Article

Applied Psychological Measurement
2017, Vol. 41(7) 530–544

DOI: 10.1177/0146621617707507
journals.sagepub.com/home/apm
SAGE

Parameter Recovery in Multidimensional Item Response Theory Models Under Complexity and Nonnormality

Dubravka Svetina[1], Arturo Valdivia[1], Stephanie Underhill[1], Shenghai Dai[1], and Xiaolin Wang[1]

Abstract

Information about the psychometric properties of items can be highly useful in assessment development, for example, in item response theory (IRT) applications and computerized adaptive testing. Although literature on parameter recovery in unidimensional IRT abounds, less is known about parameter recovery in multidimensional IRT (MIRT), notably when tests exhibit complex structures or when latent traits are nonnormal. The current simulation study focuses on investigation of the effects of complex item structures and the shape of examinees' latent trait distributions on item parameter recovery in compensatory MIRT models for dichotomous items. Outcome variables included bias and root mean square error. Results indicated that when latent traits were skewed, item parameter recovery was generally adversely impacted. In addition, the presence of complexity contributed to decreases in the precision of parameter recovery, particularly for discrimination parameters along one dimension when at least one latent trait was generated as skewed.

Keywords

multidimensional IRT, complex structure, parameter estimation recovery, nonnormality

The estimation of item parameters is an essential part of item response theory (IRT) modeling, as true values of item parameters are unknown and, thus, have to be estimated. The information about psychometric properties of items, such as difficulty or discrimination, can be useful in assessment development. For example, item estimates can be used in computerized adaptive testing (CAT), where items are matched with the current estimate of an examinee's ability level (Wainer, 1990). CAT is only one example in which the accuracy of item parameter estimation is crucial for sound modeling. Accurate IRT parameters are important in other aspects in educational measurement, including IRT scoring, linking and equating, as well as in examination of differential item functioning.[1] Research examining parameter recovery in unidimensional IRT

[1]Indiana University, Bloomington, IN, USA

Corresponding Author:
Dubravka Svetina, Department of Counseling and Educational Psychology, Indiana University, W. W. Wright Education Building, 201 N. Rose Avenue, Bloomington, IN 47405-1006, USA.
Email: dsvetina@indiana.edu

models abounds (e.g., Finch, 2010, 2011; Sheng & Wikle, 2007; Way, Ansley, & Forsyth, 1988; Wiberg, 2012), yet less is known in cases where multidimensional IRT (MIRT) models are used.

Educational tests are often designed to measure several domains; for example, the Test of English as a Foreign Language (TOEFL®) contains four subdomains: listening, speaking, writing, and reading comprehension. Each domain contains items assumed to capture performance in that domain, while all items together represent the larger construct of language knowledge. Scores are reported as an overall composite and four domain subscores. In this example, if an analyst were to analyze items from each domain separately, fitting a series of unidimensional IRT models might be sufficient. However, if an analyst were to model all items simultaneously, recognizing that the subsets of items were likely correlated or that items might measure multiple dimensions (e.g., problem-solving skills within a content domain), a MIRT model with complex structure would be more appropriate.

To that end, this article aims to add to the literature related to parameter recovery within a MIRT framework by examining a number of factors that might influence precision of item parameter estimation. Specific focus is placed on (a) nonnormality of latent trait distributions, a violation of modeling assumptions, and (b) complex item structures, a design feature of the assessments. The remainder of the article is organized as follows. First, the authors situate their work in the existing literature and provide motivation for the study. Next, the simulation study design used to answer the research inquiry is described. Then, they report results, followed by the discussion, acknowledgements of the study limitations, and directions for future research.

Background

Current literature on parameter recovery investigations includes examination of a range of MIRT models, data structures, and estimation procedures (e.g., Andreis & Ferrari, 2012; Babcock, 2009; Chalmers & Flora, 2014; DeMars, 2015). More so, some factors potentially influencing the accuracy of parameter recovery have been more extensively studied than have others. For example, the number of items, sample size, and correlation between dimensions have been frequently investigated in studies on parameter recovery (e.g., Babcock, 2009; Bolt & Lall, 2003; Knol & Berger, 1991; Zhang, 2012). Considered less frequently in MIRT have been the impact of nonnormal underlying trait distributions and the presence of items demonstrating nonsimple (complex) structure (e.g., De, Ayala, & Sava-Bolesta, 1999; Kirisci, Hsu, & Yu, 2001; Seong, 1990; Stone, 1992). For example, using compensatory two-dimensional 2 parameter-logistic (2PL) and 3 parameter-logistic (3PL) MIRT models, Finch, (2010, 2011) found larger standard errors for both item difficulty and discrimination estimates when the distribution of the latent traits was skewed; a result aligned with research in IRT. Furthermore, Finch (2010) found greater bias for difficulty estimates in the skewed case than in the normal case across different estimation approaches (unweighted least squares (ULS) and robust weighted least squares.

As several scholars have previously noted, the issue of nonnormality in practice is important to consider further as nonnormal latent trait distributions negatively affected person and item parameter recovery—that is, the more skewness was present in the latent trait distribution, the more biased estimates were found to be (cf., Finch & Edwards, 2016). The authors generally assume (multivariate) normality in educational research; however, certain subpopulations (e.g., nonnative English speakers) may be better represented by latent proficiencies that are nonnormally distributed (e.g., Monroe & Cai, 2014). Furthermore, as Sass, Schmitt, and Walker (2008) argued, while in educational and psychological sciences, the assumption of a normal distribution for a latent trait may be reasonable when respondents are sampled from normally distributed populations at random, when nonrandom sampling techniques are used to obtain samples from

normally distributed populations, the potential for nonnormal trait distributions exists. Last, there exist other constructs, such as depression, pain, or gambling, that may not be normally distributed (e.g., Preston & Reise, 2014).

