MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 2, 345–361

Using Projective IRT to Evaluate the Effects of Multidimensionality on Unidimensional IRT Model Parameters

Steven P. Reise[a], Jared M. Block[a], Maxwell Mansolf[b], Mark G. Haviland[c], Benjamin D. Schalet[d], and Rachel Kimerling[e]

[a]Department of Psychology, University of California, Los Angeles, CA, USA; [b]Departments of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [c]Department of Psychiatry, Loma Linda University School of Medicine, Loma Linda, CA, USA; [d]Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands; [e]National Center for PTSD and Center for Innovation to Implementation (Ci2i), VA Palo Alto Health Care System, Menlo Park, CA, USA

ABSTRACT

The application of unidimensional IRT models requires item response data to be unidimensional. Often, however, item response data contain a dominant dimension, as well as one or more nuisance dimensions caused by content clusters. Applying a unidimensional IRT model to multidimensional data causes violations of local independence, which can vitiate IRT applications. To evaluate and, possibly, remedy the problems caused by forcing unidimensional models onto multidimensional data, we consider the creation of a projected unidimensional IRT model, where the multidimensionality caused by nuisance dimensions is controlled for by integrating them out from the model. Specifically, when item response data have a bifactor structure, one can create a unidimensional model based on projecting to the general factor. Importantly, the projected unidimensional IRT model can be used as a benchmark for comparison to a unidimensional model to judge the practical consequences of multidimensionality. Limitations of the proposed approach are detailed.

KEYWORDS

Projective IRT; unidimensionality; multidimensionality; bifactor model

Item response theory (IRT) is a set of psychometric models that characterize the relation between an individual's standing on a latent variable (the "trait") and their probability of responding to a binary or polytomous item in a specific category. In contemporary psychometrics, IRT models are often used to conduct basic psychometric analyses, to construct measures, and for administration and scoring. For these purposes, however, IRT models make strong assumptions about item response data. Most important, the application of unidimensional IRT models assumes that the item responses are *unidimensional*; that is, they are locally independent, conditional on a single latent variable (Chen & Thissen, 1997). In other words, unidimensional IRT models require that correlations among the items are completely explained by a single common factor that reflects the intended target trait assumed to underlie and cause the reliable variation in item responses. Establishing that data are unidimensional is critical, because important applications of unidimensional IRT—interpreting the estimated item and person parameters, computerized adaptive testing, scale linking and equating, and evaluating differential item functioning—depend, in large part, on the data being consistent with the unidimensionality assumption.

The unidimensionality requirement of unidimensional IRT is a strong restriction. For measures of any complex construct, item response data are typically multidimensional, not strictly unidimensional (Humphreys, 1986; Ozer, 2001; Reckase et al., 1988, Zhang, 2007). Many measures have a latent structure where there is a dominant factor running through all the items reflecting the target trait and several nuisance dimensions reflecting common variance caused by clusters of items with similar content (Reise et al.,

2013, 2023). This structure results when psychological traits have heterogeneous manifestations (Clark & Watson, 2019; Tellegen, 1991), so that clusters of items with different content are included to capture the trait's conceptual breadth. Soto and John (2017), as one example, developed a measure of the Big Five dimensions using 15 "facet" traits (see also, Comrey (1961) and Hogan and Hogan (1995)).

If one believes that psychological constructs are hierarchically structured, such as in the Hierarchical Taxonomy of Psychopathology (HITOP; Kotov et al., 2021), then any construct above the lowest rung of the hierarchy must, by definition, have heterogenous content and, thus, include a major and some minor dimensions. Heterogeneity will increase as one moves up the construct hierarchy (Morin et al., 2016). Multidimensionality due to content facets is consistent with Gustafsson and Åberg-Bengtsson (2010, Rules 1, 2, and 3, p. 108) who argue that to measure a broad trait, one must include content from lower, more conceptually narrow trait dimensions. It is also in line with Hampson, John, and Goldberg's (1986) argument that, "Trait breadth and hierarchical organization are central structural principles in personality theory and research" (p. 37).

The tension between the unidimensionality assumption of IRT models and the multidimensional nature of psychometric data required for validity creates a conundrum. Although application of unidimensional IRT is attractive, forcing multidimensional data onto unidimensional models leads to some degree of violation of the local independence assumption. Such violation leads to bias in the item and person parameter estimates, which may spoil important IRT applications.

To address the challenges of applying a unidimensional IRT model to multidimensional data, some have adopted bifactor IRT models (Cai et al., 2011; Reise et al., 2023). The general factor in the bifactor IRT model represents the intended target dimension, and multidimensionality due to content facets is controlled for through a set of orthogonal group factors. Nevertheless, bifactor models have important limitations. First, as detailed below, the item parameter estimates in bifactor IRT models are difficult to interpret (Stucky & Edelen, 2014). Second, a bifactor IRT model is not parsimonious and is difficult to use in important IRT applications, such as scale linking, differential item functioning analysis, and computerized adaptive testing. Third, the group factors in a bifactor model sometimes consist of only a few items and, as such, are considered as contributing to "nuisance" variance rather than being of substantive interest.

In the present research, we attempt to overcome these limitations by transforming the bifactor IRT model parameters. Specifically, we create a projective unidimensional IRT (PIRT) model based on projecting to a general factor in a bifactor IRT model as previously developed (Ip, 2010a, 2010b; Ip & Chen, 2014; Ip et al., 2013; Stucky et al., 2013; Stucky & Edelen, 2014). We propose that the PIRT model has two important applications. First, it can be used as a standalone unidimensional model for IRT applications (Ip & Chen, 2014). Stucky et al. (2013), for example, demonstrate how PIRT derived from a bifactor model can be used to create unidimensional short-forms with content diversity. Kim and Cho (2020) recently applied PIRT based on a bifactor model to perform true-score equating. Second, it can be used as a benchmark comparison model to judge the practical effect of model misspecification caused by imposing unidimensional IRT models on multidimensional data. This latter application is in the spirit of Crişan et al. (2017), who argue for careful consideration of the practical consequences of model violations.

To understand the problem of multidimensional data, we provide a small set of simulations to illustrate the consequences of forcing multidimensional data into a unidimensional IRT model. These illustrative simulations provide a foundation for understanding the strengths and limitations of our proposed projective IRT modeling approach.

Multidimensional data forced into unidimensional IRT models

Consider a researcher who wishes to use a multi-item scale to assess a single, broad psychological construct. We assume that the variance on each item can be decomposed into four orthogonal parts: (a) a general trait (reflected in all the items) that represents the intended target construct, (b) group component (variance shared with a subset of content similar items), (c) a specific component (reliable, systematic variance unique to the item), and (d) random error. This is the canonical bifactor structure originally proposed by Holzinger and Swineford (1937).

Consider first the upper left panel in Table 1, which displays a bifactor structure in the factor analytic metric with one general and three orthogonal group factors. Demonstration A (Demo-A, upper left panel) has 15 items, and each item has a factor loading on the general factor of 0.60 and zero loadings on the group factors. This structure represents a unidimensional model where all items are related to the general dimension equally. Given the known relation

Table 1. The effects of forcing multidimensional data onto a unidimensional model in factor analytic and IRT metrics.

	Demonstration A – unidimensional						Demonstration B – unequal group					
	True loadings				Estimated		True loadings				Estimated	
	Bifactor model				1-factor		Bifactor model				1-factor	
Item	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	$\hat{\lambda}$	$\hat{\alpha}$	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	$\hat{\lambda}$	$\hat{\alpha}$
1	.60	0	0	0	.59	1.23	.60	.70	0	0	.92	3.91
2	.60	0	0	0	.61	1.30	.60	.70	0	0	.92	3.87
3	.60	0	0	0	.60	1.28	.60	.70	0	0	.91	3.71
4	.60	0	0	0	.59	1.24	.60	.70	0	0	.91	3.85
5	.60	0	0	0	.60	1.27	.60	.70	0	0	.91	3.74
6	.60	0	0	0	.61	1.30	.60	0	.50	0	.53	1.05
7	.60	0	0	0	.61	1.30	.60	0	.50	0	.53	1.07
8	.60	0	0	0	.61	1.30	.60	0	.50	0	.52	1.05
9	.60	0	0	0	.60	1.27	.60	0	.50	0	.53	1.06
10	.60	0	0	0	.62	1.34	.60	0	.50	0	.53	1.05
11	.60	0	0	0	.61	1.30	.60	0	0	.30	.51	1.01
12	.60	0	0	0	.59	1.25	.60	0	0	.30	.50	0.97
13	.60	0	0	0	.60	1.29	.60	0	0	.30	.48	0.94
14	.60	0	0	0	.61	1.30	.60	0	0	.30	.48	0.93
15	.60	0	0	0	.61	1.31	.60	0	0	.30	.49	0.96

