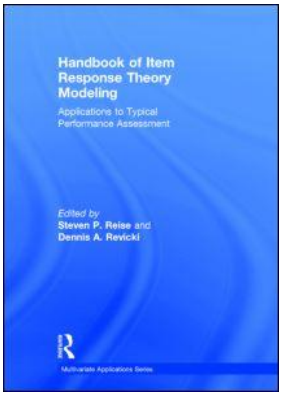

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Using Hierarchical IRT Models to Create Unidimensional Measures From Multidimensional Data

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch9
Brian D. Stucky, Maria Orlando Edelen
Published online on: 16 Dec 2014

9 Using Hierarchical IRT Models to Create Unidimensional Measures From Multidimensional Data

Brian D. Stucky and Maria Orlando Edelen

Introduction

Approaching the measurement of psychological constructs from an item analysis tradition (e.g., item response theory (IRT)) often reveals the inadequacy of single-factor or simple structure models in describing complex psychological phenomena. When test analysts closely evaluate the interrelationships among a collection of item responses it is not uncommon to find that a more complex measurement model is needed. For example, consider the seemingly well-known mental health construct of depression. Though responses to scales measuring depression are routinely treated as though there is only a single latent trait accounting for their covariance, item analysis often uncovers the presence of a single *general* dimension common to all the items, but additional *specific* dimensions that account for the uniqueness of content clusters (e.g., somatic symptoms; Irwin et al., 2010). In these modeling scenarios, some type of general hierarchical IRT model (e.g., bifactor or two-tier) may be appropriate.

This chapter reviews the structure of traditional bifactor models along with more recent advances in generalizations of bifactor model structure in the form of two-tier models that are being increasingly utilized to account for the covariance among items assessing psychological constructs. We begin by briefly reviewing the historical development of hierarchical item analysis. We then present the analytic structure of various multidimensional measurement models (for example, multidimensional IRT (MIRT; see Reckase, 2009), traditional bifactor IRT (Gibbons & Hedeker, 1992), and two-tier IRT models (Cai, 2010)) in an effort to highlight the utility and computational challenges of various modeling approaches. Next, using data from the Patient Reported Outcomes Measurement Information System (PROMIS®) adult anger, anxiety, and depression short forms (Pilkonis et al., 2011), we present a brief application of a bifactor IRT modeling process.

Finally, a general framework is offered that describes how unidimensional item subsets can be selected from a larger bifactor model (i.e., item-level dimensionality assessment). This framework may be useful in scale development scenarios where a unidimensional representation of a construct is desired, yet the data suggest multidimensionality. These item selection techniques are demonstrated by developing a hypothetical unidimensional *emotional distress* short form based on a bifactor IRT model representation of the three PROMIS® short forms. These techniques are developed as a response to the alternative of fitting unidimensional models to multidimensional data that is described in Chapter 2 of this volume.

Traditional and Restricted Bifactor Models

Holzinger and Swineford (1937) describe a bifactor pattern of loadings in which each item has a nonzero loading on exactly two factors: a general (i.e., primary) factor and a specific (i.e., secondary) factor. The item responses are conditionally independent after accounting for the general and specific dimensions. The general factor "runs through" all the items,

effectively capturing their shared content with the unifying concept. The specific factors, of which there are at least two, account for response variation that is unique or particular to item subsets. For example, the uniqueness of the specific factors may be due to content, item formatting, or other conceptual influences that result in the item responses being correlated above and beyond their association with the general dimension. Consider, for example, the PROMIS® adult anger, anxiety, and depression item banks. Though these banks were developed to measure separate constructs, the interrelationships among the items suggest the possibility of a relatively strong general dimension representative of emotional distress. In this scenario, the bifactor IRT model allows for the measurement of a unifying emotional distress dimension, along with the specific dimensions in the fashion of subscales representing distinct underlying concepts (i.e., anger, anxiety, depression). In this manner IRT conceptualizations of bifactor models have been used in psychological (Thissen & Steinberg, 2010), health (Gibbons, Rush, & Immekus, 2009; Reeve et al., 2007; Reise, Morizot, & Hays, 2007), and educational research settings (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002).

While traditional bifactor models are useful in situations where a prominent (i.e., general) dimension is comprised of multiple subscales that are of interest in their own right, *restricted* bifactor models are useful in situations where a collection of items are generally unidimensional but small clusters of items (e.g., so-called item doublets or triplets) demonstrate unintended excess covariance. Unlike traditional bifactor models, when modeling restricted bifactor models the general dimension is usually of most concern and the specific dimensions are of less interest (hence they are commonly referred to as "nuisance" dimensions). The model is *restricted* in the sense that, in addition to the general dimension, items will only have an additional specific factor loading when necessary to account for excess local dependence (LD) that is not accounted for by the single dimension. LD, which refers to a violation of the IRT assumption of unidimensionality (i.e., local independence of items), often results from clusters of items having overly similar meaning or phrasing (Steinberg, 1994). By introducing specific dimension(s) for unwanted covariation formed by item clusters, doublets, or triplets, the restricted bifactor measurement model reestablishes the IRT assumption of conditional independence.

Historical Developments in Hierarchical IRT

Bifactor models have traditionally been estimated in a confirmatory factor analytic framework. This is due in part to the bifactor's historical roots in factor analysis (see Thurstone, 1947), and in part due to the computational ease of the factor analytic approach. For ordered-categorical responses, limited-information factor analysis of the polychoric correlation matrix is relatively straightforward and additional specific factors can be added to the model without burdening the estimation. Among the advantages of this method is the abundance of model fit indices available with mean and variance adjusted weighted least squares estimation (WLSMV; Muthén, du Toit, & Spisic, 1997) or diagonally weighted least squares estimation (DWLS; Jöreskog & Sörbom, 1988) (e.g., CFI, TLI, RMSEA). However, this approach requires a complete pairwise correlation matrix, and estimates may be unstable as the sample proportions become small in the two-way frequency tables. The information is "limited" in this sense because the only available information is contained in the pairwise correlations.

In contrast to limited information factor analysis, hierarchical IRT models (including bifactor IRT models) have their roots in the full-information item analysis framework for categorical response items. Bock and Aitken (1981) first describe the relationships among item responses in a single-dimension IRT framework using marginal maximum

likelihood estimation with the expectation maximization algorithm (MML-EM), which was then extended by Bock, Gibbons, and Muraki (1988) to allow for the estimation of multiple dimensions. Following these advancements, Gibbons and Hedeker (1992) then provided a bifactor IRT model with an approach to item parameter estimation consistent with the methodology of Bock and Aitken (1981). The Gibbons and Hedeker approach capitalized on the within-item two-dimensional structure of the bifactor model to allow for the routine estimation of high-dimensional bifactor IRT models.[1]

Research Methods

The differences between various multidimensional hierarchical IRT approaches are best described by examining them in some analytic detail. In this section we review the relative advantages and disadvantages of traditional MIRT, bifactor, and two-tier IRT models. To aid in the presentation of the structure of these models we adopt a consistent notation that will be used throughout the remainder of this chapter. To begin we assume there are $i = 1, \ldots, N$ respondents and $j = 1, \ldots, n$ items with K_j response categories such that y_{ij} represents the response y from person i to item j and $\mathbf{y}$ is a n x 1 response pattern vector of person i, for simplicity we will assume i and j subscripts are implicit in the $\mathbf{y}$ response vector. For now we use Cai's (2010) notation and assume that the responses $\mathbf{y}$ are accounted for by an IRT model with an S x 1 vector of latent variables for the specific dimensions $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{is})'$ and a G x 1 vector of latent variables for the general dimensions $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iG})'$. Considering a vector of general dimensions is a departure from the traditional bifactor model but will enable analytic comparisons to more recent, generalized versions of the bifactor model.