In addition to nonnormality, this study focuses on complexity of items on assessments. Here, the authors reflect on practices over the past decade in which educational researchers have investigated what are known as 21st-century skills.[2] Constructs of such assessments may include skills related to complex problem solving, digital proficiencies, creativity, and computer and information literacy (Ercikan & Oliveri, 2016). These complex constructs may not be as easily assessed in comparison to more traditional content-based constructs, and items that reflect only one (content) domain may not be sufficient. Examples of complex items, however, can also be found in what might be considered traditional assessments. Assessments in mathematics may include items that tap into the content areas of algebra or geometry separately; however, they can also contain items that require proficiencies in both geometry and algebra, making an item complex.

The authors believe that their study expands beyond what has been studied previously (e.g., Finch, 2010, 2011; Zhang, 2012). For example, Finch (2010) examined item parameter recovery in the presence of simple structure items only, while Finch (2011) and Zhang (2012) studied complex structures with somewhat different foci. Finch (2011) examined complex structures of a few items only, while Zhang focused on the approach to estimating mixed structure assessments. In addition, to the authors' knowledge, other studies have not examined the impact on parameter recovery of a greater number of complex items or complexity imbalance—that is, modeling both a larger number of items as complex and an uneven distribution of complex items across dimensions, which is reasonably found on assessments.

Thus, the following hypotheses were formed. First, the authors hypothesized that recovery of complex items would suffer largely in the presence of nonnormal traits. Specifically, the authors anticipated that the presence of skewed latent traits would result in poorer item parameter recovery than in conditions where normal traits were assumed. In addition, it was anticipated that balanced complexity would result in better accuracy than when only a few items were (im)balanced with respect to complexity. The authors further hypothesized that discrimination parameters would be better recovered for dimensions with which they were more strongly associated in the presence of imbalanced discriminations than when balanced discriminations were modeled. In addition, it was anticipated that the correlation between the latent traits would not have a systematic impact across conditions. Namely, it was anticipated that the correlation between the normal latent traits would impact item parameter recovery negligibly, while skewness in latent traits would negatively impact the recovery. Lastly, the authors anticipated item complexity to interact with trait correlation such that in conditions where simple structure was modeled, increases in correlation would not meaningfully affect parameter recovery, while when complexity was introduced, higher correlation would negatively affect estimation accuracy.

Method

Data Generation and Study Design

The research question regarding the accuracy of parameter estimation was addressed using a Monte Carlo simulation study. To simulate the data, 30 random draws were used from the pool of estimated parameters for 62 dichotomous items from the National Assessment of Educational Progress (NAEP) Grade 4 items on reading assessment as reported in Zhang (2012; see Appendix E in Allen, Donoghue, & Schoeps, 2001). Item responses were generated as binary

and followed a two-dimensional compensatory normal ogive model, where model parameters were manipulated as described below using functions and code written in R (R Core Development Team, 2015) by the authors.[3] In addition, it was assumed that 2,000 simulees produced responses to 30 items.

Manipulated factors. To target contexts previously understudied, several factors were manipulated, including *model type* (two levels), *correlation* between dimensions (three levels), *distribution of the latent variables* (θs; three levels), *complexity* (five levels), and *discrimination balance* (two levels). The fully crossed design yielded 180 conditions with 500 replications each.

Model type. The impact of a *c* parameter on recovery in MIRT is unclear; however, unidimensional IRT literature suggests that the presence of the *c* parameter hinders recovery. Thus, data were simulated using either a two-dimensional two-parameter normal ogive (2PNO) or a two-dimensional three-parameter normal ogive (3PNO) model.

Correlation. The literature presents wide variation in the correlations between θs, with simulated correlation levels ranging from .00 to .95 with increments ranging from .20 to .50 (e.g., Bolt & Lall, 2003; Chalmers & Flora, 2014 Finch, 2010, 2011; Zhang, 2012). Empirically, Wetzel and Hell (2014) found similar correlation values between factors, with ranges from .00 to .83. The correlations between the θs in the current study were set to .00, .40, or .70.

Latent variable distribution type. The following three levels of θ distribution types were considered in the current study: when both θs were normally distributed (normal-normal), when one θ was normal and the other skewed (normal-skewed), and when both θs were skewed (skewed-skewed). For the normal-normal case, values were drawn from bivariate normal distributions with means zero, standard deviations one, and appropriate covariance (to account for the corresponding correlation level). In conditions in which θs were generated as normal-skewed, skewed data were first simulated using a shifted Gamma distribution with unit-shape and unit-rate parameters, and normal data were simulated using a standard normal distribution. Then, correlated traits were obtained by transforming the uncorrelated data using Cholesky decomposition on the desired correlation matrix, and the resulting values were rescaled to be in the typical θ range (±3). An analogous method was used for the skewed-skewed case, but using two Gamma distributions instead.[4]

Complex structure. The baseline conditions—simple structures—included cases in which all 30 items were modeled as simple [S]: Items 1 to 15 loaded on the first dimension, and Items 16 to 30 had nonzero loadings on the second dimension. Complexity was modeled in the following ways. First, balanced conditions were examined in which an equal number of items was modeled as complex. Two levels were considered here: 20% [C20] or 40% [C40] of items were modeled as complex, resulting in three or six complex items on each dimension, respectively. Second, imbalanced complexity conditions were examined in which only one dimension contained complex items. Specifically, imbalanced complexity of 20% [C20I] or 40% [C40I] was included, so that the first dimension included three or six items that were modeled as complex, respectively, while the second dimension included only simple structure items. A fuller description and visualization of complex structures can be found in the supplementary documentation file (S1).

