Unidimensional Calibrations and Interpretations of Composite Traits for Multidimensional Tests

Richard M. Luecht and Timothy R. Miller
American College Testing

A two-stage process that considers the multidimensionality of tests under the framework of unidimensional item response theory (IRT) is described and evaluated. In the first stage, items are clustered in a multidimensional latent space with respect to their direction of maximum discrimination. The separate item clusters are subsequently calibrated using a unidimensional IRT model to provide item parameter and trait estimates for composite traits in the context of the multidimensional trait space. This application is proposed as a workable compromise to some of the estimation, indeterminacy, and interpretation problems that affect the direct use of multidimensional IRT procedures for item calibration and trait estimation. The findings of a study based on simulated multidimensional data indicate that there are identifiable gains in estimation robustness and score interpretation with almost no sacrifice in goodness-of-fit using this two-stage approach to modeling composite latent traits. *Index terms: item response theory, model fit, multidimensionality, parameter estimation; model fit; multidimensionality in IRT; parameter estimation; person fit; reference composites; trait estimation.*

In recent years, there has been a growing concern about the inherent multidimensional nature of tests. That concern arises from the perspective that the number and type of cognitive or psychological processes of examinees responding to test items may vary considerably (Ansley & Forsyth, 1985; Reckase, Carlson, Ackerman, & Spray, 1986). In reaction to that concern, a number of researchers have proposed multidimensional response-based models. These multidimensional models tend to follow the parametric forms of some of the more typical unidimensional item response theory (IRT) models. For example, Reckase (1985) and Reckase & McKinley (1991) proposed a logistic two-parameter multidimensional item response theory (MIRT) model that characterizes the probability of a correct response as

$$P(u_{ij} = 1 \mid \boldsymbol{\theta}_j; \mathbf{a}_i, d_i) = [1 + \exp(-\mathbf{a}_i^T \boldsymbol{\theta}_j + d_i)]^{-1} \quad , \tag{1}$$

where

d_i is a scalar location parameter representing the difficulty of the ith item;
$\mathbf{a}_i$ denotes a vector of item discriminations; and
$\boldsymbol{\theta}_j$ is a vector of latent traits for the jth examinee.

The order of these vectors corresponds to the number of dimensions in the model. Alternative models also have been proposed (Samejima, 1974; Sympson, 1978). The common conceptual framework for all of these multidimensional models is that the response surface may span $M \geq 1$ dimensions, because the items provide simultaneous or collateral information about one or more latent traits.

Most existing methods for estimating multidimensional item parameters and examinee traits use either nonlinear factor analysis or full information factor analysis applied to the covariances or

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 3, September 1992, pp. 279-293
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/030279-$2.00

tetrachoric correlations derived from a dichotomous data matrix (e.g., Bock & Aiken, 1981; Carlson, 1987; Fraser, 1986; McDonald, 1982; Muraki & Engelhard, 1985; Wilson, Wood, & Gibbons, 1984). However, despite the development of these MIRT calibration methods, there have been few practical applications of the theory or its methods.

This lack of MIRT applications arises from a variety of nontrivial problems. For example, there appears to be little consensus on any "best" method(s) for determining and interpreting the dimensionality of the trait space with respect to both psychometric and psychological criteria (e.g., Berger, 1990; Reckase et al., 1986; Stout, 1987, 1990; Way, 1990). This problem is complicated further when an exploratory approach is used to determine the structure of the latent space. Under exploratory factor analysis, the dimensionality problem is compounded by the need to resolve issues such as factor invariance and rotational indeterminacy (e.g., Gorsuch, 1983; McDonald, 1982).

Recent research has suggested that, for tests of less than 100 items, MIRT models may not be as capable of discriminating among examinee traits as unidimensional models (e.g., Davey & Hirsch, 1990). This problem is related, in part, to the increased parameterization of the MIRT models. That is, the increased dimensionality greatly complicates the identifiability and the estimability of the additional structural and incidental parameters (e.g., Holland, 1990; McDonald, 1982).

These types of problems have hindered the development of practical applications involving MIRT models. In many instances, multidimensionality is ignored completely in favor of applying a less complex unidimensional IRT model. However, this approach conceivably sacrifices information about the trait levels of examinees and also confounds interpretation of the ensuing latent trait metric. An alternative approach to MIRT is suggested here that makes use of unidimensional IRT models in a multidimensional context. This method is referred to as the composite traits approach.

An Overview of the Composite Traits Approach

The composite traits approach is a two-stage process of test calibration and metric interpretation. It assumes that the latent structure is known or capable of being confirmed by the data. By taking this confirmatory approach, issues such as the number of trait dimensions and factor invariance are at least subject to empirical verification through confirmatory factor analysis (e.g., Dillon & Goldstein, 1984; Jöreskog & Sörbom, 1986; Kenny, 1979; Muthén, 1983).

In Stage 1, a factor analysis or MIRT analysis is performed on either the matrix of item covariances or tetrachorics. One advantage of the confirmatory approach is that the number of factors and the correlation between factors can be constrained to fixed values because the latent structure is assumed to be known. For convenience, it is assumed that the latent space is orthogonal.

