JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education

Journal of Educational Measurement
Summer 2011, Vol. 48, No. 2, pp. 146–164

Relating Unidimensional IRT Parameters to a Multidimensional Response Space: A Review of Two Alternative Projection IRT Models for Scoring Subscales

Nilufer Kahraman
National Board of Medical Examiners
Tony Thompson
Pearson

A practical concern for many existing tests is that subscore test lengths are too short to provide reliable and meaningful measurement. A possible method of improving the subscale reliability and validity would be to make use of collateral information provided by items from other subscales of the same test. To this end, the purpose of this article is to compare two different formulations of an alternative Item Response Theory (IRT) model developed to parameterize unidimensional projections of multidimensional test items: Analytical and Empirical formulations. Two real data applications are provided to illustrate how the projection IRT model can be used in practice, as well as to further examine how ability estimates from the projection IRT model compare to external examinee measures. The results suggest that collateral information extracted by a projection IRT model can be used to improve reliability and validity of subscale scores, which in turn can be used to provide diagnostic information about strength and weaknesses of examinees helping stakeholders to link instruction or curriculum to assessment results.

Unidimensional Item Response Theory (IRT) models assume that a person's ability can be estimated in a unidimensional latent space. Reservations about this assumption have brought much attention to Multidimensional Item Response Theory (MIRT), which can model the relationship between two or more latent variables and the probability of correct answers to test items (e.g., Hattie, 1985; Stout, 1990; Kirisci, Hsu, & Yu, 2001; Bolt & Lall, 2003). To date, however, MIRT models are not widely used in practice in spite of their potential advantages over IRT models, such as better model fit or better diagnostic information. One central reason is that estimating and interpreting multidimensional counterparts of unidimensional concepts becomes increasingly complex as the number of dimensions increase, requiring maximization in high-dimensional space (e.g., Finch, 2009; Reckase, 2007; Sheng, 2010).

Practitioners continue using IRT models and rely on robustness studies which show that IRT models are reasonably robust to the violation of the unidimensionality assumption under the condition that a certain (unidimensional) composite of skills accounts for examinees' performance (e.g., Ackerman, 1994; Reckase, Ackerman, & Carlson, 1988). Considering the composite trait structure as a valid and replicable combination of latent abilities underlying response data, *the composite unidimensionality assumption* implies that test items may have the capability of discriminating

with respect to multiple trait components (e.g., Mislevy & Sheehan, 1989). Item and test statistics obtained through unidimensional models, then, at best could provide information about a single continuum on which multiple dimensions are collapsed, often in some unknown way (Miller & Hirsch, 1992). It is not uncommon that psychometricians take a second look at the data from a multidimensional perspective as they check for the quality of the scoring processes and try to delineate the combination of latent abilities that individuals use to obtain a correct response and whether these combinations vary from item to item. Nonetheless, the outcomes of the MIRT analysis are often isolated from the actual scoring processes and do not necessarily relate to the targeted unidimensional interpretation of the component trait measures.

As stakeholders demand more diagnostic information (without dramatic increases to test length), test developers focus more on constructing tests that can be reliably scored to infer about various aspects of examinee proficiency (e.g., Roussos, Templin, & Henson, 2007; Goldberg & Roswell, 2001; Wiggins, 1998). For example, test specifications of integrated performance assessments often include test items that are scored for evidence of proficiency in more than one content area. There is an increasing need for alternative measurement models that can make use of such multilevel information in item response data, yet at the same time allow unidimensional interpretations and, most importantly, can be used in practice.

There are two general approaches in the literature that aim to make better use of multilevel item information in response data. The first approach focuses on improving efficiency of the estimation processes of MIRT models for better diagnostic information. Research on the topic is quite active and includes a wide range of promising applications that incorporate additional hierarchical group and domain variables into the item parameter estimation processes (de la Torre & Patz, 2005; Mislevy, 1987; Kahraman, De Boeck, & Janssen, 2009; Rijmen, 2010; Yen, 1987; Wainer, Sheehan, & Wang, 2000). The second approach focuses on obtaining useful unidimensional (i.e. subscale specific) interpretations of multidimensional test items, and it is the subject of this study.

Purpose

This study reviews two different formulations of an alternative IRT model developed to parameterize collateral item information through *unidimensional projections of multidimensional test items*. Both formulations are based on the central idea that item response data might be multidimensional in its naturally occurring trait structure yet could be usefully interpreted unidimensionally. Composite trait unique unidimensional item and ability parameters are estimated after a transformation or projection process is applied to response data. Since linear or nonlinear transformations of multidimensional data to a lower dimensionality are often referred to as projections (Lee & Verleysen, 2007), the alternative IRT models described in this paper are referred to as the projection IRT models.

The general idea of the proposed projection methods is that a test item can be calibrated with respect to a number of reference composites and is expected to retain discrimination power with respect to those that it measures as *a multidimensional* item. For example, an item in a math test measuring both geometry and algebra

subdomains can be calibrated with respect to (1) the geometry subdomain composite, and (2) the algebra subdomain composite, in addition to (3) the overall test composite. Given that this item measures a composite trait (Geometry and Algebra) rather than a single trait (Geometry or Algebra), it is expected that it will provide subdomain-specific information. The next section provides a review of the two projection methods that have been proposed in the literature and is followed by two real data applications illustrating how the projection IRT models can be used in practice.