	Demonstration C equal group						Demonstration D cross-loadings					
	True factor loadings				Estimated		True factor loadings				Estimated	
	Bifactor model				1-factor		Bifactor model				1-factor	
Item	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	$\hat{\lambda}$	$\hat{\alpha}$	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	$\hat{\lambda}$	$\hat{\alpha}$
1	.60	.50	0	0	.65	1.46	.60	.50	0	**.10**	.68	1.58
2	.60	.50	0	0	.66	1.49	.60	.50	0	0	.63	1.39
3	.60	.50	0	0	.65	1.45	.60	.50	0	0	.65	1.45
4	.60	.50	0	0	.66	1.48	.60	.50	0	0	.64	1.42
5	.60	.50	0	0	.65	1.46	.60	.50	0	0	.64	1.42
6	.60	0	.50	0	.66	1.51	.60	**.30**	.50	0	.77	2.05
7	.60	0	.50	0	.65	1.45	.60	0	.50	0	.67	1.56
8	.60	0	.50	0	.64	1.40	.60	0	.50	0	.68	1.58
9	.60	0	.50	0	.65	1.45	.60	0	.50	0	.68	1.57
10	.60	0	.50	0	.66	1.48	.60	0	.50	0	.69	1.64
11	.60	0	0	.50	.65	1.45	.60	0	**.50**	.50	.80	2.30
12	.60	0	0	.50	.65	1.45	.60	0	0	.50	.65	1.46
13	.60	0	0	.50	.65	1.47	.60	0	0	.50	.64	1.40
14	.60	0	0	.50	.65	1.44	.60	0	0	.50	.65	1.46
15	.60	0	0	.50	.64	1.44	.60	0	0	.50	.64	1.43

Note: λ are factor loadings, $\hat{\lambda}$ are estimated factor loadings; α are IRT slopes; $\hat{\alpha}$ are estimated IRT slopes. Subscripts Gen, Grp1 … Grp3 refer to general and group factors in a bifactor model. Cross loadings in boldface type.

between the ordinal factor model and the IRT model (Kamata & Bauer, 2008; Takane & De Leeuw, 1987), the equivalent slope in a unidimensional 2-parameter logistic (2PL; Equation (1)) model is shown in Equation (2):

$$P(x_i = 1|\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))} = \frac{\exp(\alpha_i\theta + \gamma_i)}{1 + \exp(\alpha_i\theta + \gamma_i)}, \quad (1)$$

where $P(x_i = 1|\theta)$ is the probability of endorsing $(x_i = 1)$ item i as a function of a continuous, normally-distributed latent variable θ, typically, specified to be mean 0 and variance 1. The α_i parameter is a slope or "discrimination" determining the steepness of the item response function; β_i is a location parameter – the point on the latent trait where the probability of endorsing the item is 0.50; γ_i is an intercept equal to $-\alpha_i\beta_i$.

$$\alpha_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}(1.7) = \frac{.60}{\sqrt{1 - .60^2}}(1.7) = 1.275 \quad (2)$$

In Equation (2), λ_i is a factor loading in the ordinal factor model (Takane & De Leeuw, 1987); thus, in a unidimensional model with a true factor loading of 0.60, the true slope in the 2PL model is 1.275. In all simulations below, we assume factor thresholds (τ_i) and IRT intercept parameters are zero for each item. We also assume that the latent variables have a standard normal distribution.

Based on the true IRT model, we simulated 10,000 cases and estimated a 2PL model using the full-information marginal maximum likelihood method available in *mirt* (Chalmers, 2012). With this large

sample size, it is not surprising that in both the factor analytic and the IRT metric, true parameters for the one-dimensional model are well recovered. Both the estimated factor loadings and IRT slope parameters are unbiased; they reflect the general factor (i.e., the intended "target" dimension that explains the correlations among the items).

In Demonstration B (Demo-B; upper right panel), each item has the same 0.60 true factor loading on the general factor. The main difference from Demo A is that we have added three orthogonal group factors of five items each that vary in loading strength (0.70, 0.50, and 0.30), and, thus, communality. We simulated 10,000 cases and fit a 2PL (see Equation (1)) using *mirt*. In Demo B, the estimated factor loadings are *not* close to the true loadings of 0.60 nor the slopes to 1.275. In short, the loadings and IRT slopes in the unidimensional models are pulled toward the first set of five items that have the highest item intercorrelations, yielding estimated slopes of approximately 3.82 for these items.

In Demo B the latent variable no longer reflects the common variance among the items. Rather, when unidimensional models are applied to multidimensional data, the item parameters and latent variable are estimating a so-called "reference composite," or what most-commonly is termed "functional dimension" (Ip et al., 2013). The functional dimension is a kind of weighted average of the dimensions. By virtue of being optimized to explain as much common variance as possible among the items, the unidimensional model overrepresents the (greater) common variance among this first set of items.

In Demo-C, as in Demos A and B, the general factor has loadings of 0.60; the group factor loadings are now all 0.50 (and, thus, equal communality). Again, simulating 10,000 cases, the unidimensional 2PL model, when fit with *mirt* (Chalmers, 2012), has factor loadings and IRT slopes that are all too high. The unidimensional model has no way of pulling apart the common variance due to general and group factors, and, thus, it assumes that all common variance is general variance.

Demo-D has the same structure as Demo-C, but it adds an additional complexity. One item on each of the group factors is specified as having a cross-loading on another of the group factors. Specifically, Item 1 has a cross-loading of 0.10 on group factor 3. Item 6 has a cross-loading of 0.30 on group factor 1, and Item 11 has a cross-loading of 0.50 on group factor 2. As we will soon illustrate, these cross-loadings can pose challenges for our proposed projection IRT approach. For now, we observe that in the present example, when we simulated 10,000 cases and fit a 2PL model with *mirt*, the loadings and IRT slopes are more biased than in Demo-C, especially for the three items with cross-loadings. The factor loading, for example, for Item 11 is now 0.80, which is a substantial distortion of its relation with the target dimension.

Unidimensionality and IRT modeling

These demonstrations make clear that, if your measurement goal is to fit a model that captures the common variance among the items, forcing multidimensional data into unidimensional models can significantly interfere with that goal. Clearly, we need to proceed cautiously when considering a unidimensional IRT model application when we know the data are multidimensional. This raises the question, when are data "unidimensional enough," such that we do not need to be concerned about applications of the unidimensional model (i.e., when can we expect practical consequences to be trivial)?

There has been considerable research exploring conditions under which one can reasonably apply a unidimensional IRT model, despite its misspecification. As Ip (2010a) nicely summarizes this research, "If there is a predominant general factor in the data, and if the dimensions beyond that major dimension are relatively small, the presence of multidimensionality has little effect on item parameter estimates and the associated ability estimates. If, on the other hand, the data are multidimensional with strong factors beyond the first one, unidimensional parameterization results in parameter and ability estimates that are drawn toward the strongest factor in the set of item responses (this tendency is ameliorated to some extent if the factors are highly correlated)" (p. 397).

But how do we know if the first factor is "strong?" In Table 2, we consider some commonly-reported indices when applied to the data from Table 1. Our first set of indices assess the first factor strength but in slightly different ways. First, is the commonly reported eigenvalue ratio (*EVR*), reflecting the ratio of the first to second eigenvalue from the tetrachoric correlation matrix. Higher values ostensibly reflect a stronger general factor. In the data from the demonstrations, this ratio is relatively large for Demo-A. All other *EVR* values are above 3, a commonly noted "benchmark" for a "strong" general factor indicating that *EVR* would not detect issues with the data from the demonstrations.

Table 2. Fit and unidimensional enough indices for demonstrations A through D.

	Demo-A	Demo-B	Demo-C	Demo-D
Factor strength				
EVR	9.42	3.35	4.29	4.41
ECV	1	.57	.59	.58
ω_h	.89	.75	.77	.74
ω_{total}	.89	.95	.94	.94
Model fit				
M_2	68.71	16,622	12,760	12,030
df	90	90	90	90
p	.95	<.001	<.001	<.001
RMSEA	0	.133	.118	.115
SRMSR	.006	.102	.084	.079
CFI	1	.844	.857	.884

Note: *EVR* is ratio of 1st to 2nd eigenvalue; *ECV* is explained common variance; ω_h is omega hierarchical; ω_{tot} is omega total; M_2 is the limited information fit statistic; *df* is degrees of freedom; *RMSEA* is root mean squared error of approximation; *SRMSR* is standardized root mean standardized residual; *CFI* is comparative fit index.

Next are two "unidimensional enough" indices derived from applying a bifactor model (Rodriguez et al., 2016). One index that quantifies general factor strength is the explained common variance (*ECV*; Reise et al., 2013; ten Berge & Sočan, 2004). The *ECV* is the percent of common variance due to the general factor. The closer to 1.0, the more unidimensional the data. These values are all above 0.50, suggesting the general factor is stronger than the group factors. Two, Zinbarg et al. (2005, 2006) proposed coefficient omega hierarchical (ω_h), which is the percent of sum score variance due to the general factor. When ω_h is high, the reliable variance in composite scores is interpretable as reflecting the general factor. Here, values range from 0.89 (Demo-A) to 0.74 (Demo-D), suggesting that the summed scores reflect a high degree of reliable variance coming from the general factor. Should these values be considered as evidence of unidimensional enough? On the other hand, as a referent, the ω_h coefficient can be compared to an index called omega total, ω_T, which estimates the reliability due to all sources of common variance. The closer ω_h is to ω_T, the more unidimensional. There are gaps of approximately 0.20 between ω_T and ω_h for the multidimensional data sets in Demos B, C, and D, suggesting significant multidimensionality.

Finally, the lower portion of Table 2 includes tests of model fit and practical fit indices that are readily available in *mirt* (Chalmers, 2012) output. M_2 (Maydeu-Olivares & Joe, 2006) is a limited information goodness-of-fit test that is not significant in Demo-A but significant in Demos B, C, and D. The root mean square error of approximation (*RMSEA*), standardized root mean squared residual (*SRMSR*), and comparative model index (*CFI*) all indicate a "poor" fit of the unidimensional model to the multidimensional datasets and a "good" fit for the Demo-A data.