Following the factor analysis or MIRT analysis, the factor loadings of the items (item discriminations under traditional MIRT analyses) are converted to direction cosines (Reckase, 1985; Reckase & McKinley, 1991). The use of direction cosines removes any confounding of the item location parameters with the discriminations and provides an angular measure of the direction of maximum discriminating power of each item with respect to the latent axes. Those direction cosines are given by

$$\cos(\hat{\alpha}_m) = \frac{\hat{a}_m}{\left(\sum_{m=1}^{M} \hat{a}_m^2\right)^{1/2}} , \tag{2}$$

for $M \geq 1$ dimensions, where $\hat{a}_m$ denotes the factor loading or MIRT discrimination parameter estimate associated with the mth trait factor.

A hierarchical cluster analysis then is performed on the direction cosine differences to identify item subsets having similar orientations in the multidimensional latent space. Miller & Hirsch (1990) previously demonstrated a successful implementation of this type of cluster analysis of multidimensional item direction cosines. The basic procedure involves the generation of a dissimilarity matrix of the angular distances between the items in the latent space, represented by the differences between the direction cosines. A hierarchical cluster analysis then is used to identify the subsets of items that share a similar direction of maximum discrimination or information in the latent space (Miller & Hirsch, 1990).

In Stage 2, each cluster or subset of items is calibrated using a unidimensional IRT model. This process essentially projects a reference composite through each item cluster. This reference composite becomes the unidimensional latent trait metric axis (Wang, 1986). In a cognitive or psychological sense, the composite traits being measured by each item cluster may be multidimensional, depending on the orientation of the reference composite in the latent space. However, in a psychometric sense, a unidimensional IRT model should fit the data, because the clustered items all provide the maximum amount of discrimination in the same general direction.

This clustering of items into subsets in terms of their directions of maximum information is a direct extension of Wang's (1986) original conception of a single reference composite being fit to the entire test. That concept is extended here by capitalizing on the direction of maximum information for blocks of items, in an effort to minimize the loss of valid collateral information that a given item theoretically may provide about more than one trait. The orientation of the reference composite (essentially the mean of the directional cosines for a given item cluster), therefore, provides the meaning of the trait metric with respect to the a priori latent space. This proposed approach is conceptually similar to Levine & Drasgow's (1982) idea of item "blocking" in an attempt to account for variable examinee traits throughout a test.

In a factor analytic sense, this two-stage approach is similar to obtaining a common factor solution, rotated to oblique simple structure. However, the distinguishing features are that: (1) use of the unidimensional model reduces the number of structural and incidental parameters that are estimated; (2) the rotation to simple structure is implicit in the procedure and does not depend on multidimensional structural model constraints; and (3) once the item clusters are formed, the correlation(s) among the latent axes are no longer of concern.

Rationale for the Empirical Study

One rationale for this study was to empirically test the traditional presumption that "more complex" models are always better. It is tempting to argue that MIRT models should fit multidimensional data better than unidimensional models. Yet, that theoretical argument ignores several important issues. For example, the accuracy and stability of empirical item parameter estimates, as well as concerns about the estimates and interpretation of the latent traits, are also valid criteria for judging the utility of an approach. In this study, the nature and extent of trade offs with respect to these types of criteria were evaluated systematically.

The present study was designed to demonstrate that this approach provides advantages over using MIRT models to estimate multidimensional traits and item operating characteristics. Therefore, a MIRT model was directly contrasted with a corresponding unidimensional IRT composite traits model applied to simulated multidimensional data. Results were evaluated in terms of parameter estimation error and bias, trait estimation issues, and model fit in response pattern predictions.

Method

Data Generation

Simulated dichotomous datasets were generated to conform to an orthogonal two-dimensional latent structure comprised of θ_1 and θ_2. Each dataset contained the simulated responses of 2,000 examinees to 50 test items. The 50 items were distributed equally in two clusters, where each cluster of 25 items was oriented with respect to an underlying reference composite. That is, each reference composite denoted the principal direction of item discriminations in the two-dimensional latent space, for the corresponding 25-item cluster. In addition, three conditions were introduced to represent different amounts of variation in the within-cluster angular dispersion of the items about the reference composites.

The two underlying reference composites were oriented, respectively, at 20° and 70° from the θ_1 axis. These particular orientations were selected strictly to maintain positive manifold in the response functions and to sufficiently distinguish the two item clusters in the latent space.

Under the three conditions of within-cluster angular dispersion, the standard deviation (SD) of α_{i1} was varied. (α_{i1} denotes the angle from the θ_1 axis at which each item was most discriminating.) For the first condition, SD_α was set to 0°. This condition forced the 25 items in each cluster to be most discriminating along the corresponding reference composite. The second condition used a within-cluster SD_α of 5°. This condition allowed the 25 items in each cluster to vary somewhat from the preassigned orientation of the reference composite. The final condition set the within-cluster angular SD at 10°. In these last two conditions, the two item clusters could overlap slightly, making the correspondence of items to reference composites (and clusters) less evident. Because the exact angular sampling distribution of item vectors was not known, a normal distribution (with mean 0 and SD_α) was used for sampling the item angles within the two clusters of 25 items.