Discrimination balance. Two different sets of item parameters were considered. Specifically, a balanced set was considered in which items had similar strengths of association across the θs. Discrimination imbalance was modeled by utilizing item discriminations for specific items that were universally higher on the first dimension than were item discriminations on the second dimension (see ''Parameter Values'' section).[5]

Parameter values. The generating difficulty parameters ranged from −1.318 to 1.204, with $\overline{X}_d$ = −.169 (SD_d = .774). Item discriminations for *balanced conditions* ranged from .504 to 1.859, with $\overline{X}_a$ = .990 (SD_a = .365), while the lower asymptote parameters ranged from .000 to .310, with $\overline{X}_c$ = .123 (SD_c = .121). *Imbalanced discrimination conditions* were created such that one standard deviation (.365) was added to the balanced conditions' discrimination parameters for Items 1 to 15; difficulty and lower asymptote parameters remained the same across conditions. Imbalance in complexity and discrimination is further elaborated in the supplementary file (S2).

Analysis

Analyses were conducted using the function *mirt* in the mirt Version 1.9 (Chalmers, 2012) package in R. The standard expectation–maximization (EM) algorithm with fixed quadrature points (21 per dimension) and a .001 convergence threshold were used in estimation.[6] Performance was evaluated by computing the average bias and root mean square error (RMSE) for the parameter estimates across 500 replications within each condition. Bias was defined as the average difference between the estimated and true values of the parameters across J items, while the RMSE was obtained by taking the square root of the mean of squared deviations of estimated parameter values about their true values.

Results

Results are reported in two sections, beginning with bias results and proceeding to RMSE results. Within each section, results are shown for simple structure items first, followed by the results for the complex items' recovery. The authors reported results for conditions with correlated θs at .40, and later draw some general conclusions based on various levels of correlation.[7]

Section I: Bias

Simple structure items. Table 1 shows the average bias for simple structure items across 2PNO (Panel A) and 3PNO (Panel B) conditions. As Panel A shows, discrimination parameters varied in terms of recovery across complexity levels as well as θ distribution type in both balanced and imbalanced discrimination conditions. Focusing on discrimination parameter recovery along the first dimension (denoted as a_1), when θs were assumed normal and there was either no complexity [S] or complexity was balanced (C20 and C40), bias was small and positive, ranging from .014 to .020. On the other hand, when complexity was imbalanced (C20I and C40I), bias was negative, ranging from −.014 to −.042 in balanced discrimination conditions. Similar patterns were noted in comparable imbalanced discrimination conditions—namely, smaller positive bias was found in S, C20, and C40 (.023 to .027) conditions, while negative bias was found in C20I (−.009) and C40I (−.058) conditions. With normal-skewed θs, patterns across complexity were similar for balanced and imbalanced cases, with bias ranging from −.002 to .027. The largest bias was found in skewed-skewed θs cases, where it ranged from −.052 to −.081. For discrimination parameter recovery along the second dimension (denoted as a_2), in normal-normal conditions, bias was positive and small, ranging from .007 to .012. However, with at least one skewed θ, bias was larger across all conditions and ranged from −.252 to −.326.

Item location (denoted as d) bias was generally negative, suggesting that estimated item locations were somewhat smaller than their true values. As in previous cases, conditions with skewed-skewed θs and with imbalanced discrimination yielded the greatest biases.

Table 1. Mean Bias for Discrimination (a), Location (d), and Lower-Asymptote (c) Parameters for Simple Structure Items.

		Panel A: 2PNO											
		a_1			a_2			d			c		
		Latent structure											
Discrimination balance	Complexity	NN	NS	SS	NN	NS	SS	NN	NS	SS	NN	NS	SS
Balanced	S	.015	.015	−.052	.008	−.258	−.259	−.002	−.014	−.089	—	—	—
	C20I	−.014	.003	−.062	.011	−.287	−.288	−.001	−.014	−.074	—	—	—
	C40I	−.042	−.001	−.055	.010	−.324	−.326	−.002	−.018	−.070	—	—	—
	C20	.014	.013	−.061	.010	−.266	−.266	.000	−.015	−.080	—	—	—
	C40	.020	.019	−.052	.007	−.252	−.254	−.001	−.011	−.082	—	—	—
Imbalanced	S	.023	.024	−.070	.009	−.258	−.260	−.004	−.019	−.130	—	—	—
	C20I	−.009	.010	−.081	.010	−.287	−.287	−.002	−.018	−.113	—	—	—
	C40I	−.058	−.002	−.071	.009	−.324	−.325	−.004	−.021	−.106	—	—	—
	C20	.023	.022	−.080	.012	−.265	−.267	−.001	−.019	−.121	—	—	—
	C40	.027	.027	−.067	.008	−.252	−.253	−.004	−.017	−.129	—	—	—
		Panel B: 3PNO											
Balanced	S	.028	.029	.339	.020	−.060	−.006	−.002	−.341	−.873	.007	.094	.192
	C20I	−.012	.011	.334	.021	−.067	−.010	.006	−.379	−.904	.003	.105	.208
	C40I	−.041	−.004	.260	.026	−.080	−.021	−.004	−.414	−.879	.009	.115	.206
	C20	.028	.029	.355	.024	−.058	−.007	−.006	−.356	−.901	.010	.107	.215
	C40	.036	.031	.381	.017	−.045	−.002	−.003	−.358	−.914	.009	.107	.209
Imbalanced	S	.036	.036	.322	.021	−.060	−.011	.001	−.334	−.830	.004	.090	.168
	C20I	−.009	.015	.307	.021	−.071	−.012	.012	−.363	−.853	.003	.100	.183
	C40I	−.061	−.005	.257	.026	−.076	−.023	−.002	−.423	−.861	.006	.112	.189
	C20	.037	.037	.333	.025	−.065	−.010	−.002	−.342	−.847	.008	.103	.186
	C40	.041	.041	.362	.019	−−.040	−.005	−.003	−.368	−.865	.008	.108	.184