The following marginal likelihood computations assume that specific dimensions $\boldsymbol{\xi}$ and general dimensions $\boldsymbol{\eta}$ are orthogonal to one another, and the joint probability density is simply the product of the separate univariate distributions for the general and specific dimensions:

$$f(\mathbf{y} \mid \boldsymbol{\eta}_i, \boldsymbol{\xi}_i) = f(\mathbf{y} \mid \boldsymbol{\eta}_i) f(\mathbf{y} \mid \boldsymbol{\xi}_i). \tag{9.1}$$

Otherwise known in the IRT literature as conditional independence, Equation (9.1) indicates that the joint likelihood of all general and specific dimensions can be computed from the product of the dimensions. The assumption is that the correlation between the item responses is due entirely to the presence of the general and specific dimensions, and that after these dimensions are accounted for the item responses should be uncorrelated. There is a large literature on the impact of violating the assumption of conditional independence in the IRT literature (Ackerman, 1989; Ansley & Forsyth, 1985; De Ayala, 1994; Drasgow & Parsons, 1983; Harrison, 1986; Luecht & Miller, 1992; Reckase, 1979; Tuerlinckx & De Boeck, 2001; Way, Ansley, & Forsyth, 1988); however, the topic is much less frequently addressed in regards to bifactor IRT and other MIRT models, primarily because these models are specifically fit in order to account for violations of the unidimensionality assumption.

1 It is worth noting that the development of hierarchical IRT models evolved alongside several other related methodologies. In particular the correlated-traits MIRT model shares many similarities with hierarchical IRT (for review, see Reckase, 2009). Likewise, item factor analytic approaches serve as a bridge between analytic approaches that are sometimes seen as strictly factor analytic or item response theory based. In particular, see Reise (2012) for a comparison of IRT and factor analytic approaches to the bifactor model, and Takane and de Leeuw (1987), Bartholomew and Knott (1999), Bolt (2005), and Wirth and Edwards (2007) for general reviews of the relationship between IRT and factor analytic models.

Multidimensional IRT Likelihoods

The marginal distribution of a standard multidimensional IRT model is comprised of a number of correlated dimensions that are commonly referred to in the MIRT literature simply as a vector of latent variables $\boldsymbol{\theta}$. However, to maintain a constant notation and to aid in later comparisons to the bifactor model we refer to the set of dimensions as $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, noting that this temporarily results in some unnecessary notational clutter. Assuming this arrangement the marginal distribution of the unrestricted MIRT model is:

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{y}_i \mid \eta_1 \ldots \eta_G, \xi_1 \ldots \xi_S) d\eta_1 \ldots d\eta_G, d\xi_1 \ldots d\xi_S. \tag{9.2}$$

The integrand containing the complete set of latent variables is typically approximated with computationally intensive Gaussian rectangular quadrature across a range of equally spaced nodes Q with corresponding weights (or heights) W given by the normal densities for the latent dimensions X_q:

$$L(\mathbf{y}) \approx \sum_{qG=1}^{Q} \cdots \sum_{q1=1}^{Q}, \sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} f(\mathbf{y} \mid X_{q1} \ldots X_{qS}, X_{q1} \ldots X_{qG}) W(X_{q1}) \ldots W(X_{qS}), W(X_{q1}) \ldots W(X_{qG}). \tag{9.3}$$

The "challenge of dimensionality," as Wirth and Edwards (2007) note, is that in this arrangement of the marginal likelihood the number of quadrature nodes evaluated grows exponentially with increases in the number of dimensions. That is, the joint likelihood function is constructed from Q^{G+S} quadrature nodes, and each node must be evaluated with a call to the likelihood function. As the size of the model grows (i.e., the number of dimensions), fitting an IRT model becomes computationally prohibitive. It is in part for this reason that MIRT models have been slow to make inroads in many substantive research contexts. This challenge is especially central to many scale development scenarios or initial item analyses when multiple subfactors, triplets, or doublets need to be specified to achieve conditional independence among the item responses. In these situations even if one chooses to construct the joint-likelihood from a modest number of the quadrature nodes, the high-dimensional structure means that the number of computations needed to integrate across $G+S$ dimensions is beyond even the speed and memory capacities of modern computers.

Full-Information Bifactor IRT Model Likelihoods

To address this computational challenge, Gibbons and Hedeker (1992) developed a full-information estimation technique based on bifactor model restrictions that reduces the integration needed by transforming the unwieldy multidimensional likelihood function into a series of more manageable two-dimensional functions. Cai, Yang, and Hansen (2011) refer to the process of integral evaluations as iterated integration:

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} \left\{ \prod_{s=1}^{S} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n} f(\mathbf{y} \mid \eta_1, \xi_s) \right] f(\xi_s) d\xi_s \right\} f(\eta_1) d\eta_1. \tag{9.4}$$

In this presentation of the traditional bifactor joint-likelihood, the bifactor model assumes only a single general dimension (η_1) and S possible specific dimensions. In the Gibbons and Hedeker (1992) arrangement of the likelihood function, the two-dimensional η_1 and ξ_S integral is repeatedly evaluated for each specific dimension. Again the integrals can be evaluated with numeric quadrature:

$$f(\mathbf{y}) = \sum_{q_{G=1}}^{Q}\left\{\prod_{S=1}^{S}\sum_{q=1}^{Q}\left[\prod_{j=1}^{n} f(\mathbf{y} \mid X_{G=1,}X_s)\right]W(X_s)\right\}W(X_{G=1}). \tag{9.5}$$

As can be seen, the number of two-dimensional integrations needed increases in multiples of the specific dimensions (i.e., $S \times Q^2$). This so-called dimensionality reduction allows even complex measurement models to be estimated with relative ease. For example, consider evaluating a six-dimensional bifactor measurement model with one general dimension and five specific dimensions and 21 quadrature nodes evenly spaced from –5 to 5 in multiples of 0.25, a reasonable number of points for many full-information analyses. Without the reduction in dimensionality inherent in bifactor model structure, using Equation (9.2) the number of evaluations needed for such a model is more than 85 million (21^6). Even with adaptive quadrature, which is useful in greatly reducing the number of necessary quadrature points and allows for the routine estimation of low-dimensional MIRT models (Schilling & Bock, 2005), the problem remains computationally intractable (Gibbons & Hedeker, 1992). However, using the dimensionality reduction technique (i.e., repeatedly integrating out the general dimension per specific dimension as in Equation (9.5)) the number of evaluations needed is reduced from Q^{G+S} in the unrestricted MIRT model to $S \times Q^2$ for the bifactor model, or in the present case to (5×21^2), a little more than 2,000.

Two-Tier Model Likelihoods

Application of the traditional bifactor model is limited to situations in which there is exactly one general dimension. This necessary restriction, although allowing for a computationally efficient algorithm, somewhat limits its applicability. Recently Cai (2010) has extended the earlier bifactor model work of Bock, Gibbons, and Muraki (1988) and Gibbons and Hedeker (1992) by developing the two-tier model (i.e., one tier for the general dimension(s) and one tier for the specific dimensions). The two-tier model may be viewed as a blending of traditional MIRT models and more restricted bifactor models: the model may include multiple correlated general dimensions, items may only load on one specific dimension, general and specific dimensions are orthogonal (and all specific dimensions are jointly orthogonal). Importantly, the added computational complexity of the model only occurs as the number of general dimensions increases:

$$f(\mathbf{y}_i) = \int_{-\infty_{\mathbf{\eta}G}}^{\infty}\prod_{s=1}^{S}\int_{-\infty}^{\infty}\left[\prod_{j\in I_s} f(\mathbf{y} \mid \boldsymbol{\eta}, \xi_s) f(\xi_s) d(\xi_s)\right] f(\boldsymbol{\eta}) d(\boldsymbol{\eta}). \tag{9.6}$$

The difference between the bifactor model Equation (9.4) and the two-tier model Equation (9.6) is that the two-tier model iterates the simultaneous integration of all general dimensions G and each specific dimension (only for those items pertaining to the specific dimension as indexed by I_s) across the number of specific dimensions. The conditional independence assumption in Equation (9.1) may be updated to incorporate two-tier model efficiency:

$$f(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\xi}) = \prod_{S=1}^{S}\prod_{j\in I_s} f(y_i \mid \boldsymbol{\eta}, \xi_s). \tag{9.7}$$

Cai (2010) restates the conditional independence assumption in Equation (9.1) for general MIRT models as the joint likelihood of the product of the general dimensions and each specific dimension across the set of specific dimensions. Using two-tier model restrictions

the number of function calls needed is now $S \times Q^{G+1}$ as per the general two-tier model, as opposed to Q^{G+S} calls for the standard MIRT model and $S \times Q^2$ calls for the traditional bifactor models:

$$f(\mathbf{y}) = \sum_{q_G=1}^{Q} \ldots \sum_{q_1}^{Q} \left\{ \prod_{S=1}^{S} \sum_{q=1}^{Q} \left[\prod_{j \in I_s}^{n} f(y \mid (X_1, \ldots, X_G)', X_q) W_q \right] \right\} W_{q_1} \ldots W_{q_G}. \tag{9.8}$$

This model extends the possible applications of hierarchical IRT by allowing for a variety of factor patterns. However, the two-tier model is still somewhat more restrictive than a general multidimensional IRT model. For example, the two-tier specific factors are uncorrelated with each other and with the general factors, and only one specific factor loading may be estimated per item. Nonetheless, the two-tier model is less restrictive than the standard bifactor IRT model given that multiple correlated general factors are permitted. With the same restrictions, other multidimensional IRT models like correlated MIRT models (Reckase, 2009), testlet IRT models (Wainer et al., 2007), and bifactor IRT models (Gibbons & Hedeker, 1992) conform to the two-tier model structure (see Cai, 2010 for more details).