Given α_{i1}, a multidimensional (MD) three-parameter normal ogive model was used to generate the simulated dichotomous response data. This normal ogive response model,

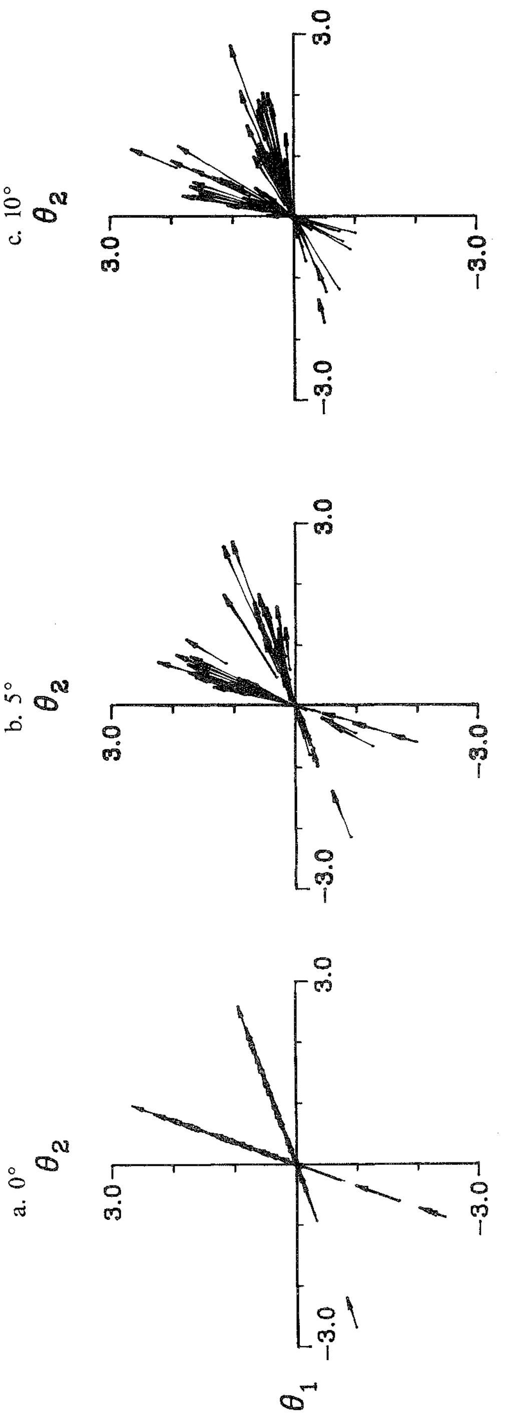
$$P_{ij} \equiv P(u_{ij} = 1 \mid \theta_{j1}, \theta_{j2}; a_{i1}, a_{i2}, d_i, c_i) = c_i + (1 - c_i)\Phi\left(\sum_{m=1}^{2} a_{im}\theta_{jm} + d_i\right) \quad , \tag{3}$$

is similar to the model in Equation 1, but adds a lower asymptote parameter, c_i. a_{im} is the mth element of the item discrimination vector, $\mathbf{a}$, and d_i is item difficulty. For each item, an a_{i1} was randomly sampled from a log normal distribution where $\text{Ln}(a_{i1})$ had a mean of 0.0 and SD of .5. The complementary second discrimination parameter, a_{i2}, was determined in closed form, given the tangent of the item vector angle. That is, $a_{i2} = a_{i1}\tan(\alpha_{i1})$. (For angles greater than 45°, the reciprocal of the tangent of the angle was used, and the a_1 and a_2 terms were interchanged.) The d_i were sampled from a standard normal distribution with a mean of 0.0 and SD of .75. These particular values of the first and second moments of the sampled distributions of MIRT item parameters were selected to correspond with the moments (hyperparameters) of the default prior distributions used by PC-BILOG for the item discriminations and item difficulties (Mislevy & Bock, 1989; see also Mislevy, 1986). In all cases, the lower asymptote was fixed at a constant value of 1/6 (i.e., $c_i = c \approx .167$).

Item vector plots for the 50 items generated under the 0°, 5°, and 10° conditions are shown in Figure 1. The angle of each item vector from either θ axis denotes the direction of maximum discrimination (i.e., information). The length of each vector graphically depicts the relative amount of discrimination in that direction. The distance from the origin to the beginning of each vector denotes the item difficulty in the direction of maximum discrimination. The two reference composites underlying the item clusters are shown at 20° and 70° in each vector plot. Figures 1a through 1c demonstrate how the increased angular dispersion of the items (the SD_α moving from 0° to 10°) potentially confounded the item clustering and, subsequently, the unidimensional (UD) composite θ calibrations involving the item projections of information onto the reference composites.