Note. Under *Complexity*, levels refer to the amount of items modeled as complex. S indicates all items are modeled as simple. C20I/C40I indicates 20%/40% of items on the first dimension are modeled as complex (imbalance in complexity w.r.t. the first dimension only), meaning 3/6 items associated with the first dimension are modeled as complex, while all items associated with the second dimension are modeled as simple. C20/C40 indicates 20%/40% of items associated with both dimensions are modeled as complex. For discrimination balance, balanced refers to situations in which items were similarly associated with their respective dimensions, while imbalanced refers to situations in which items associated with the first dimension had universally higher discrimination parameters. PNO = parameter normal ogive; NN = normal-normal (both θ distributions normal); NS = normal-skewed (first θ distribution normal and second θ skewed); SS = skewed-skewed (both θ distributions skewed).

Panel B of Table 1 shows the bias results for 3PNO conditions. Recovery of a_1 followed patterns comparable to the 2PNO conditions with one interesting departure. Namely, in skewed-skewed conditions, bias was large and positive (.257 to .381) in 3PNO conditions while it was relatively smaller and negative (−.052 to −.081) in their 2PNO counterparts. Recovery of a_2 differed in corresponding 3PNO and 2PNO conditions, except in conditions in which both θs were assumed normal; in these latter conditions, bias was relatively small and positive. In conditions with at least one skewed θ, bias was smaller in magnitude under the 3PNO model, ranging from −.080 to −.002, than under the 2PNO model in which it ranged from −.326 to −.252.

Location parameter recovery yielded similar patterns between 2PNO and 3PNO conditions, with bias becoming more extreme as distributions of θs became increasingly nonnormal. In 3PNO conditions, however, the magnitude of the bias with skewed-skewed distributions was much larger (ranging from −.914 to −.830) than in 2PNO conditions (ranging from −.130 to −.070). Bias for the c parameter followed similar patterns to those noted with discrimination and location parameters such that the smallest bias was found in conditions with normal-normal θs (never exceeding .010), somewhat larger bias was found in normal-skewed θ conditions ($\leq$ .115), and the largest bias was found in skewed-skewed conditions (.168-.215).

Complex structure items. Recovery of item parameters for complex items revealed different patterns than those noted with simple structure items, though smaller differences were found between 2PNO and 3PNO comparable conditions (see Table 2). In 2PNO conditions (Panel A), bias was generally large (.186 to .494) for a_1 in imbalanced complexity conditions with normal-normal and normal-skewed θs, while bias for a_1 was smaller in magnitude (.014 to .053) in comparable balanced complexity conditions. Conditions with skewed-skewed θs yielded no clear patterns across complexity and discrimination balance levels.

Generally, bias for location parameters in 2PNO conditions (Panel A of Table 2) was small across normal-normal and normal-skewed conditions for both balanced and imbalanced discrimination, ranging from −.029 to .044. However, with skewed-skewed θs, bias was larger and negative, ranging from −.285 to −.133. Such patterns were also found in comparable 3PNO conditions, although the magnitude of bias was larger; with skewed-skewed θs, bias exceeded .6 in magnitude. Bias for the c parameter was generally smaller for complex items (Table 2, Panel B) than for simple structure items in comparable conditions (Table 1, Panel B). Furthermore, skewed-skewed θs yielded larger bias in c than normal-normal or normal-skewed conditions.

Section II: RMSE

Simple structure items. As noted in Panel A of Table 3, RMSEs for a_1 in simple structure items in normal-normal and normal-skewed 2PNO conditions were smaller (.049-.139) than in comparable skewed-skewed conditions (.170-.237). Similar patterns were noted in 3PNO conditions (Panel B), although RMSEs were generally larger, ranging from .112 to .171 and .482 to .626, respectively. Recovery of a_2 suffered greatly in the presence of one or more skewed θs; RMSEs ranged from .294 to .445 across 2PNO and 3PNO conditions, with the highest RMSEs found in imbalanced complexity conditions. In 2PNO conditions, location parameter recovery was similar to discrimination recovery along the first dimension, with RMSEs ranging from .051 to .063 for normal-normal and normal-skewed conditions and from .116 to .191 for skewed-skewed conditions. Recovery of location parameters in 3PNO conditions worsened in the presence of skewed θs, with RMSEs for skewed-skewed θs being near or above 1. Normal-normal conditions yielded the smallest RMSEs for c parameters, ranging from .056 to .083, while the presence of discrimination and complexity imbalance yielded the poorest results.

Table 2. Mean Bias for Discrimination (a), Location (d), and Lower-Asymptote (c) Parameters for Complex Structure Items.