Consider Tables 9.1 and 9.2 that present hypothetical bifactor-like models with multiple general dimensions, which we may now call two-tier models. Table 9.1 presents a model with two general dimensions each with a *doublet* item pair identified by their specific dimension slopes constrained to equality to identify the model (i.e., analogous to a one-degree of freedom residual correlation in factor analysis). A two-tier model of this variety is useful when the data suggest two *nearly* unidimensional factors, but with the presence of LD. Modeled as such, the two-tier algorithm can estimate a specific dimension within each general dimension and the correlation between the general dimensions.

Similarly, Table 9.2 presents a model that very nearly conforms to the traditional bifactor structure; however, in this case there is a doublet involving two items (4 and 5) that load on separate specific dimensions (item 4 loads on the first specific dimension, item 5 loads on the second). To satisfy the two-tier restrictions, this doublet is specified as a separate general dimension so that each item loads on no more than a single specific dimension. Specified in this manner the likelihoods for models in Tables 9.1 and 9.2 both require $2 \times Q^3$ function evaluations, still easily manageable by modern computing standards. In

Table 9.1 Example of Two-Tier Structure: Two General Dimensions, Each With a Doublet

Item	η_1	η_2	ξ_1	ξ_2
1	a_{11}		a_{11}	
2	a_{21}		a_{21}	
3	a_{31}			
4	a_{41}			
5		a_{52}		a_{52}
6		a_{62}		a_{62}
7		a_{72}		
8		a_{82}		

Note: For the purpose of identification the slope parameters in each specific dimension are constrained to equality.

Table 9.2 Example of Two-Tier Structure: Traditional Bifactor Model With a Doublet Across Both Specific Dimensions

Item	η_1	η_2	ξ_1	ξ_2
1	a_{11}		a_{11}	
2	a_{21}		a_{21}	
3	a_{31}		a_{31}	
4	a_{41}	a_{42}	a_{41}	
5	a_{51}	a_{52}		a_{52}
6	a_{61}			a_{62}
7	a_{71}			a_{72}
8	a_{81}			a_{82}

Note: For the purpose of identification the slope parameters in η_2 are constrained to equality.

fact, it seems that the two-tier reduction in integration demands makes possible all but the most complex models that could be envisioned in item and scale analysis scenarios.

Multidimensional IRT Models

Thus far our presentation of the hierarchical IRT model joint-likelihood estimation has left the item response model implicit. In this section we fill in that gap by briefly discussing item parameter interpretation when using multidimensional IRT models (Muraki & Carlson, 1993; Reckase, 2009). Although this discussion is applicable to any IRT model, we use Samejima's (1969) graded response model (GRM) as generalized to be applicable for MIRT. Later in this chapter we describe the multidimensional GRM and discuss how to interpret the dimensionality based on the (now conditional) slope parameters. This model describes the probability of responding in item response category k or higher, where $k = 0, 1, \ldots, m$. Using general two-tier structure, the multidimensional GRM for the two-tier model describes the cumulative response category probabilities as:

$$p^*(y = 1 \mid \boldsymbol{\eta}, \xi_s) = \frac{1}{1 + \exp\left[-\left(\boldsymbol{a}'\boldsymbol{\eta} + a_s\xi_s + c_1\right)\right]}, \tag{9.9}$$

for responses in the first category. Response probabilities in the last category m are defined as:

$$p^*(y = m \mid \boldsymbol{\eta}, \xi_s) = \frac{1}{1 + \exp\left[-\left(\boldsymbol{a}'\boldsymbol{\eta} + a_s\xi_s + c_m\right)\right]}. \tag{9.10}$$

P^* traces the probability that an item response is in category k or higher conditional on the vector of general dimensions ($\boldsymbol{\eta}$) and at most one specific dimension (ξ_s) with corresponding slope parameters ($\mathbf{a}$) that describe the strength of the relationship of the item response with each latent dimension, and $m - 1$ intercept parameters (c_k). The probability of responding in a particular category, k, is the difference between the probability of responding in category k or higher and the higher response, $k + 1$, or higher:

$$p(\boldsymbol{\eta}, \xi_s) = p^*(k \mid \boldsymbol{\eta}, \xi_s) - p^*(k + 1 \mid \boldsymbol{\eta}, \xi_s). \tag{9.11}$$

Interpreting Conditional Parameter Estimates From MIRT Models

In practice using the slope parameters to interpret the dimensionality of hierarchical models is challenging because the interpretation is limited to assessing the probability of response on one dimension conditional on the model's other dimension(s). The difficulty is that unlike unidimensional models in which a slope on a general factor indicates the marginal item response relationship, a slope parameter in a multidimensional model indicates the relation of an item response with the given dimension conditional on all other dimensions for which the item loads. Hence, using the general dimension slope parameter to interpret the relation between the item and the general dimension may be confusing or misleading, because its magnitude depends on the magnitudes of the item's other slopes (on a specific dimension and any other general dimensions that may be present).

Stucky, Thissen, and Edelen (2013) use the following example to illustrate this challenge: consider two items with the following general and specific dimension slope parameters: item 1 with $a_{general} = 3$ and $a_{specific} = 2$, and item 2 with $a_{general} = 4$ and $a_{specific} = 3$. Though the conditional general dimension slope is higher in magnitude for item 2, the marginal relationship between this item and the general dimension is reduced given the item's relatively high conditional relationship with the specific dimension; as we will see shortly for this example, the net effect is that the strength of the relationship of both items with the general dimension is nearly identical. Such occurrences are not infrequent and clearly present an interpretive challenge even for those familiar with bifactor IRT models.

Marginal IRT Response Functions for Multidimensional Models

As a response to the challenge of interpreting the strength of the relationship between an item response and the general dimension in hierarchical IRT models, Ip (2010a, 2010b) and Stucky and colleagues (2013) developed marginal trace lines (i.e., the IRT response function; Lazarsfeld, 1950) for the general dimension that allow unidimensional-equivalent trace line interpretation. Generalized here for two-tier IRT models, to obtain the marginal trace line for the first general dimension η_1, one must integrate over the remaining general dimensions $(\eta_2 \ldots \eta_G)$ as well as the single specific dimension ξ_s on which the item loads. For example, in a two-tier IRT model, the marginal trace line for the first general dimension is:

$$p_{\text{Marginal}}(y \mid \eta_1) = \int f(\boldsymbol{\eta}, \xi_s)\phi(\eta_2 \ldots \eta_G)\phi(\xi_s)d\eta_2 \ldots d\eta_G d\xi_s. \tag{9.12}$$

In this example, the product of the conditional trace surface and the normal distribution density functions ϕ, integrated across the specific dimension ξ_2 and additional general dimensions $(\eta_2 \ldots \eta_G)$ is the marginal trace line for η_1, P_{Marginal}.[2] In other words, the marginal trace line is obtained by weighting the MIRT model response function by the normal distribution(s) and integrating out the specific (i.e., *nuisance*) dimension and other general dimensions. In practice this integral may be approximated computationally using quadrature.