Method

Analytical Procedure

Wang (1986) showed that when a unidimensional IRT model is fit to multidimensional response data (collapsing multiple test dimensions onto a single test composite continuum), unidimensional item parameters of items are, in fact, unidimensional projections of test items with respect to the test composite. The compound trait estimated by a unidimensional model is defined as the first eigenvector of A′A, where A is the ($n \times k$) matrix of discrimination parameters of n ($i = 1, \ldots, n$) items on k dimensions ($k = 1, \ldots, k$). Assuming the latent trait matrix, $\boldsymbol{\theta}$, is standardized as $\boldsymbol{\theta} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_i)$ is a positive definite matrix with unit diagonal elements, a linear composite θ_α of the latent variables is defined to be a standardized linear combination of $\theta_1, \theta_2, \ldots\ldots, \theta_k$ as (Wang, 1986)

$$\theta_\alpha = \boldsymbol{\alpha}^t \boldsymbol{\theta} = \sum_{k=1}^{K} \alpha_k \theta_k, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots\ldots, \alpha_k)^t$ is a constant vector with nonnegative α_j such that $\boldsymbol{\alpha}^t \boldsymbol{\Sigma} \boldsymbol{\alpha} = 1$ (i.e., Var $\theta_\alpha = 1$) (Wang, 1986, 1988; Zhang & Wang, 1998). The reference composite vector, $\boldsymbol{\alpha}$, represents the direction of composite θ_α in the m-dimensional test space, and is used to compute the Analytical unidimensional item discrimination (a_i^*) and item difficulty (b_i^*) parameter projections for composite, θ_α, using

$$a_i^* = \frac{\mathbf{a}_i^t \Sigma \alpha}{\sqrt{1 + \sigma_i^2}}, \tag{2}$$

$$b_i^* = \frac{b_i}{\sqrt{1 + \sigma_i^2}}, \tag{3}$$

$$\sigma_i^2 = \mathbf{a}_i^t \boldsymbol{\alpha} \mathbf{a}_i - \left(\mathbf{a}_i^t \Sigma \alpha\right)^2, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots\ldots, \alpha_k)^t$, $\mathbf{a}_i$ is a vector of item discrimination parameters of item i on k dimensions and Σ is a correlation matrix of $\boldsymbol{\theta}$ (i.e., a covariance matrix with

diagonal values of 1). Formulas 2, 3, and 4 show that the unidimensional item discrimination and item difficulty for a particular reference composite ($\boldsymbol{\alpha}$) are functions of the MIRT item parameters and the correlation matrix of $\boldsymbol{\theta}$.

Empirical Procedure

A series of unidimensional IRT models is fit to response data in parts that are known to measure a target trait of interest (i.e., subtests) to obtain unidimensional projections of test items (Ackerman & Davey, 1991; Davey & Hirsh, 1991; Kahraman & Kamata, 2004). First, items are calibrated with respect to the primary domain dimension that they are classified to measure. The resulting IRT item parameter estimates are referred to as in-scale item parameter estimates. Second, items that are not primarily classified to measure the target domain dimension, but are expected to have some relevant information with respect to it, are *individually* calibrated with all the target domain items while holding their already estimated in-scale item parameters constant. The resulting item parameters are referred to as out-scale item parameters.

Out-scale item parameters for a specific item (item $n + 1$) is estimated by maximizing the likelihood function (Kahraman & Kamata, 2004)

$$L(\mathbf{U} \mid \boldsymbol{\theta}, a_{n+1}, b_{n+1}) = \prod_{j=1}^{s} \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} [1 - P_i(\theta_j)]^{1-u_{ij}} \times P_{n+1}(\theta_j)^{u_{(n+1)j}} [1 - P_{n+1}(\theta_j)]^{1-u_{(n+1)j}}, \tag{5}$$

where the item response function is defined as the unidimensional two-parameter logistic IRT model, $\mathbf{U}$ is the response pattern matrix for all examinees including all in-scale items and 1 out-scale item, $\boldsymbol{\theta}$ is the ability parameter vector for all examinees, u_{ij} is the item response to the ith item by the jth person, $P_i(\theta_j)$ is the item response function for item i ($i = 1, \ldots, n$) with known item parameters (a_i and b_i for $i = 1, \ldots, n$ are held constant), and $P_{n+i}(\theta_j)$ is the item response function for item $n + 1$ with unknown item parameters (i.e., a_{n+1} and b_{n+1}). The purpose of calibrating out-scale items one at a time is to ensure that other out-scale items do not contaminate the scale of the calibration by introducing nonprimary domain dimensions.

The Empirical procedure clearly has the advantage of simplicity. Its calibration process uses unidimensional IRT models only. The Analytical procedure, on the other hand, has the advantage of precision. By grounding its calibration process directly on the results from a MIRT model, it utilizes multidimensional item information. However, the Analytical procedure is subject to the usual difficulties encountered when estimating and interpreting high-dimensional IRT models.

Real Data Application 1: Two-dimensional (2D) Item Data

Table 1 lists the 2D item parameters of the application data that might be familiar to MIRT and the Analytical projection IRT model researchers (for various illustrative MIRT and Analytical projection IRT model applications on the data, see Reckase, 1985; Reckase, 2007; Reckase & McKinley, 1991; Wang, 1986). Figure 1

Table 1
True Item Parameters for a 20-item Test

Item Number	a_1	a_2	d	MDISC	COS	Angle	MDIFF	Slope
1	1.81	.86	1.46	2.00	.90	25	−.73	.50
2	1.22	.02	.17	1.22	1.00	1	−.14	.31
3	1.57	.36	.67	1.61	.97	13	−.42	.40
4	.71	.53	.44	.89	.80	37	−.50	.22
5	.86	.19	.10	.88	.98	12	−.11	.22
6	1.72	.18	.44	1.73	.99	6	−.25	.43
7	1.86	.29	.38	1.88	.99	9	−.20	.47
8	1.33	.34	.69	1.37	.97	14	−.50	.34
9	1.19	1.57	.17	1.97	.60	53	−.09	.49
10	2.00	.00	.38	2.00	1.00	0	−.19	.50
11	.24	1.14	−.95	1.16	.21	78	.82	.29
12	.51	1.21	−1.00	1.31	.39	67	.76	.33
13	.76	.59	−.96	.96	.79	38	1.00	.24
14	.01	1.94	−1.92	1.94	.01	90	.99	.49
15	.39	1.77	−1.57	1.81	.22	78	.87	.45
16	.76	.99	−1.36	1.25	.61	52	1.09	.31
17	.49	1.10	−.81	1.20	.41	66	.67	.30
18	.29	1.10	−.99	1.14	.25	75	.87	.28
19	.48	1.00	−1.56	1.11	.43	64	1.41	.28
20	.42	.75	−1.61	.86	.49	61	1.87	.21