How useful are these "unidimensional enough," or "practical fit" indices and their associated benchmarks? We have three practical concerns that underlie our present proposal. First, the "benchmarks" for unidimensionality, when they exist at all, are somewhat arbitrary in the same way $p <= 0.05$ for "significant" or Cohen's $d = 0.20$ for a "small" effect are arbitrary: their values have no precise reference. Second, fit indices can reject models where the bias in parameter estimates has few, if any, practical consequences (Bonifay et al., 2015). Multidimensionality or lack of model fit does not necessarily suggest dire consequences in practical applications (Crişan et al., 2017). Indeed, the misfit in Demos B, C, and D may have very little effect on scaling individuals on the latent trait (i.e., scoring). Third, and most important, the values of the indices in Table 2 cannot be directly linked to any specific degree of bias in the estimated item parameters, trait level estimates, or standard errors, which are valuable to know when considering an IRT application.

In contrast, we argue that fitting a PIRT model can be used to directly judge the appropriateness of a unidimensional model and, thus, provide important insight to make more informed modeling decisions. Specifically, a comparison of the item parameter estimates, trait level estimates, and standard errors between a unidimensional IRT model and a better fitting PIRT model can tell you directly how wrong a unidimensional model is. Before illustrating such an application, we first describe PIRT and apply it to the three multidimensional datasets in Table 1 (Demo B, C, and D).

Projecting to the general factor in a bifactor model

PIRT is a class of models that attempt to create a unidimensional model out of a multidimensional space. Projection is based on the foundational work of Ip (2010a, 2010b) showing the conditions under which multidimensional models are equivalent to unidimensional models with local dependence. We describe projecting slopes (and intercepts) from a bifactor IRT model onto the general factor of a bifactor model, thus, creating a single dimension that is purified of nuisance dimensions. This single dimension captures the common factor running among the items (i.e., the general variance associated with the general factor).

For presentation simplicity, assume there is an estimated restricted bifactor IRT model for binary items, where items load on the general factor (*gen*) and one group factor (*grp*) as in Equation (3).

$$P(x_i = 1|\theta_{gen}, \theta_{grp}) = \frac{\exp(\alpha_{i(gen)}\theta_{gen} + \alpha_{i(grp)}\theta_{grp} + \gamma_i)}{1 + \exp(\alpha_{i(gen)}\theta_{gen} + \alpha_{i(grp)}\theta_{grp} + \gamma_i)}, \tag{3}$$

where, $P(x_i = 1|\theta_{gen}, \theta_{grp})$ is the probably of endorsing the item (i) as a function of trait levels on the general (θ_{gen}) and group (θ_{grp}) factors, $\alpha_{i(gen)}$ is the IRT slope on the general factor, $\alpha_{i(grp)}$ is the IRT slope on the group factor, and γ_i is the intercept.

Our goal is to estimate the 2PL PIRT model in Equation (4) using Equation (3).

$$P(x_i = 1|\theta_0) = \frac{\exp(\alpha_i^*\theta + \gamma_i^*)}{1 + \exp(\alpha_i^*\theta + \gamma_i^*)} \tag{4}$$

where, α_i^* is the slope in the PIRT model, and γ_i^* is the intercept in the PIRT model.

Treating the general factor in the bifactor model as the dimension to be projected to, we transform the slopes and intercepts from a bifactor IRT model to the PIRT model by using Equation (5):

$$\alpha_i^* = logit_i\left(\alpha_{i(gen)} + \frac{\alpha_{i(grp)}\rho\sigma_{i(grp)}}{\sigma_{i(gen)}}\right),$$
$$\gamma_i^* = logit_i\gamma_i, b_i^* = -\left(\frac{\gamma_i^*}{\alpha_i^*}\right), \tag{5}$$

where $logit_i = \left[k^2\alpha_{i(grp)}^2(1-\rho^2)\sigma_{i(grp)}^2 + 1\right]^{-1/2}$ and $k = \frac{16\sqrt{3}}{15\pi} = 0.59 = 1/1.7$, and b_i^* is the item location parameter in the PIRT model.

In a bifactor model, the general factor and group factor are orthogonal so that ρ (correlation among the latent factors) goes to 0. The group ($\sigma_{i(grp)}^2$) and general factor variances ($\sigma_{i(gen)}^2$) are 1.0. Ip and Chen (2014), Table 11.4 (p. 246) displayed an estimated bifactor model with Item 1 slopes of $\alpha_{gen} = 1.985$, $\alpha_{grp} = 0.840$ and $\gamma = 0.096$. Performing the calculations based on these values, we obtain the following values that match the values estimated by Ip and Chen.

$$\alpha_i^* = .8965(1.985) = 1.78,\ \gamma_i^* = .8965(0.096) = 0.086,$$
$$b_i^* = -\left(\frac{0.086}{1.78}\right) = -.0483.$$

Ip and colleagues are not the only researchers to consider projection in a bifactor context. Stucky et al. (2013), for example, evaluated the possibility of judging the effects of multidimensionality on unidimensional models by comparing the slope parameters in a unidimensional IRT model with the slope parameters from the general factor in the bifactor model. Such a comparison would appear to provide direct information on how much the unidimensional slopes are biased by the multidimensionality. Unfortunately, it is not that simple. The item parameters in an IRT bifactor model are *conditional* parameters and are not directly comparable to the *marginal* item parameters estimated in unidimensional IRT models. In the IRT bifactor model, the slopes represent the relation between the item and trait for people at the mean on all group factors (conditional). In contrast, the desired target parameters are the slopes that relate the item to the latent variable after integrating out the other dimensions (marginal). For a proper comparison, one must first transform item parameters from conditional to marginal through integrating out the multidimensionality; this is essentially the same as conducting a projection. This transformation is easy if a bifactor IRT structure is known. Using the above example:

Step 1. Convert IRT bifactor slopes from the general factor into a factor analytic correlational metric (see also Kamata & Bauer, 2008):

$$\lambda_{i(gen)} = \frac{\alpha_{i(gen)/1.7}}{\sqrt{1 + \sum_{p=1}^{P}\left(\alpha_{ip}/1.7\right)^2}}$$
$$= \frac{1.985_{/1.7}}{\sqrt{1 + \left(\frac{1.985}{1.7}\right)^2 + \left(\frac{0.840}{1.7}\right)^2}} = .723,$$

where P represents the number of dimensions. The 1.7 scaling factor is needed to convert from the factor analytic model (a normal probit model) into the IRT metric (a logistic model). Importantly, this equation clearly shows that the transformed loading on the general factor depends on the conditional slopes on all dimensions in the bifactor IRT model.

Step 2. Determine the standard deviation of residuals (σ_i) in the factor analytic metric.

$$\sigma_i = \sqrt{1 - \lambda_{i(gen)}^2} = \sqrt{1 - .723^2} = .69$$

Observe that the term under the square root sign is the residual variance, but it is based only on the loading on the general factor, thus, treating the common variance due to the group factors as "residual" for the purposes of obtaining marginal item parameters.

Step 3. Convert this factor analytic model back into the metric of a new unidimensional PIRT model.

$$\alpha_i^* = \left(\frac{\lambda_{i(gen)}}{\sigma_i}\right)1.7 = \left(\frac{.72}{.69}\right)1.7 = 1.78,$$
$$\beta_i^* = \left(\frac{-\gamma_i}{\alpha_{gen}}\right) = \left(\frac{-0.09}{1.985}\right) = -0.04, \tag{6}$$

$$\gamma_i^* = -\alpha_i^*\left(\beta_i^*\right) = -1.78(-0.0483) = 0.086$$

The above values for the marginal slopes are equal to those we calculated previously for the PIRT model. We showed both approaches so that readers can see the equivalences in two distinct literatures. Both sets of equations integrate out the nuisance dimensions through marginalization. The parameters of the PIRT model are directly comparable to parameters in a unidimensional IRT model, and, if the PIRT model is accurate, one can judge how "bad" the parameter estimates are in the unidimensional model, which is contaminated by multidimensionality.

The PIRT model can also be used as a stand-alone unidimensional IRT model for applications (Ip & Chen, 2014). The goal is for the slope parameters in the PIRT model to correctly reflect the relation between the general factor and the item responses, and the latent variable in PIRT then properly represents the common factor running through the items. When is this "ideal" most likely achieved? The answer, we believe, is whenever the bifactor model is properly specified and accurately estimated (see also, Bell et al. (2024) for a similar view on estimating accurate omega coefficients). We return to this critical topic in the discussion.

The simulated data reconsidered

In Demos B, C, and D, item parameters were biased, and the unidimensional model did not fit the data. We now examine these three multidimensional data sets under PIRT. Tables 3–5 are results for Demos B, C, and D, respectively. Table 3 is results for the data with the variable group factor loadings (Demo-B). The top portion shows the true generating factor loadings as well as the communality and the ECV_i (Stucky et al., 2013; Stucky & Edelen, 2014), the percent of common variance explained by the general factor for an item and an index of item-level unidimensionality. Next to the factor loadings are the equivalent IRT item parameters in a bifactor model. These were calculated using standard equations cited earlier; for example, using Equation (2) with P representing dimensions:

$$\alpha_{iP} = \frac{\lambda_{iP}}{\sqrt{1-\sum_{p=1}^{P}\lambda_{iP}^2}}(1.7).$$

In the far-right column are the correct projected slopes (1.275) that reflect the general variance with the nuisance dimensions integrated out. In the case of a known restricted bifactor model they are found most easily by:

$$\alpha_i^* = logit_i(\alpha_{i(gen)}) \quad \text{or} \quad \alpha_i^* = \frac{\lambda_{i(gen)}}{\sqrt{1-\lambda_{i(gen)}^2}}\ (1.7). \quad (6)$$

In the bottom panel of Table 3 are shown the estimated slopes in a unidimensional 2PL model. These slopes closely approximate those in Table 1. In the bottom middle panel (bifactor) are the estimated bifactor slopes from *mirt* (Chalmers, 2012). They are not exactly equal to the true slopes, but they are close. Finally, in the rightmost panels are shown the estimated PIRT model and the difference between the true (e.g., computed *via* the closed form equation above from true values of the bifactor model) and estimated PIRT parameters. The differences are close

Table 3. Demonstration B: true factor loadings and IRT slopes and estimated unidimensional, bifactor, and projected unidimensional slopes.