Logistic Approximations of Marginal Trace Lines for Multidimensional Graded Response Models

After obtaining the marginal trace lines from the general hierarchical IRT model, it is useful to obtain their logistic approximations. For these logistic functions one can obtain IRT

2 Though we demonstrate the marginal trace line for the first general dimension, in practice it could be computed for any general dimension of interest.

item parameters that result in computationally tractable approximations to the marginal trace lines.[3] As an extension of methods proposed by Ip (2010a) for 2-PL and 3-PL models, for the graded response model the technique for obtaining the discrimination parameter estimate for the marginal trace line, a_1^*, from the conditional slope parameter a_1 is to transform the MIRT slope parameters into the bifactor loading metric (λ_1), and then reverse the transformation for the general dimension to arrive at the marginal slopes (Stucky et al., 2013). In other words, for the dimension of interest, in this case the general dimension θ_1:

$$\lambda_1 = \frac{a_1 / D}{\sqrt{1 + \sum (\mathrm{a}/\mathrm{D})^2}}, \tag{9.13}$$

where λ_1 is the general dimension loading in factor analytic notation and D is the commonly used scaling constant 1.7. To simplify the notation, we use the square root of the item variance unexplained by the general latent dimension:

$$\sigma_1 = \sqrt{1 - \lambda_1^2}, \tag{9.14}$$

then the slope parameter estimate for the marginal trace line is:

$$a_1^* = \left(\frac{\lambda_1}{\sigma_1} \right) D \cdot \tag{9.15}$$

Because of the weighting process in Equation (9.12), the slope of the marginal trace line (a_1^*) is never greater in magnitude than the slope of conditional trace lines for θ_1 given values of θ_2 (that is, the trace lines described by a_1) and depending on the relationship between the conditional slopes (a_1 and a_2) the marginal slope may be much smaller.

To obtain the marginal threshold estimates for the logistic approximations of the marginal trace line for the general dimension, the GRM intercept parameters are transformed for each dimension of interest after accounting for all other dimensions. Generalized from unidimensional IRT, for the multidimensional model the marginal threshold for the general dimension is the location on θ_1 where the probability of endorsing a particular response category is 0.5 given that all specific dimensions are fixed at zero. To obtain the threshold for the general dimension the exponents in Equations (9.9) and (9.10) are set to zero (because 1/ [1 + exp (0)] = 0.5), and all specific dimensions are then fixed to be zero so that the exponent is $a_1^* \ \theta_1 - c_{\mathrm{ik}} = 0$, or rearranged for simplicity:

$$b_{ik}^* = {}^{-c_{ik}}\!\!\diagup\!\!_{a_{1i}}. \tag{9.16}$$

Here b_{ik}^* is the location on the general dimension where the probability of endorsing a particular response category is 0.5, averaged over the specific dimensions. Repeating this process for the specific dimension yields a different set of location parameters and may be helpful in understanding the compensatory nature of the model.

3 Regarding the use of the logistic distribution to approximate the normal CDF, Haley (1952) notes that the two never differ by probability values greater than 0.01. Further, Ip (2010a) provides a graphical illustration comparing marginal and logistic approximations that suggests close correspondence. In our experience, across a wide range of marginals the maximum difference in probability between the marginal and logistic approximation of the trace line is never more than about ±0.01. Given this close correspondence, it appears a logistic approximation is sufficiently accurate for most uses.

Application

In this final section we demonstrate the utility of the hierarchical IRT model as a means of describing dimensionality, but also as a way of informing the potential selection of unidimensional item subsets from a larger, multidimensional model. Using the item responses from three PROMIS® short forms (Anger, Anxiety, and Depression, as described later in this chapter), we first demonstrate that despite the high intercorrelations among the three short forms a violation of local independence occurs when the three scales are fit to a single underlying dimension. Next, we evaluate the resulting matrix of LD indices from the unidimensional model to inform the structure of a subsequent bifactor IRT model. Similar to Chapter 2 of this volume, we then consider the extent of the multidimensionality and provide some techniques for assessing the bias in parameter estimates that result from fitting a mis-specified unidimensional IRT model to multidimensional data. Finally, based on recently developed techniques for bifactor models, we explore the potential for selecting a subset of items from the bifactor model that may adequately fit a unidimensional model. As will be demonstrated, this technique serves to identify the items that are both closely representative of the general dimension and not overly influenced by the specific dimensions.

The data used in this application were collected as part of the development and evaluation of the emotional distress item banks from the PROMIS® initiative. PROMIS® is a multi-site research initiative designed to develop, evaluate, and standardize item banks for use in health outcomes research, and its framework includes calibrated item banks covering many domains of health outcomes. For the purposes of our example, we focus on the short form items from the adult emotional distress item bank domains of anger (8 items), anxiety (7 items), and depression (8 items) (Pilkonis et al., 2011). The data collection procedures used a randomized block design that assigned a subset of items to participants in order to maximize response coverage ($N = 15{,}725$). All items had the same five-point response scale with the options 0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often*, and 4 = *always*.

Prior analyses of the PROMIS® anger, anxiety, and depression short forms by Pilkonis and colleagues revealed that distinct unidimensional models closely represent each short form. However, the somewhat strong factor score correlations (depression and anxiety $r = 0.81$, depression and anger $r = 0.60$, anxiety and anger $r = 0.59$; Pilkonis et al., 2011) led Pilkonis and colleagues to suggest that a single construct perhaps labeled "internalized distress" could potentially underlie all three domains. The purpose of this application is to explore the extent to which a bifactor IRT model may best reflect the relationship among the responses to these short forms, and then subsequently to use recently developed psychometric techniques to inform the selection of a unidimensional subset of items. The concept of a unidimensional set of item responses is clearly plausible when considering anxiety and depressive symptoms. Indeed, the idea of a single construct underlying both anxiety and depression has been widely studied and various authors have suggested a range of hierarchical theories each in an attempt to conceptually merge anxiety and depressive symptoms, including "general distress" (Clark & Watson, 1991; Watson et al., 1995a, 1995b), "internalizing spectrum" (Krueger, 1999; Krueger & Finger, 2001; Krueger, McGue, & Iacono, 2001), and "anxious apprehension" (Gray, 1987). In this tradition we propose that a single general underlying dimension, *emotional distress*, may best reflect the shared variance among a subset of anxiety and depression items. The relationship of the anger items with this general emotional distress dimension will also be considered.

Initial Unidimensional IRT Model

It is often informative to begin the item analysis process by fitting a parsimonious unidimensional model in a factor analytic framework. In this approach model fit indices

and LD information (e.g., modification indices (Sörbom, 1989) in LISREL or *Mplus*) are often used to determine the appropriateness of a single-factor model or the need for a more complex hierarchical model. However, for our particular application the use of limited-information estimators (like WLSMV (Muthén, du Toit, & Spisic, 1997) in *Mplus* or DWLS (Jöreskog & Sörbom, 1988) in LISREL) is not possible because the data collection procedures used a randomized block design that resulted in bivariate missingness, predominately between anger and both anxiety and depression items.

Instead we began the model fitting process in an IRT framework by fitting the data to a unidimensional IRT model using marginal maximum likelihood in IRTPRO (Cai, du Toit, & Thissen, 2011). While the full-information model makes use of the complete individual response patterns, one loses access to traditional model fit indices that are helpful in establishing whether the item responses are characterized by a single or multidimensional model (for some recent IRT-based model fit indices, see Maydeu-Olivares & Joe (2005, 2006) and Maydeu (this volume). However, there are several IRT-based indices of LD (e.g., Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2013; Liu & Thissen, 2012; Yen, 1984) that serve the same purpose as the more commonly used factor analytic-based modification indices. As will be shown, IRTPRO provides a matrix of standardized LD χ^2 indices based on differences between the unidimensional model-implied and model-observed bivariate response frequencies (Chen & Thissen, 1997) that provide a means of identifying unmodeled multidimensionality. These indices can be very useful in evaluating the appropriateness of the unidimensional model and correspondingly the need for additional dimensions.

Evaluating Slope Parameter Magnitude

After fitting the PROMIS® anger, anxiety, and depression short form responses with a single unidimensional IRT model, it is useful to first evaluate the magnitude of the slope parameters to establish any potentially dominant item content. Table 9.3 lists the unidimensional IRT parameters sorted by magnitude of the slope parameter within each short form. Clearly the depressive symptoms items dominate the factor structure (average slope = 3.37), followed to a somewhat lesser extent by anxiety (average slope = 2.87), with the anger items being less representative of the dimension (average slope = 2.23). Taken together these results provide the first indication that a possible general dimension may be more representative of feelings of depression and anxiety than feelings of anger (i.e., irritation/annoyance).