Ten samples of 2,000 θ_1 and θ_2 values were generated under each condition. The pairs of examinee

Figure 1
Item Vector Plots of the 50 Two-Dimensional Items Used for the Simulation Under Three Conditions of Angular Dispersion Within Clusters



a. 0°

b. 5°

c. 10°

θs were sampled from a bivariate normal distribution ($\mu = [0,0]$, $\Sigma = [1,1]$, $\rho_{\theta1,\theta2} = 0$). Given the MIRT item parameters, the response probability, P_{ij}, (Equation 3) was computed for each examinee. P_{ij} then was compared to a uniform random value P^*_i where $0 \leq P^* \leq 1$. A binary item score of $u_{ij} = 1$ was assigned when $P_{ij} \geq P^*$. Otherwise, a score of $u_{ij} = 0$ was assigned. This procedure, repeated over examinee samples and item dispersion conditions, provided 30 MD dichotomous datasets.

Item Calibrations, Factor Analyses, and Cluster Analyses

Each of the 30 simulated datasets was independently analyzed using NOHARM (Fraser, 1986) on all 50 items to obtain the MD item parameter estimates. NOHARM approximates a normal ogive MD item calibration using a variant of harmonic factor analysis described by McDonald (1982). The NOHARM analyses were constrained to provide a two-factor, orthogonal solution using a varimax rotation.

Three datasets were sampled randomly from each dispersion condition, and were analyzed further using a principal axis factor analysis of the interitem phi coefficients constrained to a two-factor solution. The nine dissimilarity matrices were constructed using the direction cosines computed from both the NOHARM discrimination estimates and from the factor loadings (yielding 18 matrices). The dissimilarity matrices then were analyzed using hierarchical cluster analysis with complete linkage as described by Miller & Hirsch (1990). Although the "true" clustering of items was known and used for the subsequent UD item composite calibrations, these cluster analyses were performed to verify the integrity of the complete linkage method in identifying the item clusters using factor loadings or MD discrimination parameters. Because the "true" clusters of items were known, the pertinent criterion was how well each method replicated the "true" clusters of items in terms of cluster orientation in the θ space and identification of items.

UD IRT calibration analyses were performed on each cluster of 25 items within each of the 30 datasets, using PC-BILOG (Mislevy & Bock, 1989). The default normal ogive approximation was used for appropriately scaling the discrimination parameters. The distributions of beta hyperparameters on the lower asymptote were constrained to be extremely leptokurtic, thus essentially fixing c at .167 for all items.

Results

Integrity of the Stage 1 Clustering Procedure

The range of between-cluster average angle differences was 43° to 63° across the nine datasets in which the dissimilarity matrices were based on the NOHARM discrimination estimates. These average angle differences were computed by calculating the mean angle of each item cluster using the direction cosines (Equation 2), and then computing the difference in mean angles. In contrast, the range of between-cluster average angle differences for the dissimilarity matrices based on the factor loadings was 32° to 59°. In this latter case, there was also a minor rotational shift toward the θ_1 axis. This shift is somewhat apparent given the reported range. However, both sets of results seemed to be consistent with the actual 50° difference angle between the underlying 20° and 70° reference composites used to generate the items. In practice, either method probably would be acceptable for interpreting the relative orientation of the item clusters.

In the 0° and 5° conditions, the cluster analysis was able to replicate the "true" cluster-item mappings without error. That was not the case for the 10° condition. Under the 10° condition, the accuracy in identifying the "true" item clusters ranged from 92% to 98% (for the three analyzed matrices).

Robustness of Item Parameter Estimates

Contrasts between the MD NOHARM and UD PC-BILOG item calibrations for the 30 datasets required a comparable set of estimators denoting location and slope. The MD discrimination estimates, here limited to $\hat{a}_{i1}$ and $\hat{a}_{i2}$ for two dimensions, could not be compared directly to the UD discrimination parameter estimates, $\hat{a}_i$. Instead, the UD discrimination parameters were compared with a MD scalar discrimination index, MDISC (Reckase, 1985). MDISC is the norm of the vector of the MD discrimination parameter estimates. Therefore, for this two-dimensional case,

$$\text{MDISC}_i = \left(\sum_{m=1}^{2} \hat{a}_{im}^2\right)^{1/2} \quad . \tag{4}$$

To compensate for the confounding of direction and location present in the MD model parameter d, MDIFF (Reckase, 1985) was used as the comparative difficulty or location parameter estimate corresponding to the UD $\hat{b}_i$ estimate, where

$$\text{MDIFF}_i = \frac{-\hat{d}_i}{\text{MDISC}_i} \quad . \tag{5}$$

MDISC and MDIFF were calculated from the NOHARM MIRT parameter estimates.