Discrimination balance	Complexity	a_1			a_2			d			c		
		Latent structure											
		NN	NS	SS	NN	NS	SS	NN	NS	SS	NN	NS	SS
Panel A: 2PNO													
Balanced	C20I	.333	.197	.115	—	—	—	.021	−.002	−.225	—	—	—
	C40I	.419	.186	.087	—	—	—	−.004	−.010	−.186	—	—	—
	C20	.035	.014	−.065	.031	−.412	−.398	−.009	−.017	−.154	—	—	—
	C40	.030	.019	−.109	.034	−.433	−.430	−.008	−.015	−.133	—	—	—
Imbalanced	C20I	.200	.196	.046	—	—	—	.044	−.002	−.285	—	—	—
	C40I	.494	.237	.090	—	—	—	−.008	−.016	−.259	—	—	—
	C20	.053	.029	−.081	.039	−.469	−.443	−.014	−.029	−.200	—	—	—
	C40	.045	.031	−.122	.045	−.494	−.486	−.010	−−.022	−.174	—	—	—
Panel B: 3PNO													
Balanced	C20I	.458	.254	.564	—	—	—	−.026	−.036	−.627	.004	.001	.034
	C40I	.477	.226	.647	—	—	—	−.020	−.033	−.694	−.002	.006	.103
	C20	.041	.028	.352	.033	−.396	−.215	.000	−.029	−.660	−.003	−.003	.081
	C40	.040	.037	.331	.042	−.404	−.227	−.001	−.033	−.642	.000	.005	.115
Imbalanced	C20I	.285	.242	.547	—	—	—	.027	−.015	−.676	.004	.003	.024
	C40I	.542	.264	.607	—	—	—	−.007	−.018	−.652	−.001	.001	.056
	C20	.064	.049	.341	.050	−.453	−.280	−.009	−.035	−.663	−.003	−.003	.064
	C40	.055	.050	.320	.051	−.465	−.307	−.002	−.035	−.625	.000	.004	.090

Note. Under *Complexity*, levels refer to the amount of items modeled as complex. S indicates all items are modeled as simple. C20I/C40I indicates 20%/40% of items on the first dimension are modeled as complex (imbalance in complexity w.r.t. the first dimension only), meaning 3/6 items associated with the first dimension are modeled as complex, while all items associated with the second dimension are modeled as simple. C20/C40 indicates 20%/40% of items associated with both dimensions are modeled as complex. For discrimination balance, balanced refers to situations in which items were similarly associated with their respective dimensions, while imbalanced refers to situations in which items associated with the first dimension had universally higher discrimination parameters. PNO = parameter normal ogive; NN = normal-normal (both θ distributions normal); NS = normal-skewed (first θ distribution normal and second θ skewed); SS = skewed-skewed (both θ distributions skewed).

Table 3. Mean RMSEs for Discrimination (a), Location (d), and Lower-Asymptote (c) Parameters for Simple Structure Items.

		Panel A: 2PNO											
		a_1			a_2			d			c		
		Latent structure											
Discrimination balance	Complexity	NN	NS	SS	NN	NS	SS	NN	NS	SS	NN	NS	SS
Balanced	S	.058	.057	.173	.049	.391	.394	.055	.055	.144	—	—	—
	C20I	.058	.049	.176	.053	.412	.415	.051	.052	.116	—	—	—
	C40I	.114	.059	.191	.060	.444	.445	.058	.052	.119	—	—	—
	C20	.055	.052	.170	.050	.400	.402	.057	.054	.121	—	—	—
	C40	.058	.057	.182	.047	.373	.376	.056	.053	.128	—	—	—
Imbalanced	S	.071	.072	.211	.050	.391	.394	.061	.060	.191	—	—	—
	C20I	.059	.059	.216	.053	.412	.414	.055	.056	.167	—	—	—
	C40I	.139	.065	.237	.058	.444	.445	.063	.057	.167	—	—	—
	C20	.070	.068	.211	.051	.399	.403	.061	.060	.170	—	—	—
	C40	.073	.073	.227	.047	.373	.375	.062	.059	.187	—	—	—
		Panel B: 3PNO											
Balanced	S	.130	.133	.566	.109	.305	.316	.162	.632	1.030	.070	.184	.266
	C20I	.112	.122	.589	.113	.332	.328	.160	.679	1.066	.072	.193	.282
	C40I	.149	.132	.509	.120	.360	.351	.182	.703	1.059	.083	.207	.290
	C20	.126	.126	.580	.110	.311	.309	.159	.645	1.038	.073	.202	.291
	C40	.138	.129	.626	.105	.297	.297	.172	.655	1.060	.079	.202	.290
Imbalanced	S	.146	.143	.532	.107	.308	.309	.143	.627	.970	.056	.179	.242
	C20I	.114	.127	.532	.112	.338	.326	.141	.664	.999	.058	.187	.258
	C40I	.171	.136	.482	.126	.375	.348	.157	.731	1.022	.066	.200	.271
	C20	.139	.133	.539	.111	.303	.302	.134	.624	.971	.057	.197	.263
	C40	.150	.140	.582	.106	.299	.294	.145	.663	.994	.062	.200	.264

Note. Under *Complexity*, levels refer to the amount of items modeled as complex. S indicates all items are modeled as simple. C20I/C40I indicates 20%/40% of items on the first dimension are modeled as complex (imbalance in complexity w.r.t. the first dimension only), meaning 3/6 items associated with the first dimension are modeled as complex, while all items associated with the second dimension are modeled as simple. C20/C40 indicates 20%/40% of items associated with both dimensions are modeled as complex. For discrimination balance, balanced refers to situations in which items were similarly associated with their respective dimensions, while imbalanced refers to situations in which items associated with the first dimension had universally higher discrimination parameters. PNO = parameter normal ogive; NN = normal-normal (both θ distributions normal); NS = normal-skewed (first θ distribution normal and second θ skewed); SS = skewed-skewed (both θ distributions skewed).