Evaluating LD Indices

Next, to understand the residual relationships among item responses that are not accounted for by the unidimensional model, we use the previously mentioned Chen-Thissen LD χ^2 to identify LD as implemented in IRTPRO. The matrix of LD indices provided in HTML format is arranged in deepening shades of red to reflect increasingly large positive LD or deepening shades of blue to reflect increasingly negative LD (which is often ignorable). For demonstration purposes, in Table 9.4 we simplify the resulting LD matrix by merely labeling positive LD as "+", negative LD as "-", and blank values in cells for which there is no bivariate response coverage.[4] To provide some contrast in the LD matrix, positive values greater than nine are in boldface. If the model were sufficiently unidimensional, then we would expect the standardized χ^2 from each item pair to randomly deviate between a

4 As a result of data collection procedures that used a randomized block design, the pairwise response coverage was particularly weak between anger items and both depression and anxiety items, resulting in noticeably more blank values in the anger—depression and anger—anxiety blocks of Table 9.4.

Table 9.3 Unidimensional Item Parameter Estimates for the PROMIS® Anger, Anxiety, and Depression Short Forms

		a	b_1	b_2	b_3	b_4
Anger 1	I felt like I was ready to explode	2.51	0.35	1.18	2.17	3.35
Anger 2	I stayed angry for hours	2.44	0.45	1.39	2.35	3.35
Anger 3	I was grouchy	2.44	-0.91	0.28	1.66	3.04
Anger 4	I felt angrier than I thought I should	2.38	-0.05	0.80	1.94	2.74
Anger 5	I felt annoyed	2.27	-1.24	-0.06	1.27	2.86
Anger 6	I felt angry	2.21	-0.91	0.39	1.85	3.16
Anger 7	I was irritated more than people knew	1.81	-0.93	0.03	1.29	2.48
Anger 8	I made myself angry about something just by thinking about it	1.75	-0.50	0.61	1.93	3.11
Depr 1	I felt helpless	4.21	0.36	0.90	1.62	2.34
Depr 2	I felt worthless	4.11	0.42	0.97	1.65	2.32
Depr 3	I felt hopeless	3.96	0.47	0.97	1.66	2.4
Depr 4	I felt that I had nothing to look forward to	3.38	0.34	0.93	1.60	2.37
Depr 5	I felt unhappy	3.29	-0.68	0.22	1.19	2.14
Depr 6	I felt depressed	2.94	-0.17	0.61	1.48	2.38
Depr 7	I felt sad	2.74	-0.55	0.43	1.45	2.42
Depr 8	I felt like a failure	2.36	0.21	0.88	1.82	2.57
Anx 1	I felt uneasy	3.93	-0.27	0.58	1.54	2.47
Anx 2	I felt tense	3.06	-0.57	0.30	1.27	2.35
Anx 3	I felt nervous	2.82	-0.29	0.64	1.67	2.78
Anx 4	I found it hard to focus on anything other than my anxiety	2.64	0.43	1.23	2.13	2.96
Anx 5	I felt worried	2.61	-0.60	0.28	1.37	2.37
Anx 6	I felt anxious	2.51	-0.18	0.67	1.72	2.67
Anx 7	I felt fearful	2.49	0.38	1.21	2.16	2.99

(small) positive or negative value. Taken as a whole, then, the matrix would show no pattern or clustering of positive or negative values, but rather a random arrangement of "+" or "-" symbols. So, while it is the case that a more nuanced level of detail is lost in this "+/-" display of the matrix, this particular presentation, just by virtue of the pattern of "+" and "-" symbols, shows a clear violation of the IRT assumption of unidimensionality.

Table 9.4 LD Indices for a Unidimensional Model

		Anger								*Depression*								*Anxiety*						
	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Anger	1		+	+	+	+	+	+	+	-	-	-						-	+				-	-
	2	+		+	+	+	+	+	+	-	-	-						-	-					
	3	+	+		+	+	+	+	+	-	-	-							-					
	4	+	+	+		+	+	-	+				-											
	5	+	+	+	+		+	-	-					-	-					-				
	6	+	+	+	+	+		-	-					-	-					-				
	7	+	+	+	-	-	-		+						-					-				
	8	+	+	+	+	-	-	+							-	+				+				
Depression	9	-	-	-							+	+	+	+	+	+	+	-	-	-	-	-	-	-
	10	-	-	-						+		+	+	+	+	+	+	-	-	-	-	-	-	-
	11	-	-	-						+	+		+	+	+	+	+	-	-	-	-	-	-	-
	12				-					+	+	+		+	+	+	+	+	-	+	-	-	-	-
	13					-	-			+	+	+	+		+	+	+	+	+	+	-	-	-	-
	14					-	-	-	-	+	+	+	+	+		+	+	+	+	+	-	-	-	-
	15								+	+	+	+	+	+	+		-	-	-	-	-	-	-	-
	16									+	+	+	+	+	+	-		-	-	-	-	-	-	-
Anxiety	17	-	-							-	-	-	+	+	+	-	-		+	+	+	+	+	+
	18	+	-	-						-	-	-	-	+	+	-	-	+		+	+	+	+	+
	19					-	-	-	+	-	-	-	+	+	+	-	-	+	+		+	+	+	+
	20									-	-	-	-	-	-	-	-	+	+	+		+	+	-
	21									-	-	-	-	-	-	-	-	+	+	+	+		+	+
	22	-								-	-	-	-	-	-	-	-	+	+	+	+	+		+
	23	-								-	-	-	-	-	-	-	-	+	+	+	-	+	+	

Note: Positive LD values are noted by "+"; positive values greater than nine in bold; and negative values are noted by "-".

That is, the covariance within each short form domain is stronger than predicted by a unidimensional model, which is indicated by predominantly bolded "+" cells in these quadrants of the table. Likewise, the item covariance across short form domains is weaker than expected by the model (indicated by predominantly "-" cells in the corresponding quadrants). Thus, in this particular case the pattern of LD neatly illustrates that anger, anxiety, and depression may comprise specific dimensions in a bifactor model.

A Bifactor IRT Model for the PROMIS® Emotional Distress Short Forms

To account for the unique variance among the anger, anxiety, and depression short form items as suggested by the LD indices in Table 9.4, and in order to evaluate the shared variance across these content clusters, we next fit a traditional bifactor model (see Table 9.5). We first note that the threshold estimates appearing in Table 9.5 are presented with respect

Table 9.5 Bifactor IRT Parameter Estimates for the PROMIS® Anger, Anxiety, and Depression Short Forms

	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
Anger 1	2.41	2.1	0	0	0.43	1.45	2.67	4.08
Anger 2	1.98	1.84	0	0	0.61	1.84	3.08	4.36
Anger 3	2.41	1.32	0	0	-1.00	0.29	1.79	3.30
Anger 4	2.07	1.9	0	0	-0.05	1.03	2.48	3.49
Anger 5	2.42	1.24	0	0	-1.30	-0.07	1.33	3.00
Anger 6	2.42	2.01	0	0	-1.05	0.46	2.16	3.67
Anger 7	2.17	1.05	0	0	-0.89	0.04	1.28	2.42
Anger 8	2.01	1.57	0	0	-0.52	0.68	2.12	3.37
Depr 1	4.18	0	1.81	0	0.41	1.01	1.79	2.58
Depr 2	4.44	0	2.22	0	0.48	1.09	1.85	2.61
Depr 3	4.71	0	1.93	0	0.50	1.03	1.75	2.54
Depr 4	3.71	0	2.13	0	0.40	1.07	1.81	2.69
Depr 5	3.63	0	0.56	0	-0.66	0.23	1.18	2.13
Depr 6	3.66	0	0.64	0	-0.18	0.58	1.43	2.28
Depr 7	3.74	0	0.35	0	-0.55	0.38	1.36	2.25
Depr 8	3.46	0	1.71	0	0.24	0.90	1.82	2.53
Anx 1	4.31	0	0	1.79	-0.29	0.58	1.57	2.55
Anx 2	3.15	0	0	1.43	-0.62	0.30	1.32	2.45
Anx 3	3.62	0	0	1.93	-0.31	0.65	1.73	2.88
Anx 4	3.16	0	0	1.29	0.43	1.26	2.17	3.00
Anx 5	3.35	0	0	1.22	-0.61	0.27	1.34	2.32
Anx 6	3.32	0	0	1.48	-0.20	0.66	1.70	2.62
Anx 7	2.65	0	0	1.16	0.37	1.23	2.22	3.06

Note: Threshold parameters were computed with respect to the general dimension.

to the general dimension using Equation (9.16). As Way, Ansley, and Forsyth (1988) noted, threshold estimates should not be expected to be sensitive to dimensionality because they are essentially transformations of the relative proportions of endorsed response categories. Indeed, in this example, there is little difference in these estimates relative to the thresholds from the unidimensional model in Table 9.3, which ignores local dependence. The average absolute difference between the unidimensional and bifactor model item thresholds is 0.04 (SD = 0.04), 0.08 (SD = 0.11), 0.16 (SD = 0.19), and 0.24 (SD = 0.27) for the first through fourth thresholds, respectively. The increasing discrepancy for the higher thresholds merely indicates a loss of estimation precision at the more extreme end of the distribution.