Table 1 provides means and SDs of the MDISC_i and MDIFF_i indices along with their corresponding UD $\hat{a}_i$ and $\hat{b}_i$ parameter estimates. There was a slight downward bias in the $\hat{a}_i$ estimates compared to the "true" MDISC_i values. There was less apparent bias in the estimated MDISC_i indices. This finding is consistent with a discussion by Wang (1986), concerning the projections of the item vectors onto a UD reference composite. At the same time, there appears to be less bias in the $\hat{b}_i$ than in the MDIFF_i estimates. Correlations between the various UD and MD estimators and the "true" values were consistently greater than .96.

Table 1
Mean and SD of UD and MD Item Parameter Estimates
Based on 10 Replications of 50 Items With N = 2,000
per Replication, at Three Levels of Angular Dispersion

	0°		5°		10°	
Parameter	Mean	SD	Mean	SD	Mean	SD
MDISC						
True	1.102	.438	1.066	.446	1.165	.408
Estimated	1.130	.512	1.054	.449	1.176	.435
MDIFF						
True	.065	.797	−.025	1.033	−.008	.667
Estimated	.052	.804	−.056	1.078	−.024	.673
UD						
b	.063	.765	−.031	.984	−.019	.689
a	1.026	.363	1.016	.408	1.071	.350

Table 2 presents the means, medians, SDs, and minimum/maximum values of the empirical standard errors (SEs) of the parameter estimates, taken across replications and summarized over items. These data indicate that the SEs of the UD parameter estimates tend to be smaller than their MIRT counterparts. This reduction in SEs is most evident in terms of the item discriminations. Also noticeable are the larger SDs of the SEs for the MDIFF_i, taken across items.

Table 2
Descriptive Statistics for Empirical SEs of UD and MD Item Parameter Estimates Across Replications and Items for Three Levels of Angular Dispersion

Dispersion	Discrimination		Difficulty	
and Statistic	MDISC	UD $\hat{a}$	MDIFF	UD $\hat{b}$
0°				
Mean	.122	.075	.072	.074
Median	.082	.061	.056	.055
SD	.126	.033	.071	.046
Minimum	.032	.033	.023	.034
Maximum	.747	.151	.384	.277
5°				
Mean	.109	.086	.106	.074
Median	.090	.077	.053	.057
SD	.061	.044	.220	.060
Minimum	.034	.023	.013	.024
Maximum	.286	.249	.382	.414
10°				
Mean	.103	.077	.053	.059
Median	.088	.070	.050	.056
SD	.062	.034	.020	.019
Minimum	.026	.018	.018	.029
Maximum	.317	.161	.112	.110

The improvement in stability under the UD composite θs approach involved only half as many items per calibration as were used for the MIRT calibrations. Of course, even calibrating only 25 items with a sample of 2,000 examinees could be expected to produce fairly stable UD parameter estimates (e.g., Drasgow, 1989). That is, if either cluster had contained poor quality items—items with extreme operating characteristics or very small numbers of items or examinees—this level of stability probably would not be seen. However, a MIRT solution would not be expected to perform any better in those situations.

θ Estimation

Expected a posteriori (EAP) estimates of θ were derived for all simulated examinees in all 30 datasets (50 items × 2,000 examinees). For the vectors of MD θ estimates, a bivariate normal prior distribution was used, $G(\theta_1,\theta_2)$, and an uncorrelated two-dimensional θ space was assumed. The elements of the θ vector for each examinee were approximated by Bayes mean estimators (Mislevy, 1986) computed marginally over the joint posterior θ distributions. That is,

$$\hat{\theta}_{j1} = \sum_{k=1}^{Q} t_k f(t_k, t.) \quad , \tag{6}$$

for t_k, $k = 1, \ldots, Q$ quadrature points along the θ_1 axis, and

$$\hat{\theta}_{j2} = \sum_{l=1}^{R} t_k f(t., t_l) \quad , \tag{7}$$

for t_l, $l = 1, \ldots, R$ quadrature points along the θ_2 axis. The joint posterior distribution is given by

$$f(t_k,t_l) = \frac{L(\mathbf{U}\,|\,t_k,t_l)G(t_k,t_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{R} L(\mathbf{U}\,|\,t_k,t_l)G(t_k,t_l)} \quad , \tag{8}$$

where the likelihood, $L(\mathbf{U})$, taken across all 50 items assumes the general form,

$$L(\mathbf{U}\,|\,t_k,t_l) = \prod_{i=1}^{50} P(t_k,t_l)^{u_i}[1 - P(t_k,t_l)]^{1-u_i} \quad . \tag{9}$$

Similarly, for each examinee, two independent EAP θ estimates corresponding to the two reference composites also were computed from the two clusters of UD item parameter estimates. For each of these EAP θ estimates, the general form of the UD Bayes mean estimator is

$$\hat{\theta}_j = \sum_{k=1}^{Q} t_k f(t_k) \quad , \tag{10}$$

approximated over t_k, $k = 1, \ldots, Q$ quadrature points along the appropriate reference composite, where

$$f(t_k) = \frac{L(\mathbf{U}\,|\,t_k)G(t_k)}{\sum_{k=1}^{Q} L(\mathbf{U}\,|\,t_k)G(t_k)} \tag{11}$$

and

$$L(\mathbf{U}\,|\,t_k) = \prod_{i=1}^{25} P(t_k)^{u_i}[1 - P(t_k)]^{1-u_i} \tag{12}$$

for the 25 items in each cluster.