	True factor model						True IRT model				
	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	h^2	ECV_i	α_{Gen}	α_{Grp1}	α_{Grp2}	α_{Grp3}	α_{PIRT}
1	.60	.70	0	0	.85	.42	2.63	3.07	0	0	1.28
2	.60	.70	0	0	.85	.42	2.63	3.07	0	0	1.28
3	.60	.70	0	0	.85	.42	2.63	3.07	0	0	1.28
4	.60	.70	0	0	.85	.42	2.63	3.07	0	0	1.28
5	.60	.70	0	0	.85	.42	2.63	3.07	0	0	1.28
6	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
7	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
8	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
9	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
10	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
11	.60	0	0	.30	.45	.80	1.38	0	0	.69	1.28
12	.60	0	0	.30	.45	.80	1.38	0	0	.69	1.28
13	.60	0	0	.30	.45	.80	1.38	0	0	.69	1.28
14	.60	0	0	.30	.45	.80	1.38	0	0	.69	1.28
15	.60	0	0	.30	.45	.80	1.38	0	0	.69	1.28

Estimated IRT models ($N = 10{,}000$)

	Unidimensional	Bifactor				Projective unidimensional	
	$\hat{\alpha}_{Uni}$	$\hat{\alpha}_{Gen}$	$\hat{\alpha}_{Grp1}$	$\hat{\alpha}_{Grp2}$	$\hat{\alpha}_{Grp3}$	$\hat{\alpha}_{PIRT}$	bias
1	3.91	2.74	3.17	0	0	1.30	−0.02
2	3.87	2.72	3.15	0	0	1.29	−0.01
3	3.71	2.61	2.95	0	0	1.30	−0.02
4	3.85	2.73	3.22	0	0	1.28	0.00
5	3.74	2.61	3.07	0	0	1.26	0.02
6	1.05	1.54	0	1.24	0	1.25	0.03
7	1.07	1.59	0	1.35	0	1.25	0.03
8	1.05	1.6	0	1.48	0	1.21	0.07
9	1.06	1.55	0	1.34	0	1.22	0.06
10	1.05	1.52	0	1.39	0	1.18	0.10
11	1.01	1.41	0	0	0.59	1.33	−0.05
12	0.97	1.38	0	0	0.67	1.29	−0.01
13	0.94	1.3	0	0	0.62	1.22	0.06
14	0.93	1.35	0	0	0.69	1.25	0.03
15	0.96	1.35	0	0	0.66	1.26	0.02

Note. λ are true factor loadings; h^2 is communality; ECV_i is explained common variance for items; α are true slopes; and $\hat{\alpha}$ are estimated slopes. Subscripts Gen, Grp1 … Grp3 refer to general and group factors. Subscript Uni refers to the unidimensional model and PIRT refers to the projective model. Bias is $\alpha_{PIRT} - \hat{\alpha}_{PIRT}$.

to zero. In other words, the item parameters and the latent variable now reflect the common variance among the items due to the general factor. This illustrates that when data have a well-structured independent cluster bifactor solution, the projective dimension can be highly accurate even in the presence of substantial multidimensionality and poor fit to a unidimensional model.

Table 4 presents parallel results for Demo-C where the group factors all have the same loadings. The top portion provides the same information presented in Table 3, and the expected slope in the projective model remains 1.275. The bottom panel shows that the unidimensional 2PL model slopes ($\hat{\alpha}_{uni}$) are upwardly biased. In the right-hand panels are the PIRT model ($\hat{\alpha}_{pirt}$) estimates and the bias. Once again, when the multidimensionality is well structured, the PIRT model recaptures the common dimension very well. Through PIRT, we have obtained a unidimensional model in which the parameters and the latent variable reflect the common variance due to the general factor.

Table 4. Demonstration C: true factor loadings and IRT slopes and estimated unidimensional, bifactor, and projected unidimensional slopes.

	True factor model						True IRT model				
	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	h^2	ECV_i	α_{Gen}	α_{Grp1}	α_{Grp2}	α_{Grp3}	α_{PIRT}
1	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
2	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
3	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
4	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
5	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
6	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
7	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
8	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
9	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
10	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
11	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
12	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
13	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
14	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
15	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28

Estimated IRT models ($N = 10{,}000$)

	Unidimensional	Bifactor				Projective unidimensional	
	$\hat{\alpha}_{Uni}$	$\hat{\alpha}_{Gen}$	$\hat{\alpha}_{Grp1}$	$\hat{\alpha}_{Grp2}$	$\hat{\alpha}_{Grp3}$	$\hat{\alpha}_{PIRT}$	Bias
1	1.46	1.62	1.36	0	0	1.26	0.02
2	1.49	1.65	1.32	0	0	1.30	−0.02
3	1.45	1.61	1.4	0	0	1.25	0.03
4	1.48	1.61	1.23	0	0	1.31	−0.03
5	1.46	1.61	1.32	0	0	1.27	0.01
6	1.51	1.67	0	1.39	0	1.29	−0.01
7	1.45	1.58	0	1.37	0	1.23	0.05
8	1.40	1.52	0	1.38	0	1.18	0.10
9	1.45	1.58	0	1.38	0	1.23	0.05
10	1.48	1.64	0	1.47	0	1.24	0.04
11	1.45	1.58	0	0	1.3	1.25	0.03
12	1.45	1.59	0	0	1.34	1.25	0.03
13	1.47	1.64	0	0	1.44	1.25	0.03
14	1.44	1.58	0	0	1.34	1.24	0.04
15	1.44	1.58	0	0	1.34	1.24	0.04

Note. λ are true factor loadings; h^2 is communality; ECV_i is explained common variance for items; α are true slopes; and $\hat{\alpha}$ are estimates slopes. Subscripts Gen, Grp1 … Grp3 refer to general and group factors. Subscript Uni refers to the unidimensional model, and PIRT refers to the projective model. Bias is $\alpha_{PIRT} - \hat{\alpha}_{PIRT}$.

The data in Tables 3 and 4 are highly multidimensional (see Table 2), but the structure of that multidimensionality is *via* independent clusters. Table 5 is the projective results for Demo-D, where the independent cluster structure was contaminated by cross-loadings for three items (1, 6, and 11). To compute the true slope in the projected model, we now need to accommodate two group factor slopes. In the (conditional) IRT metric, an additional term in the *logit* scaling factor to accommodate two group factor slopes (*grp* and *crossload*).

$$\alpha_i^* = \left[k^2\alpha_{grp}^2 + k^2\alpha_{crossload}^2 + 1\right]^{-\frac{1}{2}}(\alpha_{gen}).$$

When that is done, the true slopes in the PIRT are still 1.275. Alternatively, in the factor loading metric, the true PIRT slope can be estimated using Equation (7):

$$\alpha_i^* = \frac{\lambda_{i(gen)}}{\sqrt{1-\lambda_{i(gen)}^2}}(1.7) = 1.275.$$

In the bottom panel of Table 5 are the estimated PIRT results and bias based on estimating a restricted bifactor model. The results are not as accurate as before, and the average bias is −0.15. The latent variable measured here is captured best by Item 11—the item with the highest communality due to the cross-loading. The problem here is not the cross-loadings, per se, but rather in the algorithm used in the estimation of restricted bifactor models. Bifactor IRT models are estimated using full-information methods (Gibbons et al., 2007; Gibbons & Hedeker, 1992) that use quadrature nodes and weights to specify a normally distributed latent variable. In multidimensional models, the number of quadrature points increases exponentially with the number of dimensions, which makes estimation challenging. The algorithm developed by Gibbons and Hedeker (1992) allows each item to load on only one group factor. Consequently, the number of dimensions per item is at most two, making estimation of canonical bifactor IRT models feasible. When there are more than two loadings per item, however, suppressing common variance associated with the cross-loading must create a distortion somewhere else in the model. In Demo-D, this distortion manifests as inflated loadings on the general and group factors for items with cross-loadings and downwardly biased loadings for other items.

Table 5. Demonstration D: true factor loadings and IRT slopes, and estimated unidimensional, bifactor, and projected unidimensional slopes.