Next, we consider the challenge of interpreting conditional slope parameters in bifactor IRT models. Comparing the general dimension slope estimates in Table 9.5 to the unidimensional estimates in Table 9.3, we note that, without exception, the slope parameters for the general dimension of the bifactor model are larger in magnitude than those from the

Table 9.6 Factor Loadings and Marginal Slope Parameters Based on a Bifactor IRT Analysis

	Factor loadings					*Marginal slopes*			
	λ_1	λ_2	λ_3	λ_4	I-ECV	a_1	a_2	a_3	a_4
Anger 1	0.66	0.58	0	0	0.56	1.49	1.21		
Anger 2	0.62	0.58	0	0	0.53	1.34	1.21		
Anger 3	0.75	0.41	0	0	0.77	1.93	0.76		
Anger 4	0.63	0.58	0	0	0.54	1.38	1.21		
Anger 5	0.75	0.39	0	0	0.79	1.93	0.72		
Anger 6	0.68	0.56	0	0	0.60	1.58	1.15		
Anger 7	0.73	0.36	0	0	0.80	1.82	0.66		
Anger 8	0.66	0.51	0	0	0.63	1.49	1.01		
Depr 1	0.86	0	0.37	0	0.84	2.87		0.68	
Depr 2	0.85	0	0.42	0	0.80	2.74		0.79	
Depr 3	0.88	0	0.36	0	**0.86**	3.15		0.66	
Depr 4	0.81	0	0.46	0	0.76	2.35		0.88	
Depr 5	0.90	0	0.14	0	**0.98**	3.51		0.24	
Depr 6	0.90	0	0.16	0	**0.97**	3.51		0.28	
Depr 7	0.91	0	0.08	0	**0.99**	3.73		0.14	
Depr 8	0.82	0	0.41	0	0.80	2.44		0.76	
Anx 1	0.87	0	0	0.36	**0.85**	3.00			0.66
Anx 2	0.82	0	0	0.37	0.83	2.44			0.68
Anx 3	0.81	0	0	0.44	0.77	2.35			0.83
Anx 4	0.83	0	0	0.34	**0.86**	2.53			0.61
Anx 5	0.85	0	0	0.31	**0.88**	2.74			0.55
Anx 6	0.83	0	0	0.37	0.83	2.53			0.68
Anx 7	0.79	0	0	0.35	0.84	2.19			0.64
ECV	0.79	0.11	0.05	0.05					

unidimensional model. This apparent slope inflation is a (misleading) result of the estimates from the bifactor IRT model reflecting the conditional relationships among the dimensions. Note that the factor analytic parameterization of this structure may be advantageous in certain situations as it results in somewhat simpler interpretations of parameter estimates. For this reason, evaluating the strength of the general dimension in bifactor models using standard rules of thumb based on experience with unidimensional models will inevitably lead to misinterpretation. For example, depression items 3 and 2 (*I felt hopeless* and *I felt worthless*) have the two largest slopes on the general dimension in the bifactor model (4.71 and 4.44, respectively); however, interpreting the strength of their association with the general dimension must be made given their relationship with the specific dimension. To ease this interpretational challenge, the columns to the right in Table 9.6 provide the marginal trace line slope parameters for the general and specific dimensions that were computed from the conditional bifactor parameter estimates using Equation (9.15).

Evaluating the marginal slope parameters in Table 9.6 provides a new, perhaps simplified interpretation of the bifactor results. Because the conditional parameter estimates have been transformed, the values in Table 9.6 can be evaluated as if they are in the metric of the univariate IRT model. In doing so, our perspective on the items' structure may need adjustment. Returning to depression items 3 and 2, the slope parameters for these items reflect a reduced association with the general dimension when going from the conditional to marginal estimates (the item 3 slope is reduced from 4.71 to 3.15, and the item 2 slope is reduced from 4.44 to 2.74). In fact, after evaluating the marginal slope parameters it is clear that the general dimension is most closely represented by depression items 7, 6, and 5 (*I felt sad* (3.73), *I felt depressed* (3.51), and *I felt unhappy* (3.51), respectively). This shift in content emphasis suggests that the unidimensional model, which ignored LD, was oriented toward items measuring more severe aspects of depression (for example, the items reflecting feeling "helpless" and "worthless" had the highest unidimensional IRT slope parameter magnitudes).[5] In contrast, the bifactor IRT model, as hypothesized, accounts for each short form's unique content (i.e., the more severe symptom aspects) via the specific dimensions and accounts for information shared across short forms via the general dimension, which appears to be representative of a less severe symptom expression akin to emotional distress (for example, feeling "unhappy" or "sad").

Evaluating the Magnitude of Multidimensionality in the PROMIS® Emotional Distress Bifactor IRT Model

When fitting bifactor IRT models it is often useful to consider the relative strength or weakness of the general dimension with respect to the specific dimensions. These comparisons not only aid in the interpretation of the general dimension, but also indicate sets of items for which the multidimensionality is relatively weak and possibly ignorable. In particular we find the Explained Common Variance (ECV) index to be a useful indicator of unidimensionality (Reise, Moore, & Haviland, 2010; ten Berge & Socan, 2004). When computed based on the percentage of common variance across all items that is explained by the general dimension, the ECV serves as an index of unidimensionality:

$$\mathrm{ECV} = \frac{\sum \lambda^2_{\mathrm{Gen}}}{\left(\sum \lambda^2_{\mathrm{Gen}}\right) + \left(\sum \lambda^2_{\mathrm{Spec}_K}\right)}. \tag{9.17}$$

5 Chapter 2 of this volume deals exclusively with the degree to which multidimensionality impacts or distorts the estimation of unidimensional IRT parameter estimates.

Using the factor loadings in Table 9.6, the ECV for the general dimension is 0.79. Our experience suggests that ECV values of approximately 0.85 or higher are needed to consider a set of items sufficiently unidimensional to warrant a one-factor model (Stucky et al., 2013; Stucky et al., 2014).[6] The ECV index may also be computed for each specific dimension in the bifactor model to establish the uniqueness of each, simply by replacing the numerator in Equation (9.17) with the specific dimension of interest. In this example, the ECVs are low for the depression and anxiety item sets (specific-dimension ECVs = 0.05 for both, see Table 9.6), as these items largely dominate the general dimension, leaving little remaining unique variance for the specific dimensions. Notice, however, that the anger items represent a somewhat unique construct as evidenced by their low loadings on the general dimension and somewhat higher specific-dimension loadings (ECV = 0.11).

The Impact of Ignoring Multidimensionality

Despite the telltale pattern of LD indices from the unidimensional model that was initially fit, the ECV for the general dimension of the bifactor specification does indicate the presence of a strong dimension underlying this group of items. This naturally leads one to consider the consequences of ignoring the multidimensionality that is present in these items. Is it in fact true that a unidimensional model cannot sufficiently account for the covariances among these items? While it is well known that ignoring the specific dimensions leads to artificially inflated estimates of score reliability (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989), the severity of the local independence violations necessary to affect score reliability remains largely unknown, and in this particular example, the severity of the violation is very much in the "gray area."