Note. See citations for the source of this data.

plots vector presentations of the items along with measurement directions implied by the test reference composite (all items) and by the reference composites of the first and the last ten items, which will be referred to as Domain 1 and Domain 2 in the following text. The tick dashed line shows the measurement direction of the test reference composite (with a 37-degree angle with θ_1 axis), the dotted dashed line shows the measurement direction of the Domain 1 reference composite (with a 16-degree angle with θ_1 axis), and the dotted line shows the measurement direction of the Domain 2 reference composite (with a 72-degree angle with θ_1 axis). The angle formed by Domain 1 and Domain 2 measurement directions translate to a correlation of .56.

In the figure, the direction of the item vector indicates the direction in space the item best measures, while the length of the vector is a scaled value of multidimensional discrimination and indicates how discriminating the item is in that direction of the space. The distance from the origin to the beginning of each vector denotes the item difficulty in the direction of maximum discrimination. The figure shows that the Domain 2 items (last 10 items; the upper right vector cluster) has more difficult and discriminating items that do not closely cluster around the domain composite when compared to Domain 1 items (first 10 items, the lower left vector cluster). It is clear that these item parameters have a confounding of item difficulty and dimensionality and that a unidimensional IRT model would fit Domain 1 better than it would fit Domain 2 response data.

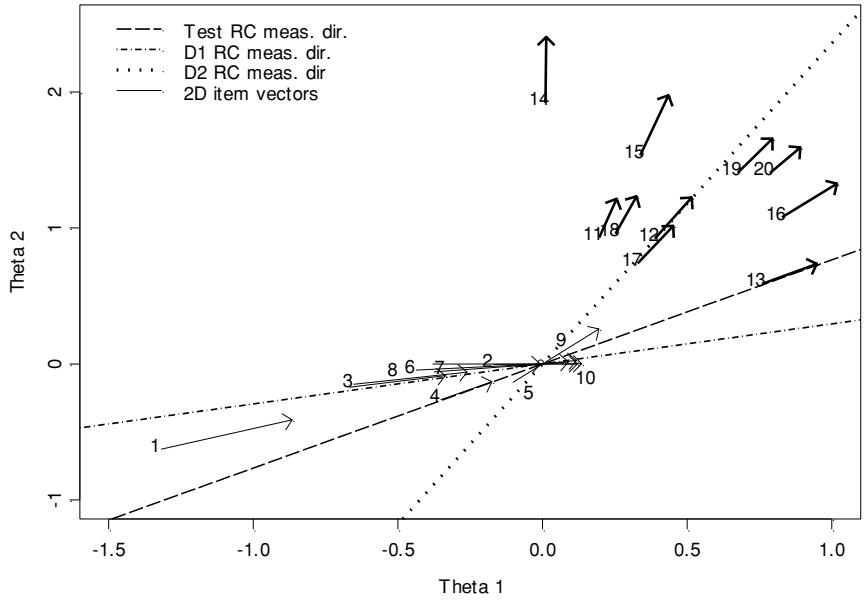

Figure 1. True item vector locations and reference composite measurement directions.

To illustrate how unidimensional Analytical and the Empirical projection IRT model parameters would be calibrated, binary item response data for 1,000 examinees were generated using the true parameters described above over 30 replications.

Parameter Estimation

Item parameter estimates were obtained using the two-parameter logistic (2PL) models. In this model, the probability that a person j answers item i correctly is given by (Reckase & McKinley, 1991)

$$P(yij = 1 \mid \theta_j) = \frac{\exp\left(\sum_{k=1}^{K} \mathrm{a}_{ik}\theta_{jk} - di\right)}{1 + \exp\left(\sum_{k=1}^{K} \mathrm{a}_{ik}\theta_{jk} - di\right)}, \tag{6}$$

where y_{ij} is the score on item i ($i = 1, \ldots, \mathrm{I}$) by person j ($j = 1, \ldots, \mathrm{J}$), a_i is a vector of item discrimination parameters $\mathrm{a}_i = (a_{i1}, a_{i2}, \ldots, a_{ik})$', d_i is a scalar parameter related to item location, and θ_j is a vector of ability parameters from a multivariate normal distribution for person j on k dimensions ($k = 1, \ldots, \mathrm{k}$), $\theta_j = (\theta_{\mathrm{j}1}, \theta_{j2}, \ldots, \theta_{jk})$. For model identification purposes, means (vectors) and variances (matrix) of θ were restricted to zero and one (identity matrix), respectively. The model in (6) reduces to a unidimensional 2PL model when $k = 1$.

In the following text, unidimensional item parameter calibrations with respect to primary and nonprimary subtest reference composites will be referred to as *in-scale* and *out-scale* parameterizations, respectively. Clearly, this distinction is not

meaningful at the test level because all test items would be in-scale items. All computations were done using Splus (Version 7.0) except for the 2D MIRT models for which Mplus (Muthén & Muthén, 2001) was used.[1] Marginal maximum likelihood estimators were obtained through an expectation maximization algorithm (Bock & Aitkin, 1981) for all models. A difference of less than .00001 between parameter estimates of consecutive iterations was used as convergence criteria. The integrals over the latent variables were approximated with Gaussian quadrature with 20 quadrature points for the unidimensional models and 15 quadrature points per dimension for the 2D MIRT models. For application 2, expected posterior estimates were computed.