	True factor model						True IRT model				
	λ_{Gen}	λ_{Grp1}	λ_{Grp2}	λ_{Grp3}	h^2	ECV_i	α_{Gen}	α_{Grp1}	α_{Grp2}	α_{Grp3}	α_{PIRT}
1	.60	.50	0	**.10**	.62	.58	1.65	1.38	0	0.28	1.27
2	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
3	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
4	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
5	.60	.50	0	0	.61	.59	1.63	1.36	0	0	1.28
6	.60	**.30**	.50	0	.70	.51	1.86	0.93	1.55	0	1.27
7	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
8	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
9	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
10	.60	0	.50	0	.61	.59	1.63	0	1.36	0	1.28
11	.60	0	**.50**	.50	.86	.42	2.73	0	2.27	2.27	1.28
12	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
13	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
14	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28
15	.60	0	0	.50	.61	.59	1.63	0	0	1.36	1.28

Estimated IRT models ($N = 10{,}000$)

Unidimensional	Bifactor					Projective unidimensional	
	$\hat{\alpha}_{Uni}$	$\hat{\alpha}_{Gen}$	$\hat{\alpha}_{Grp1}$	$\hat{\alpha}_{Grp2}$	$\hat{\alpha}_{Grp3}$	$\hat{\alpha}_{PIRT}$	bias
1	1.58	1.61	3.17	0	0	**1.20**	**0.07**
2	1.39	1.44	3.15	0	0	1.07	0.21
3	1.45	1.49	2.95	0	0	1.13	0.15
4	1.42	1.48	3.22	0	0	1.10	0.18
5	1.42	1.44	3.07	0	0	1.09	0.19
6	2.05	2.51	0	1.24	0	**2.48**	**−1.21**
7	1.56	1.99	0	1.35	0	1.68	−0.40
8	1.58	1.92	0	1.48	0	1.74	−0.46
9	1.57	1.88	0	1.34	0	1.72	−0.44
10	1.64	1.94	0	1.39	0	1.80	−0.52
11	2.30	2.71	0	0	0.59	**2.27**	**−0.99**
12	1.46	1.49	0	0	0.67	1.07	0.21
13	1.40	1.43	0	0	0.62	1.04	0.24
14	1.46	1.5	0	0	0.69	1.08	0.20
15	1.43	1.44	0	0	0.66	1.08	0.20

Note. λ are true factor loadings; h^2 is communality; ECV_i is explained common variance for items; α are true slopes; and $\hat{\alpha}$ are estimates slopes. Three items with cross-loadings in boldface type. Subscripts Gen, Grp1 … Grp3 refer to general and group factors. Subscript Uni refers to the unidimensional model and PIRT refers to the projective model. Bias is $\alpha_{PIRT} - \hat{\alpha}_{PIRT}$.

Assume in Demo-D (see Table 5) that we knew the true generating solution so that that bifactor model was estimated correctly. To transform this model into a PIRT model, we need only the general factor loading and the communality. We will illustrate with Item 11. Specifically, first take the square root of the communality (h_i^2) minus the general factor loading squared and treat that value as if it were the only group factor loading. We will label that the pseudo group factor, $\lambda_{i(grp')}$:

$$\lambda_{i(grp')} = \sqrt{h_i^2 - \lambda_{i(gen)}^2} = \sqrt{.86 - .36} = .71.$$

Now the implied IRT slope on the general and (pseudo) group (grp') factor is:

$$\alpha_{i(gen)} = \frac{.60}{\sqrt{1-.86}}(1.7) = 2.73 \;\; \alpha_{i(grp')} = \frac{.71}{\sqrt{1-.86}}(1.7) = 3.2,$$

and with these estimates, the slope of the PIRT model would be:

$$logit = \frac{1}{sqrt(1 + .588^2(3.212^2))} = .47,$$

$$\alpha^*_{i(pirt)} = logit(\alpha_{i(gen)}) = 1.275.$$

Thus, the estimated slope for Item 11 in the projected model based on a hypothetical, accurately-estimated bifactor model, would equal 1.275—the correct value. This demonstrates that the incorrect value of 2.27 in Table 4 is solely a product of the estimation bias caused by forcing a group factor loading to be zero, when it should have been .50. The critically important point here is that essential to accurate projection in a bifactor model is the accurate estimation of all the multidimensionality. We return to this topic in the discussion, but now demonstrate PIRT with real data.

Real data example

The data for this study are from 7,122 respondents to the 23-item PROMIS Healthcare Engagement item bank (Schalet et al., 2021). Sample details are provided in Schalet et al. (2021). For the purposes of these analyses, due to sparse data, we collapsed the two lowest (of five) categories into a single category and scored the items 0 to 3. Provisional item content is displayed in Supplemental Material. Based on previous work, the items can be partitioned into three content clusters: (1) Self-Management (SM), (2) Collaborative Communication (CC), and (3) Healthcare Navigation (HN). The results are divided into three sections: (1) restricted bifactor analysis and two comparison modeling sections evaluating (2) item parameters and (3) person parameters and standard errors. This data set was selected for the present illustration because it reasonably fits a bifactor model (see below), whereas in practice, it will be modeled and scored *via* unidimensional models.

Confirmatory item response theory

We first estimated a restricted bifactor IRT model, namely, the bifactor version of the graded response model (Samejima, 1997) using full-information marginal maximum likelihood methods. The graded response model extends the 2PL described previously to three or more response categories. In this context, there were four slope parameters α_{Gen} and $\alpha_{HN}, \alpha_{CC}, \alpha_{SM}$ (one for each of the latent factors corresponding to

each of the group factors, *HN*, *CC*, and *SM*), and three category intercept parameters γ_k ($k = 1$, 2, and 3). The model was estimated using *mirt* (Chalmers, 2012) with one intercept (or location) for each of the three boundaries between response options: 0 vs 1,2,3; 0,1 vs 2,3; and 0,1,2 vs 3. The fit statistics for the restricted bifactor model are, M_2 = 3,392 ($df = 161$) $p < 0.001$, $AIC = 328{,}542$, $SABIC = 328{,}967$, $RMSEA = 0.053$, $SRMSR = 0.036$, and CFI = .97. Coefficient ω_T value was 0.95, ω_H was 0.87, and ECV was 0.79, suggesting a strong common dimension. In the columns on the left in Table 6 are the estimated factor loadings from the restricted bifactor model. In the right panels are the estimated IRT bifactor slopes. Note that these slopes are in a conditional metric (slopes when all other factors are at their mean value, zero) making them challenging to interpret.

Evaluating the item parameters

Now we consider PIRT as a comparison model to judge the adequacy of a unidimensional model. In the left set of columns in Table 7 are the item parameter estimates for a unidimensional graded response model (Samejima, 1997) estimated with *mirt*: M_2 = 5261 ($df = 184$, $p < 0.001$), $AIC = 335{,}282$, $SABIC = 335{,}622$, $RMSEA = 0.06$, $SRMSR = 0.06$, and CFI = .95. The unidimensional model appears to fit reasonably well, but there is a large difference in M_2, *AIC, and SABIC* between the bifactor and unidimensional models, supporting the idea that the bifactor is superior. In the right set of columns in Table 7 are the PIRT parameters. In the PIRT model, the latent trait is the general factor from the bifactor model but now with marginal instead of conditional item parameters, such that the group factors have been integrated out. For almost all items, the slopes in the unidimensional model are higher or equal to the slopes in the PIRT model. This result is expected because the former are inflated by multidimensionality, whereas the latter have any multidimensionality integrated out. The slopes decreasing the most from the unidimensional model to PIRT model are items 10, 12, and 15 (all health plan items), which have the lowest ECV_i in the restricted solution of 0.70, 0.61 and 0.67, respectively. These three items all ask specifically about obtaining something from a "*provider*." On the other hand, for two items, 18 and 21, the slopes in the PIRT model are larger than in the unidimensional IRT model. Interestingly, these items are from different content domains, but both items contain the phrase "*pros and cons of treatment.*" Observe also that these items had ECV_i statistics of 0.97 and 0.98 in the restricted solution.

A relatively simple way to evaluate the effect of multidimensionality (i.e., how inaccurate the unidimensional

Table 6. Restricted bifactor solution in factor analytic and IRT metrics.

	Factor analytic					item response theory				
Item	$\hat{\lambda}_G$	$\hat{\lambda}_{HN}$	$\hat{\lambda}_{CC}$	$\hat{\lambda}_{SM}$	h^2	ECV_i	$\hat{\alpha}_{Gen}$	$\hat{\alpha}_{HN}$	$\hat{\alpha}_{CC}$	$\hat{\alpha}_{SM}$
1	.62	0	0	.24	.44	.87	1.41	0	0	0.54
2	.71	.29	0	0	.59	.86	1.89	0.78	0	0
3	.66	0	0	.34	.55	.79	1.67	0	0	0.86
4	.62	0	0	.37	.52	.74	1.54	0	0	0.91
5	.55	.27	0	0	.38	.81	1.20	0.58	0	0
6	.62	.27	0	0	.46	.84	1.43	0.63	0	0
7	.65	.24	0	0	.47	.88	1.51	0.56	0	0
8	.61	0	0	.44	.56	.65	1.56	0	0	1.13
9	.69	0	0	.33	.58	.81	1.80	0	0	0.86
10	.72	.47	0	0	.75	.70	2.44	1.59	0	0
11	.76	.33	0	0	.69	.85	2.34	1.00	0	0
12	.67	.53	0	0	.73	.61	2.18	1.74	0	0
13	.68	.42	0	0	.64	.72	1.92	1.20	0	0
14	.72	0	.40	0	.67	.76	2.14	0	1.19	0
15	.66	.46	0	0	.65	.67	1.89	1.31	0	0
16	.72	0	.38	0	.67	.78	2.15	0	1.14	0
17	.77	0	.22	0	.65	.92	2.22	0	0.64	0
18	.72	0	0	.13	.54	.97	1.81	0	0	0.32
19	.65	0	0	.44	.61	.68	1.77	0	0	1.2
20	.78	.31	0	0	.71	.87	2.47	0.97	0	0
21	.84	.12	0	0	.72	.98	2.72	0.39	0	0
22	.72	0	.39	0	.67	.77	2.15	0	1.17	0
23	.75	0	.39	0	.72	.78	2.40	0	1.26	0

Note. $\hat{\lambda}_G$ is estimated loading on general or group factors $\hat{\lambda}_{HN...SM}$ where HN is Healthcare Navigation, CC is Collaborative Communication, and SM is Self Management; h^2 is communality. $\hat{\alpha}_{Gen}$ is the estimated slope parameter for the general factor; $\hat{\alpha}_{HN...SM}$ are the estimated slope parameters for the group factors.