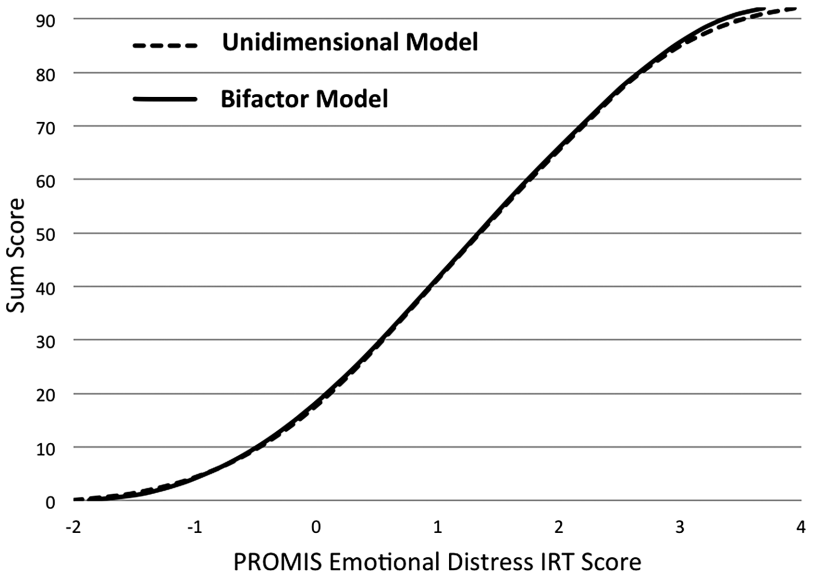

Figure 9.1 Ignored (minor) multidimensionality has negligible impact on score estimates.

6 In studies investigating the effects of fitting unidimensional models to multidimensional data Reise and colleagues (Reise, Scheines, Widaman, & Haviland, 2013; Bonifay et al., under review) evaluate the relationship between the ECV index and various other model fit and factor strength indices (e.g., the Dimensionality Evaluation to Enumerate Contributing Traits index (DETECT; Kim, 1994; Zhang & Stout, 1999), the Percentage of Uncontaminated Correlations (PUC; Reise et al., 2013), and *omegaH* (McDonald, 1999)). Note that coefficient omega and the ECV index are available in the *psych* package for the statistical software R (Revelle, 2013).

One way of identifying the practical impact of ignored multidimensionality on scoring is to compare IRT score and standard error estimates based on the univariate IRT model that ignores multidimensionality with those from the general dimension of the bifactor IRT model that accounts for multidimensionality (i.e., the general dimension from the bifactor model scores will be compared against the single dimension scores from the univariate IRT model). We did this by computing IRT scores (i.e., *expected a posteriori* (EAP)) from summed scores using the recursive algorithm for unidimensional (Thissen, Nelson, Rosa, & McLeod, 2001; Thissen, Pommerich, Billeaud, & Williams, 1995) and bifactor models (Cai, 2010) as implemented in the scoring module of the computer software IRTPRO.[7] Figure 9.1 presents these EAPs in the form of (overlapping) test characteristic curves. Similar to findings by Yen (1993), it appears that minor violations of the assumption of local independence produce mostly negligible differences in scaled score estimates. To further evaluate the impact on scoring we computed the absolute value of the average difference in IRT scores between the univariate and bifactor general dimension model-based summed scores weighted by the probability of obtaining a given summed score. Based on this approach the average EAP difference between the two scores across the underlying continuum is 0.03, confirming the relatively minor impact of ignoring multidimensionality on scoring.

However, when evaluating the impact of ignored multidimensionality on score precision, the effects are more pronounced. Figure 9.2 shows the inflated score precision of the unidimensional model across the summed scores of the underlying dimension. The difference in marginal reliability is 0.07, which indicates that even when violations of the univariate IRT model's assumption of local independence appear minor by most accounts, the resulting bias in score precision can be large enough to potentially mischaracterize the utility of the scale. Because IRT estimates of score precision are based on the model parameters, the bias in score precision is the result of the unidimensional model's overestimation of the strength of the relationship between (at least some) items and the underlying

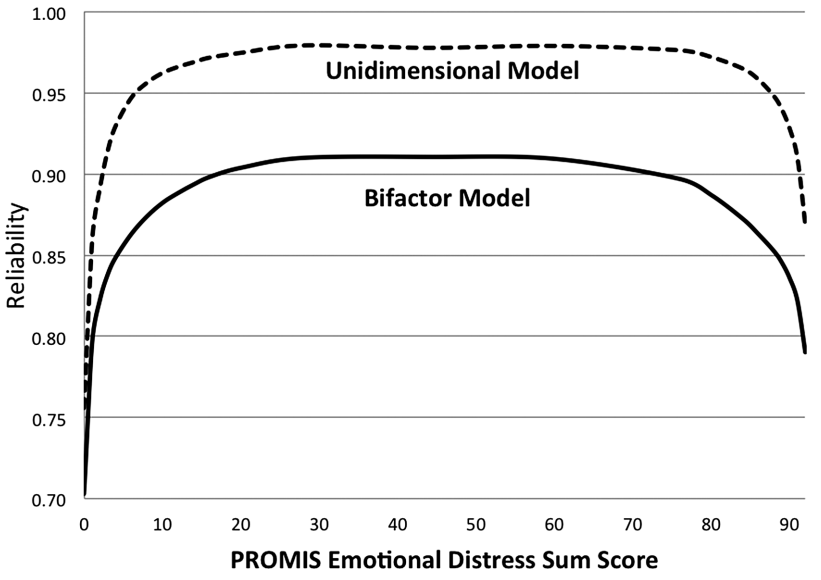

Figure 9.2 Ignored multidimensionality leads to inflated score precision.

7 A DOS-based computer program that computes IRT scores from summed scores is also freely available from the second author (orlando@rand.org).

dimension. In other words, the magnitudes of some of the slope parameters in the unidimensional model are exaggerated.

Identifying a Unidimensional Subset of Items

While our analyses thus far have shown the problematic effects of ignored multidimensionality, the results have also demonstrated the prominence of the general dimension (high ECV value for bifactor general dimension) in comparison to relatively weak specific dimensions. This leads to the question: Is there a unidimensional subset of items that represent the general dimension, emotional distress? While the high ECV value obtained earlier indicates that *some* subset of items may be fit with a unidimensional model, it does not indicate which items should be selected or which short form content domains should be represented in the unidimensional item subset. To inform the extent to which items from a given short form are likely to contribute to this unidimensional subset, it is useful to return to the ECV index. However, instead of computing a single ECV using the full collection of items, we compute separate *within-domain* ECVs for each short form. That is, Equation (9.17) is used to compute an ECV based only on the general and specific factor loadings for the items in a given short form (e.g., the within-domain ECV for anger is computed using only the eight anger items' general and specific factor loadings from the original bifactor model).

In this context the *within-domain* ECVs have the same interpretation as before, but now their relative magnitudes indicate which particular content domains are most representative of the general dimension, and indeed which short forms should contribute most to the unidimensional subset of items (i.e., short forms with higher within-domain ECVs). Moreover, it may not be appropriate to include items in the unidimensional item subset from content domains with low within-domain ECVs, as the combination of these items with items from other dimensions is unlikely to result in a unidimensional set of item responses. In this example, the within-domain ECV values for the anger, anxiety, and depression domain subsets are 0.65, 0.87, and 0.84, respectively. Noting that ECV values > 0.85 generally reflect a sufficiently unidimensional item set, the within-domain ECV values in this example indicate that covariance among items selected from the depression and anxiety domains may form an emotional distress construct that can be adequately represented by a unidimensional model.[8]

After establishing that the general dimension is most representative of the depression and anxiety domains, we now take a closer look at the items within these domains in an effort to select a unidimensional subset of items that maximizes the strength of the general dimension (emotional distress) while minimizing the impact of the specific dimensions. To aid in the selection of items we have recently used another variation of the ECV that is defined for each item (I-ECV). The I-ECV is calculated as the ratio of the item-level variance accounted for by the general factor to the total item-level variance accounted for by the general and specific dimensions (i.e., the numerator and denominator of Equation (9.17) use only the general and specific factor loadings from a single item). The I-ECV indicates the extent to which an item is representative of the general dimension alone—values near one indicate an item that only reflects the general dimension whereas increasingly smaller values reflect stronger associations with the specific dimension (Stucky, Thissen, & Edelen, 2013).