Complex structure items. Table 4 presents the recovery of item parameters in terms of RMSEs for complex structure items. Similar to the results for bias, an interesting pattern of recovery for a_1 was observed. Namely, in both 2PNO and 3PNO conditions, when data were modeled as normal-normal and normal-skewed, the highest RMSEs were found in imbalanced complexity conditions, where values ranged from .217 to .563 in 2PNO conditions and from .305 to .656 in 3PNO conditions. The poorest a_1 recovery was noted in 3PNO skewed-skewed conditions, and the best recovery was noted in 2PNO balanced complexity conditions.

Recovery of a_2 was best in normal-normal conditions with balanced discrimination under the 2PNO model, with RMSEs of .076 and .082 for C20 and C40 complexity conditions, respectively. Slight degradation in performance was noted in imbalanced discrimination conditions compared to their balanced discrimination counterparts under both the 2PNO and 3PNO models, and larger RMSEs were noted in conditions with at least one skewed θ.

As Panel A of Table 4 shows, the recovery of the location parameter under the 2PNO model was most successful in normal-normal and normal-skewed conditions, with RMSEs typically in the .050s and .060s. In skewed-skewed cases, for both balanced and imbalanced discrimination, RMSEs were highest in imbalanced complexity conditions, while the lowest RMSEs were found in balanced complexity conditions where 40% of items were complex. Panel B shows similar patterns, in that the RMSEs were highest in skewed-skewed conditions (typically ranging in the .700s) as compared with normal-normal or normal-skewed conditions (typically ranging in the .100s). Considering the underlying model, recovery of location parameters was better under the 2PNO than under the 3PNO model across all conditions.

RMSEs for c parameters (Panel B of Table 4) ranged from .019 to .145 across all complexity, θ distribution, and discrimination balance levels. No particular patterns were noted except that as the generating θs moved away from normal-normal, recovery typically suffered.

Discussion

The present study considered several understudied factors that potentially affect the accuracy of item parameter recovery in compensatory MIRT models. One largely consistent finding across the studied conditions and evaluation criteria was a rather clear influence of latent trait distributions. Although there were exceptions, item parameter recovery was generally poorer when one and/or both θs were generated as skewed in comparison to their normal-normal counterparts. These observations align with the study's hypotheses with regard to the adverse influence of skewed latent traits on item parameter recovery and with the (M)IRT literature (e.g., Finch, 2010; Kirisci et al., 2001; Seong, 1990; Stone, 1992).

Findings from the unidimensional IRT literature indicating that parameter recovery was generally hindered by the presence of a lower asymptote parameter was mirrored in some respects in the current study, although differences were noted as well. From both a bias and RMSE perspective, when both θs were skewed, recovery of a_1 and the location parameter was markedly poorer in the 3PNO case than in the 2PNO case. In addition, RMSEs were greater for these parameters in the 3PNO than in the 2PNO case regardless of the θ distributions. In contrast, parameter recovery was generally poorer for a_2 in the 2PNO than in the 3PNO case with at least one skewed θ.

Considering parameter estimation for complex items, recovery of a_1 was generally poorer in imbalanced complexity conditions compared with balanced complexity conditions. In addition, aligned with Finch's (2011) study, the recovery a_1 in imbalanced complexity conditions was generally poorer for complex items in comparison to their simple item counterparts. However, in contrast to Finch's findings, location and c parameter recovery was better for complex items

Table 4. Mean RMSEs for Discrimination (a), Location (d), and Lower-Asymptote (c) Parameters for Complex Structure Items.

		Panel A: 2PNO											
		a_1			a_2			d			c		
		Latent structure											
Discrimination balance	Complexity	NN	NS	SS	NN	NS	SS	NN	NS	SS	NN	NS	SS
Balanced	C20I	.349	.220	.249	—	—	—	.080	.059	.269	—	—	—
	C40I	.496	.226	.225	—	—	—	.062	.055	.241	—	—	—
	C20	.084	.078	.176	.076	.445	.431	.064	.057	.220	—	—	—
	C40	.081	.078	.179	.082	.475	.477	.065	.056	.174	—	—	—
Imbalanced	C20I	.238	.217	.212	—	—	—	.138	.067	.305	—	—	—
	C40I	.563	.273	.213	—	—	—	.066	.061	.291	—	—	—
	C20	.109	.093	.194	.088	.499	.475	.071	.069	.272	—	—	—
	C40	.100	.089	.195	.093	.528	.523	.073	.062	.219	—	—	—
		Panel B: 3PNO											
Balanced	C20I	.531	.332	.687	—	—	—	.150	.148	.715	.029	.033	.077
	C40I	.609	.305	.748	—	—	—	.139	.121	.764	.027	.033	.127
	C20	.154	.151	.447	.125	.436	.295	.131	.161	.732	.027	.043	.099
	C40	.131	.139	.414	.121	.448	.307	.112	.135	.699	.025	.039	.145
Imbalanced	C20I	.354	.313	.681	—	—	—	.163	.120	.741	.025	.023	.057
	C40I	.656	.334	.689	—	—	—	.118	.102	.699	.019	.022	.072
	C20	.180	.174	.456	.147	.490	.349	.133	.163	.740	.025	.039	.082
	C40	.148	.148	.415	.133	.501	.365	.112	.130	.682	.022	.035	.121