Table 7. Graded response model slopes and locations for the unidimensional and projective IRT models.

	Unidimensional model				PIRT model			
	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}^*$	$\hat{\beta}_1^*$	$\hat{\beta}_2^*$	$\hat{\beta}_3^*$
1	1.41	−2.63	−1.46	0.42	1.34	−2.72	−1.52	0.43
2	1.96	−1.64	−0.90	0.23	1.72	−1.76	−0.96	0.25
3	1.47	−2.67	−1.35	0.08	1.49	−2.64	−1.36	0.07
4	1.41	−1.42	−0.34	1.22	1.36	−1.46	−0.35	1.25
5	1.29	−2.35	−1.53	−0.18	1.13	−2.56	−1.67	−0.20
6	1.48	−1.64	−0.80	0.44	1.34	−1.75	−0.85	0.47
7	1.57	−0.88	0.07	1.13	1.44	−0.93	0.07	1.18
8	1.32	−1.56	−0.33	1.06	1.29	−1.59	−0.34	1.08
9	1.60	−1.90	−0.79	0.52	1.60	−1.90	−0.80	0.52
10	2.34	−1.38	−0.74	0.23	1.78	−1.56	−0.83	0.26
11	2.46	−1.63	−0.87	0.31	2.01	−1.76	−0.93	0.34
12	2.05	−1.49	−0.84	0.10	1.52	−1.73	−0.97	0.13
13	1.97	−1.47	−0.70	0.33	1.57	−1.65	−0.78	0.37
14	1.85	−1.85	−1.05	0.02	1.75	−1.91	−1.11	0.01
15	1.94	−1.74	−1.07	−0.15	1.49	−1.98	−1.22	−0.17
16	1.92	−1.73	−0.98	0.16	1.78	−1.81	−1.04	0.15
17	2.05	−2.05	−1.20	−0.06	2.08	−2.05	−1.21	−0.06
18	1.60	−2.08	−1.01	0.31	1.78	−1.98	−0.97	0.29
19	1.44	−1.82	−0.66	0.98	1.45	−1.83	−0.66	0.98
20	2.51	−1.21	−0.49	0.58	2.14	−1.28	−0.52	0.62
21	2.56	−1.53	−0.68	0.44	2.65	−1.52	−0.68	0.44
22	1.87	−1.75	−0.85	0.29	1.77	−1.81	−0.91	0.28
23	1.98	−2.48	−1.72	−0.62	1.93	−2.50	−1.75	−0.64

Note: $\hat{\alpha}$ are slope parameters and $\hat{\beta}_1$ to $\hat{\beta}_3$ are locations parameters in the unidimensional model; $\hat{\alpha}^*$ are slope parameters and β_1^* to β_3^* are locations parameters in the PIRT (Projective IRT) model.

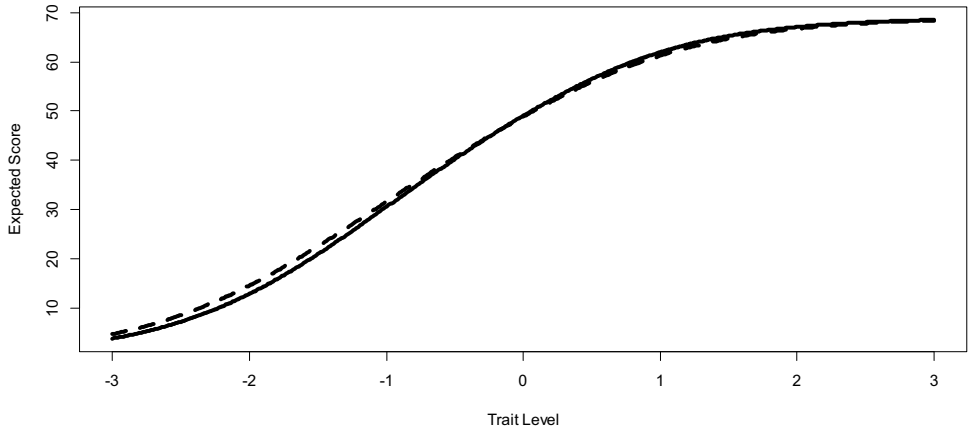

Figure 1. Test response curves for unidimensional (solid) and PIRT (dashed) models.

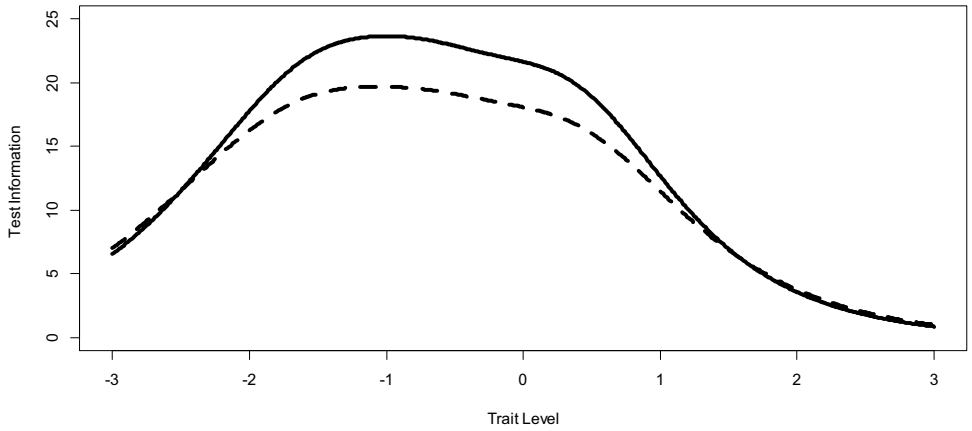

Figure 2. Test information curves in unidimensional (solid) and PIRT (dashed) models.

model is) is to compute test response curves (TRCs) and test information curves (TICs) for the two models (Ackerman et al., 2019; Stucky & Edelen, 2014). A TRC is the sum of item response curves for a scale, and it conveys the relation between trait level and expected score on a test. Differences between TRCs suggest that at a fixed trait level, there are different expected scores for the two models, reflecting potential differential test functioning. TRCs for the unidimensional and PIRT models are shown in Figure 1. TRCs for the two models show nearly perfect overlap suggesting that the overall test functioning is highly similar in the two models. Distortion caused by multidimensionality does not appear to be large enough to cause appreciable bias in the item parameters. In this sense, the data are "unidimensional enough."

On the other hand, Figure 2 displays the TICs for each model. A TIC reflects how much precision a set of items provides across the trait range. Figure 2 reveals that standard error estimates in the unidimensional model are too small. This is a clear concern for some applications of the unidimensional model.

Evaluating the person parameters

In this section, we consider the effects of multidimensionality on the estimated trait levels and standard errors. The most important potential problems with the unidimensional IRT model are that the item parameters may be biased by multidimensionality (e.g., inflated slopes), and, thus, the standard errors of trait level estimates may be too small due to the failure to take into account local dependence between items during scoring.

If we assume that the PIRT model is a closer approximation to the true model, the item parameter estimates will be more accurate than in the unidimensional model. The accurate computation of standard errors in the PIRT model, however, can be problematic (Stucky et al., 2013). If there is "hidden" local dependence in the PIRT model, adjustments may need to be made to the standard errors, depending on the strength of the group factors (Ip & Chen, 2012, 2014). Ip and Chen (2012, 2014) have recommended using sandwich estimators for scoring and computing standard errors in PIRT models. These estimators,

Table 8. Raw score to EAP trait level conversion tables for unidimensional and Lord-Wingersky 2.0 scoring.