Evaluating Similarities of the Empirical and the Analytical Procedure Produced by In-scale and Out-scale Item Parameters

Mean differences (between models) and standard errors (SEs, within models) were computed over 30 replications by the following formulae

$$\text{Mean Difference } (\delta_j) = \frac{1}{30}\sum_{h=1}^{30}\left(\hat{\delta}_{Ej} - \hat{\delta}_{Aj}\right), \tag{7}$$

and

$$SE\,(\delta_j) = \sqrt{\frac{1}{30-1}\sum_{h=1}^{30}\left(\hat{\delta}_{jh} - \frac{\sum_{h=1}^{30}\hat{\delta}_{jh}}{30}\right)^2}, \tag{8}$$

where $\hat{\delta}_{Ej}$ and $\hat{\delta}_{Aj}$ are the Analytical and Empirical unidimensional item parameter a or b for item j, and $\hat{\delta}_{jh}$ is the Analytical or Empirical unidimensional item parameter a or b obtained for item j in the hth replication ($h = 1, \ldots, 30$).

The reliabilities were computed from the marginal error variance (Sireci, Thissen, & Wainer, 1991) and setting the mean and the variance of θ to be zero and 1, respectively:

$$\hat{\rho} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2}, \tag{9}$$

where

$$\sigma_e^2 = \frac{1}{\int_{-\infty}^{\infty} I\,(\theta)\, g\,(\theta)\, d\theta}. \tag{10}$$

Results

Table 2 shows average SEs and mean differences for the projected item parameter estimates obtained by the Analytical and the Empirical procedures over replications.

Table 2

Average Standard Errors (SEs) and Mean Differences (Empirical – Analytical) for the Projected Unidimensional Item Parameters Obtained from the Analytical and the Empirical Procedures over 30 Replications

SEs for the Analytical Procedure										
	Test*				Subtest 1**			Subtest 2**		
Par.	*d*	*a*	*b*		*d*	*a*	*b*	*d*	*a*	*b*
Mean (Std. Dev.)	.08 (.03)	.11 (.03)	.12 (.08)	Domain 1	.06 (.01)	.12 (.04)	.05 (.02)	**.06 (.01)**	**.08 (.04)**	.15 (.05)
				Domain 2	**.10 (.02)**	**.08 (.02)**	**.46 (.26)**	.10 (.02)	.13 (.04)	**.12 (.04)**
SEs for the Empirical Procedure										
	Test				Subtest 1			Subtest 2		
Par.	*d*	*a*	*b*		*d*	*a*	*b*	*d*	*a*	*b*
Mean (Std. Dev.)	.10 (.02)	.14 (.03)	.11 (.04)	Domain 1	.09 (.02)	.14 (.05)	.07 (.01)	**.02 (.02)**	**.02 (.01)**	.03 (.01)
				Domain 2	**.02 (.01)**	**.01 (.01)**	**.04 (.02)**	.11 (.02)	.15 (.04)	**.12 (.05)**
Mean Differences (Empirical – Analytical)										
	Test				Subtest 1			Subtest 2		
Par.	*d*	*a*	*b*		*d*	*a*	*b*	*d*	*a*	*b*
Mean (Std. Dev.)	−.10 (.16)	−.14 (.13)	.23 (.28)	Domain 1	−.20 (.19)	−.05 (.06)	.12 (.10)	**−.06 (.19)**	**−.19 (.27)**	−.03 (.17)
				Domain 2	**−.18 (.16)**	**−.07 (.11)**	**.77 (.83)**	.03 (.05)	−.05 (.03)	**.01 (.04)**

*20 items.

**10 items.

For clarity, the table lists the means (standard deviations) over items (over 20 items at the test level and over 10 items at subtest level). Expected SEs are reported first and show that both the Analytical procedure deriving unidimensional projections from multidimensional item parameters and the Empirical procedure estimating projection item parameters from subsequent unidimensional IRT models were subject to some estimation error. Mean SEs of the test-level unidimensional projections for the Analytical approach were .08 and .11 and were slightly lower than those of the Empirical approach (.10 and .14) for item location and discrimination parameters, respectively. For the Analytical procedure, mean SEs of subtest-level unidimensional item location projections were in the range of .06 and .10 and were lower for Domain 1 item parameters. For the Empirical procedure, mean SEs of the subtest-level unidimensional item location projections were in the range of .02 and .11 and were lower for in-scale items. Mean Differences (Empirical – Analytical) were in the range of .03 and –.20 for item location and –.05 and –.19 for item discrimination parameters. These differences suggest that, when compared to the Analytical procedure, the Empirical procedure had slightly larger item discrimination parameters and slightly smaller item difficulty parameters (Mean Differences were in the range of –.03 and .77 for item difficulty parameters).

To remove differences that might be due to scale differences alone, re-scaling was considered. A closer look at the parameter estimates at the item level, however, indicates that observed minor differences were mainly due to a few items. Figures 2 and 3 plot the average discrimination and difficulty parameters obtained by the Empirical and Analytical procedures for *individual* test items over 30 replications. These figures show that differences between item parameter estimates produced by the Empirical and the Analytical projection IRT models are very small except for a few outlier parameter estimates. These are out-scale discrimination parameter estimates for items 1, 3, and 9, and out-scale difficulty parameter estimates for items 14 and 15. True Analytical projection item parameters (computed from the true 2D item parameters listed in Table 1) are also shown in the figure and show that the Analytical procedure is reasonably free from estimation bias. A joint interpretation of these outlier unidimensional item parameters with the true 2D item parameters suggest that (1) the Analytical procedure is more likely to produce rather extreme item parameter estimates for items that are too easy or have near zero discrimination with respect to one of the component test dimensions, and (2) the Empirical procedure might lead to slightly lower item difficulty and higher item discrimination parameters when subtest items are too easy or not discriminating enough. These results indicate that observed discrepancies are minor and might be, at least in part, due to difficulties encountered fitting a multidimensional model to response data in the Analytical procedure (e.g., do all items load into the same number of dimensions?) or trying to capture multidimensional item-level information using unidimensional models in general (e.g., are we projecting relevant item information?).