Descriptive statistics for the "true" θ_1 and θ_2 used to generate the simulated data and both the MD and UD EAP estimates are reported in Table 3. The means in Table 3 represent the mean θs for $N = 2{,}000$ examinees for all 10 datasets under each simulation condition. The SEs of the means are the empirical SDs of the within dataset mean θs taken over the 10 datasets under each condition (the expected value was approximately .02 for this sample size). Finally, the median SD is the median of the within dataset SDs across datasets.

Although Table 3 does not indicate major anomalies in the θ estimates, the MD EAP estimates had smaller SEs of the means and smaller median SDs for both $\hat{\theta}_1$ and $\hat{\theta}_2$. That apparent reduction in the variance of the estimated bivariate θ distribution is relevant in the context of the covariance between the MD EAP estimates, as discussed below.

Table 4 provides median, minimum, and maximum Pearson product-moment correlations between the θ_1 and θ_2 estimates and their true values for each condition, for both MD and UD EAP estimates. The nonzero correlations between the MD EAP estimates, combined with the reduced within-sample SDs of the MD EAP estimates (see Table 3), suggest a discernible bias in the variance-covariance matrix of the EAP scores. This bias is a function of the collateral information present in the MIRT item parameter estimates. Muraki & Engelhard (1985) implied that a varimax rotation might resolve this type of bias for EAP scores (a problem that affects factor scores in general—see Harman, 1976). However, because the varimax-rotated NOHARM discriminations were used for these EAP scores, it is obvious that the bias in variance-covariance matrices did not disappear here.

In terms of the UD composite θs, the observed nonzero correlations between the EAP $\hat{\theta}_1$ and $\hat{\theta}_2$

Table 3
Mean, SE of Mean, and Median SD of True and Estimated UD and MD EAP θ Estimates at Three Levels of Angular Dispersion

Statistic	0°	5°	10°
Mean			
True MD θ_1	-.002	0.000	.006
EAP MD θ_1	-.010	-.001	-.003
EAP UD θ_1	.006	.009	.003
True MD θ_2	.001	.002	-.004
EAP MD θ_2	.008	-.001	-.002
EAP UD θ_2	.004	.010	-.002
SE of Mean			
True MD θ_1	.016	.018	.020
EAP MD θ_1	.007	.002	.003
EAP UD θ_1	.021	.012	.014
True MD θ_2	.035	.011	.012
EAP MD θ_2	.005	.002	.002
EAP UD θ_2	.011	.009	.017
Median SD			
True MD θ_1	1.007	1.000	1.011
EAP MD θ_1	.890	.901	.907
EAP UD θ_1	.962	.935	.955
True MD θ_2	1.005	1.006	1.010
EAP MD θ_2	.834	.873	.879
EAP UD θ_2	.933	.943	.954

are both informative and, in a certain sense, quite consistent with the MD generating structure used in the simulation. That is, the correlations between the UD EAP θ estimates tend to closely approximate the cosine of the angle between the cluster-generating reference composites [cos(50°) = .643]. The UD estimates correctly approximated the orientations of the latent θ axis to which the items were

Table 4
Median, Minimum, and Maximum Correlation Between True θ_1 and θ_2 and Their Estimates at Three Levels of Angular Dispersion

Dispersion and Variables	Median	Minimum	Maximum
0°			
True θs	-.008	-.054	.005
MD EAP Est.	.199	.180	.223
UD EAP Est.	.666	.651	.682
5°			
True θs	-.020	-.041	.004
MD EAP Est.	.141	.120	.163
UD EAP Est.	.561	.535	.574
10°			
True θs	-.018	-.039	.006
MD EAP Est.	.121	.109	.164
UD EAP Est.	.601	.591	.631

most sensitive. In fact, the observed correlations of UD EAP estimates in Table 4 are almost identical to the cosines of the average angular differences between the item clusters. Those cosine differences were $\cos[\bar{\alpha}_{1(0°)}] = .67$, $\cos[\bar{\alpha}_{1(5°)}] = .53$ and $\cos[\bar{\alpha}_{1(10°)}] = .60$. This is an important finding that suggests that the composite traits approach unidimensionally achieves a type of oblique simple structure. However, the correlation between traits can be computed post hoc from the composite θ estimates. There is no need to estimate additional structural parameters or transformation matrices.

Model Fit

Estimation of the UD θs along the reference composites somewhat complicates the matter of goodness-of-fit. Due to the oblique reference composites, the UD composite θ metrics are not simply nested subsets of the MD θ metrics. As a result, the use of likelihood ratio χ^2 tests are not valid here because the latent metrics as well as the parameterizations of the models are different. Instead, two separate approaches were taken to evaluate the goodness-of-fit of the models to the data.