EAP trait level estimates						Standard errors of measurement					
Score	U	LW	Score	U	LW	Score	U	LW	Score	U	L-W
0	−3.36	−3.24	36	−0.69	−0.66	0	0.44	0.49	36	0.21	0.35
1	−3.10	−2.98	37	−0.64	−0.61	1	0.39	0.45	37	0.21	0.35
2	−2.93	−2.82	38	−0.59	−0.56	2	0.36	0.43	38	0.21	0.35
3	−2.78	−2.68	39	−0.54	−0.51	3	0.34	0.41	39	0.21	0.35
4	−2.65	−2.56	40	−0.49	−0.46	4	0.32	0.40	40	0.21	0.35
5	−2.54	−2.45	41	−0.43	−0.41	5	0.30	0.39	41	0.22	0.35
6	−2.44	−2.35	42	−0.38	−0.36	6	0.29	0.38	42	0.22	0.35
7	−2.34	−2.26	43	−0.33	−0.31	7	0.28	0.37	43	0.22	0.35
8	−2.26	−2.18	44	−0.28	−0.26	8	0.27	0.37	44	0.22	0.35
9	−2.18	−2.10	45	−0.22	−0.21	9	0.26	0.36	45	0.22	0.35
10	−2.10	−2.03	46	−0.17	−0.16	10	0.26	0.36	46	0.22	0.35
11	−2.03	−1.96	47	−0.11	−0.10	11	0.25	0.36	47	0.22	0.36
12	−1.96	−1.89	48	−0.06	−0.05	12	0.24	0.36	48	0.22	0.36
13	−1.89	−1.83	49	0.00	0.00	13	0.24	0.35	49	0.22	0.36
14	−1.83	−1.76	50	0.06	0.06	14	0.23	0.35	50	0.23	0.36
15	−1.77	−1.70	51	0.12	0.12	15	0.23	0.35	51	0.23	0.36
16	−1.71	−1.65	52	0.18	0.18	16	0.23	0.35	52	0.23	0.36
17	−1.65	−1.59	53	0.24	0.24	17	0.22	0.35	53	0.23	0.36
18	−1.59	−1.53	54	0.31	0.31	18	0.22	0.35	54	0.24	0.36
19	−1.54	−1.48	55	0.38	0.37	19	0.22	0.35	55	0.24	0.36
20	−1.48	−1.43	56	0.45	0.44	20	0.22	0.35	56	0.24	0.37
21	−1.43	−1.38	57	0.52	0.51	21	0.22	0.35	57	0.25	0.37
22	−1.38	−1.33	58	0.60	0.58	22	0.21	0.35	58	0.25	0.37
23	−1.33	−1.28	59	0.68	0.66	23	0.21	0.35	59	0.26	0.38
24	−1.28	−1.23	60	0.76	0.74	24	0.21	0.35	60	0.27	0.38
25	−1.23	−1.18	61	0.86	0.83	25	0.21	0.35	61	0.28	0.39
26	−1.18	−1.13	62	0.96	0.92	26	0.21	0.35	62	0.29	0.39
27	−1.13	−1.08	63	1.07	1.03	27	0.21	0.35	63	0.31	0.40
28	−1.08	−1.03	64	1.19	1.14	28	0.21	0.35	64	0.32	0.41
29	−1.03	−0.99	65	1.32	1.27	29	0.21	0.35	65	0.35	0.43
30	−0.98	−0.94	66	1.48	1.42	30	0.21	0.35	66	0.37	0.44
31	−0.93	−0.89	67	1.66	1.58	31	0.21	0.35	67	0.40	0.47
32	−0.88	−0.85	68	1.90	1.80	32	0.21	0.35	68	0.44	0.50
33	−0.83	−0.80	69	2.24	2.14	33	0.21	0.35	69	0.52	0.56
34	−0.79	−0.75				34	0.21	0.35			
35	−0.74	−0.70				35	0.21	0.35			

Note: U is unidimensional; LW is Lord-Wingersky 2.0; score is summed score.

unfortunately, are not implemented in any standard software of which we are aware, so we rely on other solutions in the present illustration.

Specifically, recall that the latent variable in the PIRT model is equal to the general factor in the bifactor transformed to a marginal unidimensional model representation. As a consequence, one way of properly scoring individuals and estimating their standard errors is to use the bifactor model for scoring. Scoring using the bifactor model, however, can be improved by using a new updated algorithm named Lord-Wingersky 2.0 (L-W 2.0). Detailed descriptions of the technical aspects of the L-W 2.0 approach to scoring are beyond our scope (see Cai, 2015; Huang & Cai, 2021). At the heart of the method is the production of a summed score to *EAP* trait-level estimate conversion table (Orlando et al., 2000; Rosa et al., 2001; Thissen et al., 1995).

Table 8 displays the summed score to trait level estimate table and standard errors for the unidimensional model and the bifactor model (general factor) with standard errors as obtained from *flexmirt* using L-W 2.0 (Cai, 2013). As expected, scores on the general trait are more spread out in the unidimensional IRT model relative to the bifactor IRT model. The differences, however, are small. Of importance, standard errors are too small in the unidimensional IRT model. In Figure 3, we display this result in terms of conditional reliability (1—conditional error variance). Clearly, the unidimensional model overestimates the scale's precision; standard errors are lower than they should be. Overall, the marginal reliability of the trait level estimates in the unidimensional model and bifactor model with L-W 2.0 were estimated to be 0.93 and 0.86, respectively.

Discussion

Unidimensional IRT models are commonly applied to multi-item measures designed to assess individual differences on a single target construct. Such models are ideal when the item response data are unidimensional—locally independent based on a single latent factor (Chen & Thissen, 1997). When item response

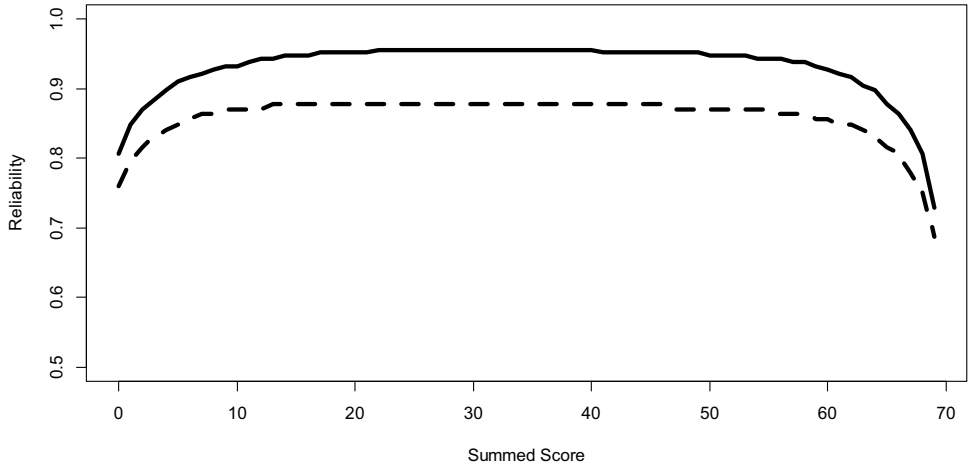

Figure 3. Reliability curves in unidimensional (solid) and PIRT model (dashed): Lord-Wingersky 2.0 scoring.

data are not unidimensional, for example, due to the inclusion of multiple content clusters (see Soto & John, 2017), unidimensional IRT models can be problematic. The central issues are that item parameters may be biased by multidimensionality (e.g., slopes too high), which results in biased trait level estimates and standard errors that may be severely biased (too small). The resulting latent trait may not represent the common variance among all the items (i.e., the intended construct), but instead may represent a reference composite or functional dimension (Ip et al., 2013; Strachan et al., 2022), whose de facto definition (based on unidimensional IRT loadings) may differ greatly from the construct definition intended for the target dimension.

Recognizing these problems caused by forcing a unidimensional model onto multidimensional data, many researchers have proposed alternative models. In the present paper, we implemented a bifactor IRT model to control for the biasing effects of multidimensionality. Specifically, we created a unidimensional PIRT model (Ip & Chen, 2014; Stucky et al., 2013) based on projecting to the general factor in a bifactor model, integrating out the group factors. To the degree that the bifactor model fits the data, the unidimensional PIRT model would be expected to provide more accurate item parameter estimates. In turn, these item parameters are in a marginal IRT metric so that they are comparable to the item parameters in the unidimensional IRT model. We used the comparison of the unidimensional IRT model with the PIRT model to judge the degree to which multidimensionality is biasing item parameters in the unidimensional model. This approach may be more useful for judging the practical significance of model misspecification, compared to relying on statistical indices of fit or first-factor strength (so called "unidimensional enough" indices). Using test response curves, the comparison between PIRT and the unidimensional model suggested that in the present empirical dataset, the unidimensional model provides essentially the same results, in terms of expected scores conditional on trait level, as the PIRT model.

PIRT is a unidimensional model with built in local dependence (Ip, 2010a, 2010b). This "hidden" local dependence is critically important for the computation of standard errors. Consequently, standard errors need to be modified if we wish to use them as a comparison model to the unidimensional model standard errors. We used the bifactor model as the comparison and used a summed score to *EAP* conversion method based on Lord-Wingersky 2.0 (Cai, 2015), available in *flexmirt* to calculate standard errors. Our results showed that standard errors in the unidimensional model were too small. In the future, alternative methods to compute standard errors are expected to be available in standard software, including Ip and Chen (2012, 2014) sandwich estimators.

Beyond its application as a comparison model, the PIRT model can also serve as a stand-alone model to be used for IRT modeling applications, such as for basic psychometric analysis and short-form creation (Stucky et al., 2013), computerized adaptive testing, equating and linking, and differential item functioning analysis. These applications rest on the unidimensionality assumption of IRT being true; that is, they all require an invariant measurement scale. In theory, the PIRT dimension should have an invariance property unobtainable otherwise (Ip & Chen, 2014; Strachan et al., 2021). In practice, a PIRT model can only be considered "purified" of multidimensionality to the degree that the multidimensionality is successfully

modeled; unmodeled multidimensionality, such as that caused by response sets, cross-loadings (e.g., Demo-D), correlated group factors, or correlated residuals, might not be adequately represented by a restricted bifactor model. It is important to report model diagnostics, such as local independence tests, and patterns of factor loadings/slopes, to strengthen the claim that the restricted bifactor is a reasonable comparison model.

Limitations of bifactor models

By accommodating multidimensional data structures, PIRT greatly expands the range of data sets that can be fit to unidimensional models. Obtaining a well-estimated and accurate restricted bifactor model is critical for the present comparison method to operate effectively.