Item Parameter Projections

Item vector presentations in Figure 1 and item parameter estimates plotted in Figures 2 and 3 reveal that an item's projected unidimensional discrimination power

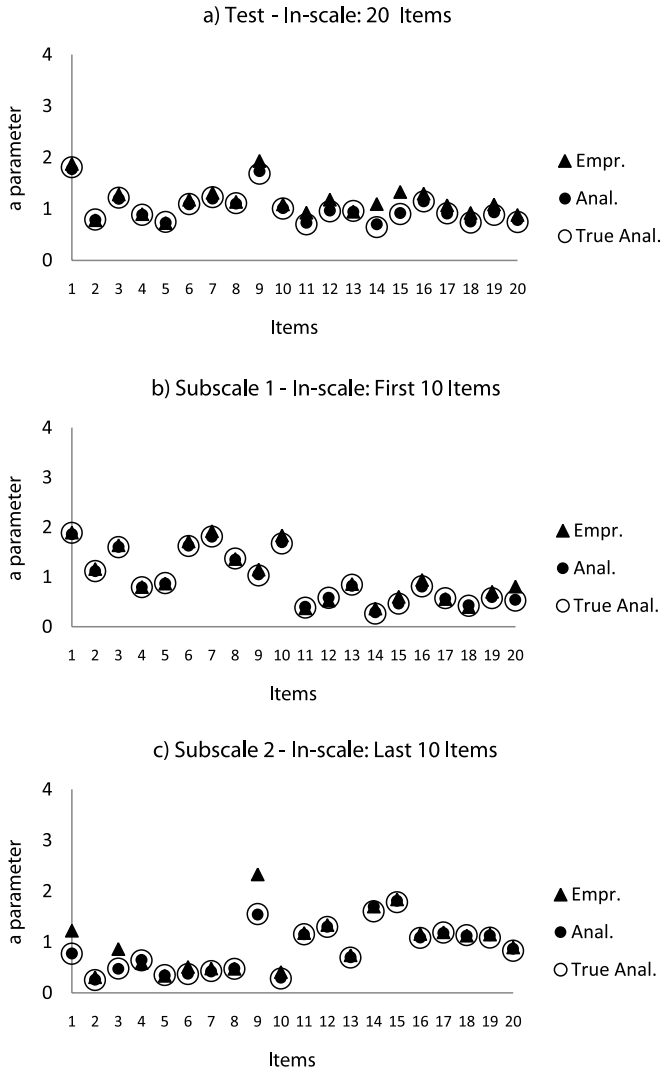

Figure 2. Analytical and Empirical discrimination parameter estimates.

reflects where the item is located in the 2D space relative to the composite of interest. Given that an items' proximity in the latent space to a test composite should coincide with test specifications, the results confirm that items are often most discriminating in their home domain and lose most of their discrimination power outside of that domain. However, there are two items in this test that illustrate how some test items may retain less discrimination power in their own subdomains. It can be seen in Figure 1 that, when projected, item 13 in Subtest 2 retains more discrimination power with respect to the Domain 1 reference composite (.825) than it does with respect to the Domain 2 reference composite (.699) while being most discriminating on the

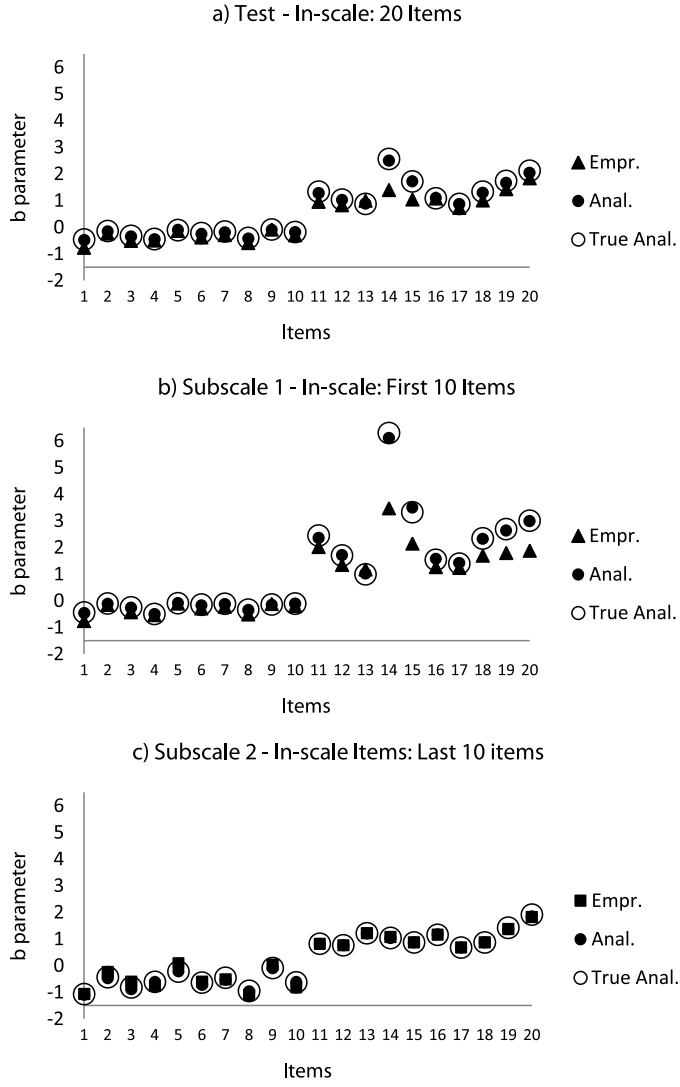

Figure 3. Analytical and Empirical difficulty parameter estimates.

test reference composite (.946). Item 9 in Subtest 1 shows a similar pattern. When projected, item 9's discrimination power with respect to the Domain 2 reference composite (1.538) is higher than its discrimination power with respect to the Domain 1 reference composite (1.055), while the item is most discriminating on the test reference composite (1.734). Visual inspection of Figure 1 indicates that, in fact, items 13 and 9 are more in alignment with the direction of the test reference composite than they are with their respective subdomain reference composites. This implies that a correct response to either of these two items is likely to require an equally weighted combination of skills defined by the two subdomains.

