important attributes designated by the test specifications used to establish the construct validity of the test.

Future Directions

MIRT research is expanding in several directions. One direction is in computing vectors for distractors of multiple-choice items. This research seeks to determine the trait composites being measured by the various distractors. A second area concerns the plotting of examinees in the latent plane who are attracted to different distractors. Is there any relationship between examinees' estimated traits and their tendency to make different types of errors?

A third direction involves the use of collateral information. Collateral information is knowledge obtained about one trait by an item (or a collection of items) that are measuring a dimensionally different, yet correlated trait. For example, in an adaptive testing setting a unidimensional estimation of PA/EA level could also yield estimates of scores on the IA/CG and PG/T content areas. Thus, a subsequent adaptive test for IA/CG would begin with an estimated θ and would ideally require fewer items to obtain an estimate with a specified level of precision. A third adaptive test for PG/T would also begin with an estimated θ based on both PA/EA and IA/CG items, requiring even fewer items to complete the estimation process.

Finally, another important area this research needs to consider is how to graphically illustrate the various concepts and relationships beyond the two-dimensional solutions. For example, are there any graphical techniques that can be used to display response surfaces for a five-dimensional solution?

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory items. *Applied Psychological Measurement, 13,* 113–127.

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. *Journal of Educational Measurement, 24,* 67–91.

Ackerman, T. A. (1994a). Creating a test information profile in a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257–275.

Ackerman, T. A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255–278.

American College Testing Program. (1989). *ACT assessment.* Iowa City IA: Author.

Computer Associates International, Inc. (1989). *DISSPLA 10.0* [Computer software]. Garden City NY: Author.

Davey, T., & Oshima, T. C. (1994, April). *Linking multidimensional item calibrations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68,* 363–373.

Fraser, C., & McDonald, R. P. (1988). *NOHARM II: A FORTRAN program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer software]. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston MA: Kluwer Nijhoff.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Hsu, Y. (1995). Item parameter estimation of a two-dimensional generalized MIRT model (Doctoral dissertation, University of Illinois, 1995). *Dissertation Abstracts International, 57,* 1584.

Kim, H. R., & Stout, W. F. (1994, April). *A new index for assessing the amount of multidimensionality and/or simple structure present in test data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Lim, C. (1993). An application of the joint maximum likelihood estimation procedure to a two-dimensional case of Sympson's noncompensatory IRT model (Doctoral dissertation, University of Iowa, 1993). *Dissertation Abstracts International, 54,* 2549.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement, 15,* 109–127.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42,* 193–198.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Stout, W. F. (1987). A nonparametric approach to assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Sympson, J. B. (1978). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wilson, D., Wood, R., & Gibbons, R. (1987). *TESTFACT* [Computer program]. Mooresville IN: Scientific Software.

Author's Address

Send requests for reprints or further information to Terry Ackerman, Department of Educational Psychology, University of Illinois, 210 Education Building, 1310 S. Sixth Street, Champaign IL 61820-6990, U.S.A. Email: tackerma@ux1.cso.uiuc.edu.