The first approach was to compute unstandardized root-mean-square residuals (RMSR) between the observed dichotomous responses and the predicted item "true scores" in an adaptation of a model-fit index described by Wright & Stone (1979). For the MD model, RMSR was computed as

$$\mathrm{RMSR} = \left\{ N^{-1} n^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} - P_i(\hat{\theta}_{j1}, \hat{\theta}_{j2})]^2 \right\}^{1/2} \quad , \tag{13}$$

conditioning the item "true scores" in each examinee's MD EAP θ estimates. For the UD model, RMSR was computed, matching items to the appropriate reference composite, as

$$\mathrm{RMSR} = \left\{ N^{-1} n^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n(1)} [u_{ij} - P_i(\hat{\theta}_{j1})]^2 + \sum_{j=1}^{N} \sum_{i=1}^{n(2)} [u_{ij} - P_i(\hat{\theta}_{j2})]^2 \right\}^{1/2} \quad . \tag{14}$$

The fact that these RMSR indices are susceptible to even minor fluctuations in fit was actually desired, given these simulation data.

Under the second approach to evaluating model fit, a standardized log likelihood was computed for each examinee, under both models, using their corresponding MD or UD EAP θ estimates. Because this "appropriateness" index (Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982) usually approximates a normal (0,1) distribution, it was possible to identify as "aberrant" those examinees whose standardized log likelihood exceeded a 95% confidence interval of l_0 (i.e., $|l_z| > 1.96$) under the MD model, under the UD model, or under both models. Larger discrepancies in identifying aberrant examinees, especially in the case of the UD model, would tend to indicate poor fit.

A summary of the RMSR results is shown in Table 5. One rather clear implication of the RMSR indices is that there was little difference in fitting the observed responses between the MD and UD models. The UD model may have fit slightly better under the 0° condition, because that condition essentially replicated two purely UD, but correlated, subtests. Similarly, as the item sensitivity moved away from the reference composite (toward 10°), there was a very slight tendency for the MD model to fit better.

The model-fit summary in Table 6 shows the empirical proportions of examinees identified as aberrant under the MD model, under the UD model(s), and under both models. These latter results indicate that no important differences existed between models. In general, both models identified the same individuals as aberrant. Furthermore, the proportions of cases selected exclusively under one model or the other were essentially equal.

Table 5
Mean and SD of RMSR for UD and MD Fit for Three Degrees of Angular Dispersion

Dispersion and Statistic	MD Fit	UD Fit
0°		
Mean	.4089	.4080
SD	.0011	.0011
5°		
Mean	.4027	.4024
SD	.0010	.0008
10°		
Mean	.3999	.4015
SD	.0010	.0011

Discussion and Conclusions

The results of these simulations indicated that separate unidimensional reference composites in a two-dimensional orthogonal latent space could be adequately fit to clusters of items having similar angular orientations in the space, even when there was a fairly large within-cluster angular dispersion of the item discriminations. When the items within each cluster were calibrated unidimensionally, the resulting discrimination and difficulty parameters were more accurate and stable than corresponding multidimensional parameters.

The unidimensional composite trait model also tended to fit the data quite well, and was essentially indistinguishable from the multidimensional model in terms of model fit indices. Trait estimation along the unidimensional reference composite metrics appeared useful for two reasons. First, the obtained estimates did not show the variance reduction present in the multidimensional trait estimates (i.e., estimation bias due to a covariance in the multidimensional trait estimates as a function of the collateral information in the items). Second, the product-moment correlations between the unidimensional trait estimates seemed to adequately represent the underlying structure of the multidimensional tests. To achieve similar results multidimensionally would have required an oblique solution that would have increased the parameterization of the model by adding a correlation matrix to the model and created additional sources of potential estimation error.

These results suggest that for tests having identifiable clusters of items, the theoretical case for MIRT models may actually overestimate the magnitude of information loss and overstate the degradation in fit under more parsimonious IRT models. This does not, however, imply that multidimensionality can be ignored completely. Rather, by capitalizing on derived knowledge of the direction

Table 6
Proportions of Examinees Identified as Aberrant by the Model Fit Indices Under the UD, MD, or Both Models for Three Degrees of Angular Dispersion

Dispersion	Selected by Both Models	Selected by Only MD Model	Selected by Only UD Model
0°	.0185	.0004	.0009
5°	.0194	.0010	.0010
10°	.0178	.0010	.0009

of maximum information of items and then clustering items having similar orientations, it was demonstrated that composite trait metrics could be accurately calibrated while retaining some relationship to the multidimensional structure of the test.

These results are important for three reasons.

1. The study demonstrated that it is feasible to make use of available unidimensional IRT procedures, as well as available factor and cluster analytic techniques, to calibrate multidimensional test data. This capability provides some direction toward a practical and workable solution to multidimensional test construction and analysis problems within the framework of existing resources.
2. It was shown that it is possible to retain multidimensional interpretations of composite traits, even though the test calibrations are conveniently unidimensional. That is, the common problem of assuming unidimensionality for an entire test is avoided by restricting that assumption to items shown to directionally cluster together in the multidimensional space. This approach further retains the essential ''structure'' of a multidimensional test through the item cluster orientations and correlations between trait estimates.
3. Some empirical evidence was provided that the stability of the unidimensional item parameter estimates, for tests having well-formed clusters of items, may actually be better than their multidimensional counterparts, with little degradation in model fit or reduction in the accuracy of trait estimates, while using fewer items.