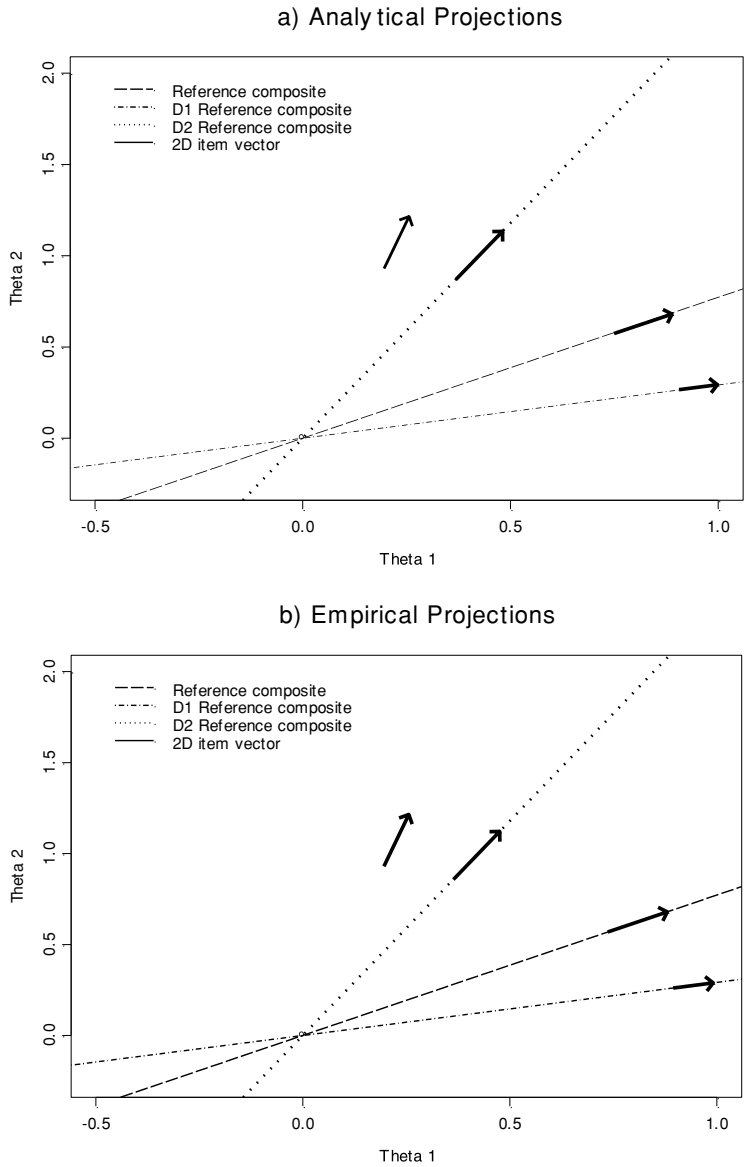

Figure 4. Analytical and Empirical unidimensional parameters for item 11.

Figure 4 shows the 2D vector presentation for one of the test items, item 11 ($a_1 = .24$, $a_2 = .86$, and $d = -1.46$) along with its unidimensional projections obtained by the Analytical procedure (Figure 4a) and the Empirical procedure (Figure 4b). It can be seen that, with both projection procedures, item 11's discrimination with respect to a given reference composite, in-scale domain or not, is proportional to the length of the item's projection on that reference composite. Gradually shortened vector lengths in this figure illustrate that the discrimination of an item on a

Table 3
Reliability Estimates

Domain of Inference	IRT Model		Projections Obtained from Simulated Data	
			Analytical Procedure	Empirical Procedure
	Items	*n*	ρ_x	ρ_x
Test	Test-scale	20	.893	.907
Subtest 1	In-scale	10	.881	.882
	In-scale + Out-scale	20	.895	.899
Subtest 2	In-scale	10	.821	.825
	In-scale + Out-scale	20	.865	.896

trait composite diminishes as the angle between the item and the unidimensional reference composite increases.

Reliabilities

Reliabilities over replications are summarized in Table 3 illustrating that the Analytical and the Empirical procedures produce very similar results. The reliabilities of the conventional IRT model *test scores* were, on average, .893 and .907 for the Analytical and the Empirical procedures, respectively. Reliabilities of the *subdomain scores* were improved with the use of out-scale items, especially for Subdomain 2. Using the Analytical procedure produced out-scale item parameters, the reliabilities of *subdomain scores* increased from .881 to .895 for Subdomain 1 and from .821 to .865 for Subdomain 2. Using the Empirical procedure produced out-scale item parameters, the reliabilities of *subdomain scores* increased from .881 to .895 for Subdomain 1 and from .821 to .865 for Subdomain 2. A careful visual inspection of Figure 2 shows one possible explanation of why Subdomain 2 scores might have benefited more from out-scale item projections than Subdomain 1 scores. It can be seen that one-third of all Subdomain 1 items are actually located near the Subdomain 2 reference composite, and when projected would help improve the test information function of Subtest 2 scores.

Real Data Application 2: Checklist Data with External Examinee Variables

Response data from administrations of two separate checklist forms used to measure the data gathering (DG) component of a clinical skills examination were analyzed by both projection procedures. Table 4 lists the 2PL 2D MIRT parameters for Checklist A and Checklist B estimated from student samples of 6,490 and 3,877. Both DG scales included dichotomously scored items from two subdomains: history taking (HT) and physical examination (PE), measuring medical students' ability to take case-relevant history and perform a PE, respectively. Both skills are considered very important in real life practice of medical doctors and are targeted to be measured as important performance components. However, the primary measure of interest with clinical skills examination is the examinee's overall proficiency in gathering relevant data, either through HT or PEs. This is an example where the