Note. Under *Complexity*, levels refer to the amount of items modeled as complex. S indicates all items are modeled as simple. C20I/C40I indicates 20%/40% of items on the first dimension are modeled as complex (imbalance in complexity w.r.t. the first dimension only), meaning 3/6 items associated with the first dimension are modeled as complex, while all items associated with the second dimension are modeled as simple. C20/C40 indicates 20%/40% of items associated with both dimensions are modeled as complex. For discrimination balance, balanced refers to situations in which items were similarly associated with their respective dimensions, while imbalanced refers to situations in which items associated with the first dimension had universally higher discrimination parameters. PNO = parameter normal ogive; NN = normal-normal (both θ distributions normal); NS = normal-skewed (first θ distribution normal and second θ skewed); SS = skewed-skewed (both θ distributions skewed).

than for their simple item counterparts in 3PNO conditions when at least one of the θs was skewed.[8]

An interesting observation was noted when examining recovery of complex items. Namely, when traits were assumed normal-normal, recovery of discrimination parameters was reasonable in balanced complexity conditions. However, in comparable imbalanced complexity conditions, a_1 parameters were recovered very poorly, generally resulting in larger bias/RMSEs than in conditions with at least one skewed θ.

Generally, the authors found support for their hypothesis regarding the accuracy of item discrimination parameter recovery in (im)balanced discrimination conditions. In particular, better recovery of item discriminations was observed in imbalanced conditions than in balanced conditions. With respect to the impact of correlation between the θs, results were quite consistent across the three studied levels of correlation, although some general observations can be noted. For example, consistent with Finch (2011), in unreported results, with increased correlation, bias and RMSE values were typically (but not uniformly across all conditions) larger for discrimination parameters associated with complex structures; however, the same pattern did not hold for location parameters, which tended to be the most accurately recovered at .70 correlation. Furthermore, complex items' locations were best recovered in conditions of .40 correlation, while the poorest recovery was at .70 correlation. Last, it was found that the presence of nonnormal θs tended to yield poorer recovery across item parameters; however, varied levels of latent trait correlations did not seem to uniformly impact the parameter recovery.

As with any simulation study, the design choices of this naturally limit generalizations, warranting further considerations. For example, it would be important to study parameter recovery in MIRT in the presence of more than two dimensions. As highlighted by the previously mentioned example of the vocational interest inventory considering six dimensions, five of which contained at least one complex item (Wetzel & Hell, 2014), higher dimensionality situations exist, and it is unclear how well parameters would be recovered in these contexts. Furthermore, the issue of missing data was not examined, yet it is reasonable to expect some missingness to be present in educational data. Finally, it was assumed that the correct model(s) were fit and that a priori knowledge with respect to how items loaded on the dimensions was available. In practice, this may not always be the case, and, thus, exploring the accuracy of parameter recovery within a more exploratory framework might be necessary.

In addition to these limitations, all of which should be considered in future research, the authors note that while their study focused on item parameter recovery, it would also be important to consider the impact of various factors related to person parameter recovery. This might be especially important in light of the estimation of MIRT models in which the presence of fewer individuals in the tail portions of the latent continuum may result in larger bias than in central locations with many individuals and in which the presence of skewed θ may possibly lead to poorer person parameter recovery. Furthermore, from a test development perspective, this study suggests that the presence of complex items may not be problematic in general, but imbalance with respect to the complexity may produce less precise estimates for certain types of parameters. As stated previously, the current study focused on two main issues: the presence of nonnormality (which can be thought of as a violation of modeling assumptions) and the presence of complex items (which would not be considered a modeling assumption violation). This distinction was reiterated to highlight that test developers may consider not only modeling assumption violations but other factors that may influence estimation when designing tests.

As previously eluded to, work within (unidimensional) IRT suggests the shape of latent trait distributions may impact both the recovery of person and item parameters. Unfortunately, the typical estimation methods included as default options in traditional IRT software may fall short in mitigating the impact of nonnormal θs on parameter recovery, since the true θs are often

assumed normal. Scholars have recently pointed to the importance of the choice of estimation procedure or method when investigating parameter recovery, including attention to alignment between the estimation procedure and assumptions about the underlying trait distribution(s). The authors briefly discuss below several methods that may be more appropriate in the presence of skewed θs, including Ramsey Curve and parametric and nonparametric Bayesian methods.

Ramsey Curve Item Response Theory (RC-IRT; Woods, 2006, 2008) has been proposed in IRT literature as an alternative estimation method that yields more accurate parameter estimates with skewed data as it estimates the distribution of the θs rather than assuming it is normal. Similarly, Sass et al. (2008) applied maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP) estimation approaches to examine θ estimation under conditions of normal and skewed true ability distributions and found that using Bayesian methods tended to work better with skewed data. Furthermore, Finch and Edwards (2016) found that RC-IRT and nonparametric Bayesian estimation (NBE; San Martin, Jaran, Rolin, & Mouchart, 2011) outperformed more traditional estimation procedures including marginal ML, MAP and Bayesian Markov chain Monte Carlo (with normal priors) in the presence of nonnormal θs, even though these methods tended to perform equally well in the presence of normal θs.

In addition to the above discussion regarding attention to the suitability of the estimation procedure, the authors briefly note that the purpose of the assessment—that is, broader validity considerations—should be considered in modeling as well. Namely, the presence of biased item and/or person estimates may perhaps be more tolerable in contexts considered low-stakes than in those that one might consider high-stakes. These high-stakes contexts, in which high precision and accuracy in parameter recovery are essential, are not only bound to the educational arena (e.g., CAT), but may also include contexts such as the clinical study of depression, in which proper estimation of one's depression is crucial for the development of a sound (and appropriate) treatment plan. Thus, the intended use and interpretation of scores from assessments are also important considerations when evaluating the amount of (im)precision in the estimation procedure.