Implications for Future Research

Future research should investigate the limitations of this technique. For example, if items cannot be accurately assigned to clusters or the within-cluster variance becomes quite large, it may be more appropriate to use a full-information MIRT model. Another research possibility concerns the potential effect of different levels of correlation between the reference composites. In cases of highly correlated reference composites, it is likely that a single composite trait metric, derived using all the test items, might provide the best results.

Another area of application for which the composite trait procedures may be useful is the construction of multidimensional parallel tests. In this application, the composite traits approach might provide a solution to deriving near-parallel subtests. This follows from recent evidence that multidimensional test parallelism can be achieved unidimensionally (Ackerman, 1991). Ackerman used computerized item selection techniques developed by Luecht & Hirsch (1990, 1991, 1992) for building unidimensional tests, even though the tests supported a multidimensional trait structure. He then demonstrated that the resulting test forms achieved a reasonable degree of parallelism in terms of multidimensional criteria. It is likely that by incorporating the composite traits approach, the relative degree of multidimensional parallelism might be improved further. Although more work on this topic is required, Ackerman's initial results are encouraging.

Another area in which the composite trait procedures may be useful is in linking test calibrations. The unidimensional treatment of the composite traits seems to simplify some of the problems of placing trait estimates and item parameters from different forms of multidimensional tests on the same scale(s). For example, by using anchor items within clusters to establish the composite trait metrics corresponding to existing base subtest reference composites (through estimates of the unidimensional examinee posterior distributions of the composite traits), new nonanchor items measuring similar composite traits could be individually calibrated to the same metrics. Any remaining parameter scaling would involve a linear transformation of the unidimensional discriminations and difficulties (e.g., Mislevy & Bock, 1989).

The results and conclusions suggested by the present study must be tempered with caution. The limited simulation in this study and its restriction to an orthogonal, two-dimensional trait space reduce the generalizability of the results. Nevertheless, the results are encouraging with respect to suggesting a set of practical solutions for dealing with some of the current problems of multidimensional tests.

References

Ackerman, T. A. (1991, April). *An examination of the effect of multidimensionality on parallel forms constructions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.

Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37–48.

Berger, M. P. F. (1990, April). *On the assessment of dimensionality in item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Boston MA.

Bock, R. D., & Aiken, M. (1981). Marginal maximum likelihood estimation for item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–458.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Report 87-19). Iowa City IA: ACT.

Davey, T., & Hirsch, T. M. (1990, June). *Examinee discrimination as a measure of test data dimensionality.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications.* New York: Wiley.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Fraser, C. (1986). *NOHARM II: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: Center for Behavioral Studies, The University of New England.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale NJ: Erlbaum.

Harman, H. H. (1976). *Modern factor analysis.* Chicago: University of Chicago Press.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55,* 5–18.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of structural relationships by maximum likelihood and least squares methods.* Chicago: National Educational Resources.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42–56.

Luecht, R. M., & Hirsch, T. M. (1990). *Computerized test construction using an average growth approximation of target information functions* (Research Rep. No. 90-6). Iowa City IA: ACT.

Luecht, R. M., & Hirsch, T. M. (1991, June). *Computerized test construction of parallel forms for problem-linked items.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick NJ.

Luecht, R. M., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16,* 41–63.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

Miller, T., & Hirsch, T. M. (1990, June). *Cluster analysis of angular data in applications of multidimensional item response theory.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. J., & Bock, R.D. (1989). *PC-BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Muraki, E., & Engelhard, G. (1985). Full-information factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417–430.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 43–65.

Reckase, M. D. (1985). The difficulty of test items that measure more than one dimension. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Canada.

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

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111–121.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 79–98.

Stout, W. F. (1990). A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Sympson, J. B. (1978). A model for testing with 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.

Wang, M. M. (1986, April). *Fitting a unidimensional model to multidimensional item response data* (ONR Rep. 042286). Iowa City IA: University of Iowa.

Way, W. D. (1990). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models.* Paper presented at the annual meeting of the American Educational Research Association, Boston MA.

Wilkerson, L. (1988). *SYSTAT: The System for Statistics* [Computer program]. Evanston IL: SYSTAT, Inc.

Wilson, D., Wood, R., & Gibbons, R. (1984). *TESTFACT: Test scoring and item factor analysis.* Mooresville IN: Scientific Software.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago IL: Mesa Press.

Author's Address

Send requests for reprints or further information to Richard M. Luecht, Support Technological Applications and Research, American College Testing, 2201 North Dodge Street, P.O. Box 168, Iowa City IA 52243-0168, U.S.A.