Problems in estimating a bifactor model

There are two prevailing concerns regarding bifactor models. The first lies in the interpretation of general and group factors (Bonifay et al., 2017), especially under conditions in which the group factors cannot be considered as "exchangeable" (Eid, 2020; Eid et al., 2017; Heinrich et al., 2020). In this framework, alternative models such S − 1 type models, are preferred for representing so-called "multi-faceted" constructs (Eid, 2020). In S − 1 type models, the general factor is defined by anchoring it through a content cluster. Doebler and Doebler (2022) describe exploratory PIRT models that project to specific subdomains of interest to the investigator. These PIRT models seem highly comparable to the goals of the S − 1 model to anchor a dominant dimension though a subdomain, but we are unaware of any research that has drawn this connection.

A second concern is that there are frequent problems in the estimation of bifactor models, such as factor collapse, replicability, and unexpected negative loadings (e.g., Heinrich et al., 2020). The critical insight from the PIRT equations and our demonstration is that the slopes (or loadings) for the general and group factors should be accurate to allow for proper adjustment of the slopes in the PIRT model to represent the general factor.

The most critical factor in applying PIRT is the degree to which the structure conforms to a restricted bifactor model (Reise et al., 2011); in other words, the degree to which there is an independent cluster structure with no cross-loadings. If there is variance due to cross-loadings, the PIRT model may not properly account for multidimensionality. Zhang et al. (2023) observed that forcing cross-loadings to zero in a bifactor model resulted in identification issues, among other problems. Thus, in situations where PIRT is most needed, one must evaluate whether the data are "bifactor enough," such that an accurate multidimensional model can be estimated and then transformed. The evaluation of the "fit" of a bifactor model, prior to PIRT, is an important step.

Conclusion

To know how "wrong" you are, you must have some standard for what is "right."

Our central goal was to introduce and illustrate a PIRT model based on a bifactor model and to use this model to judge the degree to which item and person parameters (including standard errors) are wrong when multidimensional data are forced into a unidimensional model. In this concluding section, we review the interplay of the unidimensional, the bifactor, and the projective models and consider the basic question, what does one do in practice?

We first consider the comparison between the unidimensional IRT and the PIRT model. If the data are multidimensional and contain both general and nontrivial group factors, and the bifactor model fits the data better, the PIRT item parameters will be less biased and should be used in practice. The PIRT model is superior because the item parameters have been corrected for the multidimensionality.

In contrast, neither the unidimensional nor the basic PIRT model may properly control for the local independence violations caused by content clusters when estimating standard errors. For both models, standard errors will be too small. We recommend using the Lord-Wingersky 2.0 algorithm for scoring hierarchical models to produce more accurate standard errors.

The latent variable in the PIRT model is the same latent variable as the general factor in the bifactor model. Why not simply use the bifactor model in practice? The PIRT model is superior for four reasons. First, in a bifactor IRT model, the item parameters are on a conditional metric, making them more difficult to interpret substantively—they need to be marginalized to make sense and to be comparable to a unidimensional model. Second, psychometric information in multidimensional space is very complicated and difficult to interpret when there are more than two dimensions. Third, the bifactor IRT model cannot be

easily used for basic IRT applications, such as linking, computerized adaptive testing, and differential item functioning testing. Fourth, typically, there are few items on the group factors, so that the researcher could not reliably score individuals on them. The group factors tend to be merely a device to control for nuisance variance, so it is better to simply integrate them out. For these reasons, we believe the PIRT model is much easier to work with in applied settings. We, also emphasize that the PIRT model is only as good as the accuracy of the bifactor model[1] on which it is based. If the bifactor model is poorly estimated, a PIRT application cannot be justified.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant 1I01HX002317 from the United States (US) Department of Veterans Affairs Health Services Research and Development Service and from the Loma Linda University Department of Psychiatry.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

[1]More accurate is that projection IRT depends on correctly modeling the multidimensionality, regardless of its structure. Multidimensionality does not need to conform to a canonical bifactor structure to perform projection. For purposes of this research, however, we limited our work to the restricted bifactor IRT case.

References

Ackerman, T., Ma, Y., & Ip, E. (2019). Comparison of three unidimensional approaches to represent a two-dimensional latent ability space. In *Quantitative psychology: 83rd annual meeting of the psychometric society* (pp. 195–204). Springer International Publishing.

Bell, S. M., Chalmers, R. P., & Flora, D. B. (2024). The impact of measurement model misspecification on coefficient omega estimates of composite reliability. *Educational and Psychological Measurement*, *84*(1), 5–39. https://doi.org/10.1177/00131644231155804

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 504–516. https://doi.org/10.1080/10705511.2014.938596

Cai, L. (2013). *flexMIRT Version 2.0: Flexible multilevel item analysis and test scoring (computer software)*. Vector Psychometric Group LLC.

Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, *80*(2), 535–559. https://doi.org/10.1007/s11336-014-9411-3

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248. https://doi.org/10.1037/a0023350

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Comrey, A. L. (1961). Factored homogeneous item dimensions in personality research. *Educational and Psychological Measurement*, *21*(2), 417–431. https://doi.org/10.1177/001316446102100215

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*(6), 439–455. https://doi.org/10.1177/0146621617695522

Doebler, A., & Doebler, P. (2022). Rotate and project: Measurement of the intended concept with unidimensional item response theory from multidimensional ordinal items. *Multivariate Behavioral Research*, *57*(1), 40–56. https://doi.org/10.1080/00273171.2020.1794776

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child*

Psychology, *48*(7), 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19. https://doi.org/10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. https://doi.org/10.1007/BF02295430

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). American Psychological Association. https://doi.org/10.1037/12074-005

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. *Journal of Personality and Social Psychology*, *51*(1), 37–54. https://doi.org/10.1037/0022-3514.51.1.37

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, *27*(7), 1429–1447. https://doi.org/10.1177/1073191118803738

Hogan, R. T., & Hogan, J. (1995). *Manual for the Hogan Personality Inventory* (2nd ed.). Hogan Assessment Systems.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Huang, S., & Cai, L. (2021). Lord–Wingersky algorithm version 2.5 with applications. *Psychometrika*, *86*(4), 973–993. https://doi.org/10.1007/s11336-021-09785-y

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, *71*(2), 327–333. https://doi.org/10.1037/0021-9010.71.2.327

Ip, E. H. (2010a). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*(Pt 2), 395–416. https://doi.org/10.1348/000711009X466835

Ip, E. H. (2010b). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement*, *34*(7), 467–482. https://doi.org/10.1177/0146621610364975

Ip, E. H., & Chen, S. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement*, *36*(7), 581–601. https://doi.org/10.1177/0146621612452778

Ip, E. H., & Chen, S. (2014). Using projected locally dependent unidimensional models to measure multidimensional response data. In S. P. Reise, & Revicki, D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 226–251). Routledge. https://doi.org/10.4324/9781315736013

Ip, E. H., Molenberghs, G., Chen, S., Goegebeur, Y., & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*(4), 534–562. https://doi.org/10.1080/00273171.2013.796281

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Kim, K. Y., & Cho, U. H. (2020). Approximating bifactor IRT true-score equating with a projective item response model. *Applied Psychological Measurement*, *44*(3), 215–218. https://doi.org/10.1177/0146621619885903

Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The hierarchical taxonomy of psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology*, *17*(1), 83–108. https://doi.org/10.1146/annurev-clinpsy-081219-093304

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 116–139. https://doi.org/10.1080/10705511.2014.961800

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*(3), 354–359. https://doi.org/10.1037/1040-3590.12.3.354

Ozer, D. (2001). Four principles of personality assessment. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 671–688). Guilford Press.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*(3), 193–203. https://doi.org/10.1111/j.1745-3984.1988.tb00302.x

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 329–348). Guilford Press.

Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, *71*(4), 684–711. https://doi.org/10.1177/0013164410378690

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items–scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 265–304). Routledge.

Samejima, F. (1997). Graded response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.

Schalet, B. D., Reise, S. P., Zulman, D. M., Lewis, E. T., & Kimerling, R. (2021). Psychometric evaluation of a patient-reported item bank for healthcare engagement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *30*(8), 2363–2374. https://doi.org/10.1007/s11136-021-02824-2

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Strachan, T., Cho, U. H., Ackerman, T., Chen, S., de la Torre, J., & Ip, E. H. (2022). Evaluation of the linear composite conjecture for unidimensional IRT scale for multidimensional responses. *Applied Psychological Measurement*, *46*(5), 347–360. https://doi.org/10.1177/01466216221084218

Strachan, T., Cho, U. H., Kim, K. Y., Willse, J. T., Chen, S., Ip, E. H., Ackerman, T. A., & Weeks, J. P. (2021). Using a projection IRT method for vertical scaling when construct shift is present. *Journal of Educational Measurement*, *58*(2), 211–235. https://doi.org/10.1111/jedm.12278

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.) *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 201–224). Routledge. https://doi.org/10.4324/9781315736013

Stucky, B. D., Thissen, D., & Orlando Edelen, M. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, *37*(1), 41–57. https://doi.org/10.1177/0146621612462759

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl, Vol. 1. Matters of public interest; Vol. 2. Personality and psychopathology* (pp. 10–35). University of Minnesota Press.

ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625. https://doi.org/10.1007/BF02289858

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49. https://doi.org/10.1177/014662169501900105

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, B., Luo, J., Sun, T., Cao, M., & Drasgow, F. (2023). Small but nontrivial: A comparison of six strategies to handle cross-loadings in predictive models. *Multivariate Behavioral Research*, *58*(1), 115–132. https://doi.org/10.1080/00273171.2021.1957664

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*(2), 121–144. https://doi.org/10.1177/0146621605278814