As noted above, in the current study, the authors utilized the standard EM estimation procedure as the default option in the mirt package (Chalmers, 2012) in R. However, given recent developments, their results should be interpreted with some caution, as the appropriateness of EM estimation in the presence of skewed latent traits warrants further investigation. As Finch and Edwards (2016) suggested, the interaction between the distributional characteristics of both people and items should be taken into account when selecting an estimation procedure. At the same time, the choice of estimation approach may be limited by the availability of various procedures in standard software and/or by the information needed to obtain more precision (e.g., knowledge of the prior distribution in the implementation of a Bayesian method). Although a handful of studies discussed above suggest that a Bayesian estimation approach, such as NBE, may be favorable in some contexts (e.g., in the presence of skewed θs), the promise of these methods ought to be extended and evaluated in the context of multidimensionality. Choosing an estimation procedure that will yield accurate results is critical given the practical implications and wide use of IRT in education, particularly with recognition that poor parameter estimation can pose threats to appropriate inferences and sound assessment endeavors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online supplementary material are available at http://journals.sagepub.com/doi/suppl/10.1177/0146621617707507.

Notes

1. For example, when estimating person parameters, item parameters are often considered ''fixed''; thus, inaccurate estimation of item parameters may lead to inaccurate person parameter estimates. Examples of complex test structures outside of education can be found.
2. For example, Wetzel and Hell (2014) investigated the underlying dimensionality of a vocational interest inventory based on Holland's RIASEC model; the authors found that a multidimensional model with complex structure provided superior model fit than (separate) unidimensional models. The number of complex items ranged from one to five.
3. Data generation and analysis code are available upon request from the first author.
4. In normal-normal conditions, skewness for the first (and second) θ ranged from $-.004$ ($-.005$) to .005 (.007), while kurtosis ranged from 2.986 (2.986) to 3.009 (3.012), respectively. In normal-skewed conditions, skewness for the first (and second) θ ranged from $-.004$ (.712) to .008 (2.006), while kurtosis ranged from 2.987 (3.008) to 4.447 (9.088), respectively. Last, in skewed-skewed conditions, skewness for the first (and second) θ ranged from 1.399 (2.000) to 1.971 (2.005), while kurtosis ranged from 5.875 (9.007) to 8.721 (9.050), respectively.
5. Discrimination imbalance was studied because while items with high(er) discriminations are preferred, items that are not as psychometrically desired yet whose content is important may be included on an assessment.
6. The EM algorithm is considered generally as an effective estimation method with few dimensions (Chalmers, 2012); thus, the EM estimator was deemed appropriate to use in the current study. All replications converged successfully.
7. Complete tabulated results can be obtained by sending a request to the first author.
8. The authors note that a direct comparison between their and Finch's study is not entirely possible, as reported results in Finch's study were averaged across rather than broken down by the dimension. In addition, construction of item complexity differed in the two studies. These are potential factors (alongside estimation procedure choices) that may explain some of the differences in findings across the two studies.

References

Allen, N. L., Donoghue J. R., Schoeps, T. L. (2001). The NAEP 1998 technical report No. NCES 2001-509. Washington, DC: National Center for Educational Statistics. Retrieved from https://eric.ed.gov/?q=the+NAEP+1998+Technical+report&id=EJ646496

Andreis, F., & Ferrari, P. A. (2012). Missing data and parameters estimates in multidimensional item response models. *Electronic Journal of Applied Statistical Analysis*, *5*, 431-437.

Babcock, B. (2009). *Estimating a noncompensatory IRT model using a modified Metropolis algorithm* (Doctoral dissertation). Retrieved from http://purl.umn.edu/58172

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

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, *38*, 339-358.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the Nominal Response Model. *Applied Psychological Measurement*, *23*, 3-19.

DeMars, C. E. (2015). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement*, *76*, 231-257.

Ercikan, K., & Oliveri, M. O. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education*, *29*, 310-318.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*, 10-26.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*, 67-82.

Finch, H., & Edwards, J. M. (2016). Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement*, *76*, 662-684. doi:10.1177/001316441560418

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

Knol, D. L., & Berger, M. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457-477.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm. *Educational and Psychological Measurement*, *74*, 343-369.

Preston, K. S. J., & Reise, S. P. (2014). Estimating the Nominal Response Model under nonnormal conditions. *Educational and Psychological Measurement*, *74*, 377-399.

R Core Development Team. (2015). R: A language and environment for statistical computing (Version 3.2.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/

San Martin, E., Jaran, A., Rolin, J.-M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, *76*, 385-409.

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*, 65-88. doi:10.1080/08957340701796415

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299-311.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *33*, 620-639.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1-16.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, *12*, 239-252. doi:10.1177/014662168801200303

Wetzel, E., & Hell, B. (2014). Multidimensional item response theory models in vocational interest measurement: An illustration using the AIST-R. *Journal of Psychoeducational Assessment*, *32*, 342-355. doi:10.1177/0734282913508244

Wiberg, M. (2012). Can a multidimensional test be evaluated with unidimensional item response theory? *Educational Research and Evaluation*, *18*, 307-320.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*, 253.

Woods, C. M. (2008). Ramsay-curve item response theory for the three-parameter logistic item response model. *Applied Psychological Measurement*, *32*, 447-465.