Table 4
Checklist Items

Checklist A					Checklist B				
Sub-domain	Item	a_1	a_2	d	Sub-domain	Item	a_1	a_2	d
HT	1	.217	.532	.907	HT	1	.238	.053	1.505
HT	2	.064	1.076	.803	HT	2	.060	.344	2.009
HT	3	.000	.779	–.135	HT	3	.678	.335	2.380
HT	4	.025	.838	–1.015	HT	4	.118	.000	.087
HT	5	1.429	.681	.999	HT	5	1.239	.000	1.015
HT	6	.000	.712	–.168	HT	6	1.675	.068	1.845
HT	7	.003	.157	2.745	HT	7	.356	.000	–.194
HT	8	.076	.131	–1.120	HT	8	.715	.231	3.937
PE	9	.001	.393	.241	HT	9	.299	.045	.699
PE	10	.084	.114	1.958	HT	10	.902	.262	4.070
PE	11	.513	.036	–.285	HT	11	.405	.177	.862
PE	12	.000	.320	–.030	HT	12	.960	.217	2.677
PE	13	1.464	.000	–.456	PE	13	.105	.547	.214
PE	14	.212	.000	.052	PE	14	.293	1.627	2.602
PE	15	1.275	.000	–.441	PE	15	.359	1.718	3.316
PE	16	1.170	.030	–1.994	PE	16	.041	.291	–.589
PE	17	1.190	.000	–.336	PE	17	.000	.917	–1.123
					PE	18	.000	.805	–2.112

aggregation of two mildly related measures is considered to be more important than either separate measure. Checklist A is a 17-item test with 8 HT and 9 PE items. Checklist B is an 18-item test with 12 HT and 6 PE items. Checklist A illustrates a nonorthogonal case where the correlation between the HT and the PE subdomain latent traits is .4, while Checklist B illustrates an approximately orthogonal case (trait correlations less than .1). The application summarized below compares reliabilities of test and subtest scores produced with and without projected item parameters as well as their correlations with external criteria.

Table 5 lists the reliabilities computed for Checklist A and Checklist B scores. The results show that adding out-scale items into the scoring procedure improved the reliabilities of *subdomain scores* for Checklist A. With the Analytical procedure, the reliabilities of subscale scores increased from .61 to .67 and from .71 to .75 for subdomain 1 and subdomain 2 scores, respectively. These differences correspond to a 2-item increase (a 25% increase in total subtest length) for subtest 1 and a 3-item increase (a 22% increase in total subtest length) for subtest 2. Similar to results obtained for the previous application presented, the improvement was slightly more pronounced when the Empirical procedure was used. Subscale reliabilities increased from .65 to .74 and from .69 to .76 for subdomain 1 and subdomain 2 scores, respectively. These differences correspond to a 4-item increase (a 50% increase in total subtest length) for subtest 1 and a 5-item increase (a 44% increase in total subtest length) for subtest 2.

Table 5
Reliability Estimates for the Real Data Illustrations: Checklists A and B

Domain of inference	Items	n	Analytical Procedure ρ_x	Empirical Procedure ρ_x
	Checklist A			
Test (Checklist A)	Test-scale	17	.75	.76
Subtest 1 (HT)	In-scale	8	.61	.65
Subtest 1 (HT + PE)	In-scale + Out-scale	17	.67	.74
Subtest 2 (PE)	In-scale	9	.71	.69
Subtest 2 (PE + HT)	In-scale + Out-scale	17	.75	.76
	Checklist B			
Test (Checklist B)	Test -scale	18	.54	.56
Subtest 1 (HT)	In-scale	12	.65	.59
Subtest 1 (HT + PE)	In-scale + Out-scale	6	.66	.60
Subtest 2 (PE)	In-scale	6	.55	.55
Subtest 2 (PE + HT)	In-scale + Out-scale	12	.57	.56

For Checklist B there was no improvement in subscore reliabilities when out-scale items were used. This is consistent with expectations for Checklist B; subtest composites were approximately orthogonal. The consequence of adding out-scale items into the scoring procedure was very minor (changes in reliabilities were worth less than a .5 item increase in subtest length). These results nicely illustrate that adding out-scale items to subscale item calibrations will not automatically increase test reliabilities due to added items (added out-scale items) if these items do not provide relevant information.

Validity Evidence

DG test and subtest scores obtained with both approaches were also evaluated by examining their relationship to external criterion measures: clinical knowledge (CK), communication and interpersonal skills (CIS), and documentation (DOC) scores. CK scores result from a separate standardized multiple-choice exam, while both CIS (a separate scale completed by the patients after a PE) and DOC (a patient note that each examinee completes after an encounter with the patient) scores result from the Clinical Skills Examination administrations. Table 6 and Table 7 show the resulting correlations for Checklist A and Checklist B, respectively. Both tables list correlations between external criteria and test and subtest DG scores computed by the IRT (without out-scale items) and Analytical and Empirical projection IRT models (with out-scale items) separately. Clearly, none of the external measures considered represented a measure of the outcome of interest that the DG checklists intended to measure; yet it is believed that they did represent related proficiencies. It seems reasonable to assume that, for example, an examinee seriously lacking in CK will have difficulty making the necessary decisions required to focus their patient history and PE in support of an appropriate differential diagnosis.

Table 6
Correlations with External Criteria for the Real Data Illustration: Checklist A

		Subdomain 1 (HT)		Subdomain 2 (PE)	
External Criteria	Test-scale	In-scale (HT)	In-scale + Out-scale (HT + PE)	In-scale (PE)	In-scale + Out-scale (PE + HT)
		Analytical			
CK	.18	.13	.19	.17	.19
CIS	.17	.16	.21	.17	.19
DOC	.25	.18	.25	.23	.25
		Empirical			
CK	.19	.12	.18	.17	.19
CIS	.19	.16	.21	.17	.19
DOC	.26	.17	.25	.23	.25

Table 7
Correlations with External Criteria for the Real Data Illustration: Checklist B

		Subdomain 1 (HT)		Subdomain 2 (PE)	
External Criteria	Test-scale	In-scale (HT)	In-scale + Out-scale (HT + PE)	In-scale (PE)	In-scale + Out-scale (PE + HT)
		Analytical			
CK	.18	.12	.14	.12	.14
CIS	.28	.25	.27	.15	.19
DOC	.18	.16	.17	.07	.10
		Empirical			
CK	.18	.12	.13	.12	.14
CIS	.29	.25	.26	.15	.19
DOC	.19	.16	.16	.07	.10

Table 6 reveals that the improvement observed for Checklist A might not be of practical importance for Subdomain 2. This is not surprising since CK, CIS, and DOC scores of examinees are expected to correlate higher with HT proficiency measured by Subdomain 1 items than PE proficiency measured by Subdomain 2 items. These results show that: (1) examinee subdomain-level scores produced by the Empirical and the Analytical projection IRT models correlated approximately the same with the external criteria (differences less than .01); (2) examinee test-level scores produced by the Empirical projection IRT model correlated slightly higher with the external criteria when compared to those produced by the Analytical projection IRT model; and (3) as expected, with both formulations nonorthogonal Checklist A correlations improved when projection IRT models were used while Checklist B correlations remained almost the same.

These results nicely illustrate that reliabilities of subscale scores or their correlations with external criteria will improve due to added out-scale items when and to the degree which out-of-scale items measure the target subscale construct, i.e., are construct-relevant. These results are very important not only because they show that both projection IRT models can be used to extract more information from responses to multidimensional test items, but also that this can be done without any serious drawbacks.

Summary and Discussion

The purpose of this article was to review two promising alternative projection IRT models that were developed to take advantage of construct-relevant multidimensionality. Overall, real data applications provided in this study show that both projection IRT models can be useful tools in practice for extracting useful information from multidimensional test items.[2] The results presented in this study are very important for practitioners who need alternative methods to provide meaningful subscale scores. This paper provides step-by-step illustrations of how projection IRT models can be applied to real data in practice.

Both the Analytical and Empirical projection IRT models are well studied with simulation studies. However, to date there have not been any applications of the methodology to real data where external examinee variables were available for comparison. This is the first time that projection IRT methodology is used to answer the question "how accurate are the projection IRT model-produced scores in comparison to the traditional IRT model-produced scores?" The correlations with external examinee measures (presented in Tables 6 and 7) clearly favor the projection IRT models and indicate that reliabilities and correlations with external criteria did not change for Checklist B subtest scores with approximately orthogonal subdomain traits but did improve for the Checklist A subtest scores with nonorthogonal subdomain traits. These results provide much-needed validity evidence for projection IRT methodology and suggest that both projection IRT models would help improve the precision and the accuracy of the scoring processes when and if response data are rich with component-relevant information without any serious drawbacks. When judged to be construct-relevant, collateral information extracted by the projection models can be used to improve reliability of subscale scores, which in turn can be used to provide diagnostic information about strength and weaknesses of examinees helping stakeholders to link instruction or curriculum to assessment results.

The illustrations in this paper show that perhaps the best use for collateral information would be in test settings where it would be possible to confirm that individual items discriminate among examinees with respect to more than one important test dimension. Instead of relying on the correlations between two subscale scores, it is recommended that response data are analyzed first to confirm that test items actually show a complex factorial structure. Projection IRT model applications are not recommended when (1) test items do not show a complex factorial structure or (2) observed complex factorial structure cannot be linked to test construct defined by test blueprints. Although definitive conclusions regarding the usefulness of the Empirical unidimensional projection IRT model cannot be yet made, it can be stated that the

Analytical unidimensional projection IRT model is an effective approach to reduce composite dimensionality of multidimensional response data and that the Empirical unidimensional projection IRT model is a promising procedure that can provide a close approximation to it.

Notes

[1]Splus and Mplus code can be obtained from the first author upon request.

[2]A three-dimensional application of the IRT projection methodology is also available from the first author upon request.

References

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

Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive measurement of multiple abilities*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameter estimates: Application of EM algorithm. *Psychometrika*, *46*, 443–459.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414.

Davey, T., & Hirsch, T. M. (1991, April). *Examinee discrimination as measurement properties of multidimensional tests*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

de la Torre, J. (2009). DINA Model and parameter estimation: A didactic. *Journal of Educational Measurement*, *34*, 115–13.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational Measurement*, *30*, 295–311.

Finch, H. (2009). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*, 10–26.

Goldberg, G. L., & Roswell, B. S. (2001). Are multiple measures meaningful? Lessons from a statewide performance assessment. *Applied Measurement in Education*, *14*, 125–15.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Kahraman, N., De Boeck, P., & Janssen, R. (2009). Modeling DIF in complex response data using test design strategies. *International Journal of Testing*, *9*, 151–166.

Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, *28*, 407–426.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162.

Lee, A. J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. New York, NY: Springer.

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education*, *5*, 193–211.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81–91.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661–679.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus: Statistical analysis with latent variables* (Version 2). Los Angeles, CA: Muthén & Muthén.

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

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–641). Amsterdam, The Netherlands: North-Holland.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193–203.

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

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*, 293–311.

Sheng, Y. (2010). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413–43.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

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

Wainer, H., Sheehan, M., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional response data*. Paper presented at the ONR Contractors Conference, Gatlinburg, TN.

Wang, M. (1988, April). *Measurement bias in the application of a unidimensional model to multidimensional item response data*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Wiggins, G. (1998). *Educative assessment*. San Francisco, CA: Jossey-Bass.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Zhang, J., & Wang, M. (1998, April). *Relating reported scores to latent traits in a multidimensional test*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Authors

NILUFER KAHRAMAN is a Senior Measurement Scientist, Measurement Consulting Services, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; nkahraman@nbme.org. Her primary research interests include latent trait models.

TONY THOMPSON is a Senior Research Scientist, Psychometric and Research Services, Pearson, 2510 N. Dodge St., Iowa City, IA, 52245; tony.thompson@pearson.com. His primary research interests include item response theory models and computer adaptive testing.