However, note that the magnitudes of the I-ECVs are entirely dependent on the content of the other items included in the specific dimension. Therefore, in a sense the I-ECV is a way to identify mostly general content items among a set of items with similar specific

8 The possibility of merging depression and anxiety item content to form an emotional distress unidimensional subset of items is somewhat further justified given the strong correlation between the anxiety and depression dimensions ($r = 0.87$).

content of various valence. In other words, the I-ECV provides the same indication of unidimensionality at the individual item level as the ECV does at the scale level. In practice the I-ECV is a useful aid in the selection of unidimensional subsets of items. Our experience using this index to select unidimensional items from bifactor IRT models suggests that choosing items with relatively high general factor loadings and I-ECV values greater than 0.80 or 0.85 will typically yield a fairly unidimensional item set that reflects the content of the general dimension.

Item Selection and Results for the Seven-Item Emotional Distress Short Form

Based on this criterion we selected the seven items from Table 9.6 (in boldface) with I-ECVs greater than 0.85. This item selection process resulted in a subset of items not overly dominated by either depression or anxiety, but rather balance a strong relation with the general dimension and weak relations with their respective specific dimensions, making them an ideal set to represent emotional distress. Note that, as suggested by the low within-domain ECV for anger, the eight anger items' I-ECVs indicated they would not be suitable for inclusion in this unidimensional subset.

To evaluate the appropriateness of this unidimensional model to characterize the covariance among the selected subset of item responses, we fit a one-factor model in *Mplus* using the limited information estimator WLSMV. The results suggest the subset of items is unidimensional, $\chi^2 = 205$, $df = 14$, $p < 0.01$, CFI = 0.995, TLI = 0.992, RMSEA = 0.039. In addition we note that while the initial bifactor model had an ECV of 0.79, indicative of a non-unidimensional model, the ECV of 0.91 computed from only the revised seven emotional distress dimension items clearly indicates a unidimensional item set.

Finally, it is important to ensure that the seven selected items appropriately reflect the reliability of scores for the emotional distress dimension as defined by the general factor of the original bifactor model. To evaluate the bias in scores that may have resulted in treating the seven items as a single dimension, we compared two separate estimates of the

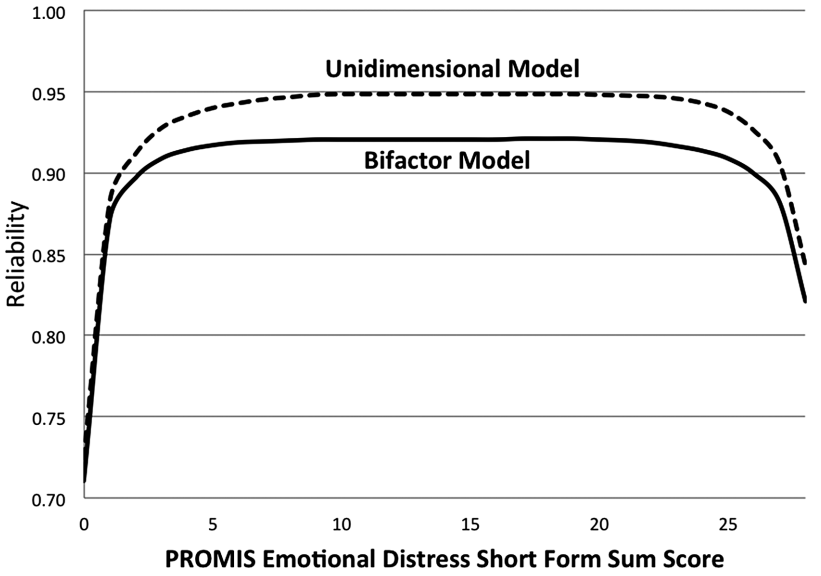

Figure 9.3 Score reliability for the seven-item emotional distress short form.

reliability of IRT scores computed from summed scores: 1) "bifactor-based" reliabilities computed for the general dimension using the item parameters from the original bifactor model (see Table 9.5), and 2) "IRT-based" reliabilities computed from a unidimensional seven-item IRT item calibration (see Figure 9.3). Because fitting unidimensional models to multidimensional data results in both biased parameter estimates (see Chapter 2 of this volume) and biased estimates of score precision, as previously discussed in this chapter, we would expect that a severe violation in unidimensionality would result in inflated estimates of score precision (as seen in Figure 9.2). Note that while Figure 9.3 indicates modest differences in score reliability, indicating some inflation in score precision, the average difference in short form score reliability for the two approaches is only 0.02 (in contrast to a difference of 0.07 obtained in the full item set), which is negligible for most scoring purposes.

Summary

The first part of this chapter reviewed a number of issues regarding the use of hierarchical IRT measurement models along with more recent generalizations (namely the two-tier model) that, with little added complexity, can accommodate less restrictive measurement models with multiple general and specific dimensions. We note that though informative, these hierarchical measurement models are challenging to interpret unless the estimated conditional parameters are transformed into the more familiar metric of the univariate IRT model (i.e., marginal trace lines).

Following this background, the second part of this chapter illustrates the versatility of the hierarchical model both as a means of describing the measurement properties of a set of items, and as a means of aiding in the selection of unidimensional subsets of items from multidimensional data. Following Chapter 2 of this volume, which describes how multidimensionality affects the parameter estimation of unidimensional IRT models, we conclude by presenting some novel psychometric techniques that may be useful in minimizing these well-known problems. We note that when used carefully these techniques do not result in (serious) biases in estimates of unidimensional item parameters or score reliabilities, though further work is needed to establish general guidelines. Taken together the marginal trace line item parameters ECV and I-ECV indices serve as useful tools for the test analyst with the goal of establishing a unidimensional scale, when faced with multidimensional data. We hope we have effectively demonstrated the utility of the general hierarchical model both as a means of describing the dimensionality of an item set and also as a way of gaining insight into the potential for alternative measurement structures.

References

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

Ansley, T.M., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 39–48.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bolt, D. (2005). Limited vs. full information estimation in IRT modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics: A festschrift to Roderick P. McDonald* (pp. 27–72). Mahwah, NJ: Lawrence Erlbaum Associates.

Bonifay, W.E., Reise, S.P., Scheines, R., & Meijer, R.R. (under review). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the *DETECT* Multidimensionality Index.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L., du Toit, S.H.C., & Thissen, D. (2011). IRTPRO Version 2: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221–248.

Chen, W.H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Clark, L.A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*, 316–336.

De Ayala, R.J. (1994). The influence of multidimensionality on the graded response model. A*pplied Psychological Measurement, 18*, 155–170.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189–199.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 3, 423–436.

Gibbons, R.D., Rush, A.J., & Immekus, J.C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research, 43*, 401–410.

Gray, J.A. (1987). *The psychology of fear and stress* (2nd ed.). Cambridge, UK: Cambridge University Press.

Haley, D.C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error Technical Report No. 15* (Office of Naval Research Contract No. 25140, NR-342-022). Stanford University: Applied Mathematics and Statistics Laboratory.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*, 91–115.

Holzinger, K.J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54.

Ip, E.H. (2010a). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement*, *34*, 467–482.

Ip, E.H. (2010b). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Irwin, D.E., Stucky, B.D., Langer, M.L., Thissen, D., DeWitt, E.M., Lai, J.S., Varni, J., Yeatts, K., & DeWalt, D.D. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*, 595–607.

Jöreskog, K.G., & Sörbom, D. (1988). *PRELIS: A program for multivariate data screening and data summarization. A pre-processor for LISREL* (2nd ed.). Mooresville, IN: Scientific Software.

Kim, H. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation. University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 55–12B*, 5598. Retrieved from: http://hdl.handle.net/2142/19110.

Krueger, R.F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*, 921–926.

Krueger, R.F., & Finger, M.S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment, 13*, 140–151.

Krueger, R.F., McGue, M., & Iacono, W.G. (2001). The higher order structure of common DSM mental disorders: Internalization, externalization, and their connections to personality. *Personality and Individual Differences, 30*, 1245–1259.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al. (Eds.), *Measurement and prediction*. Princeton NJ: Princeton University Press.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*, 254–274.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*, 670–688.

Luecht, R.M., & Miller, T.R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*, 279–293.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732.

McDonald, R.P. (1999) *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Muraki, E., & Carlson, J.E. (1993). *Full-information factor analysis for polytomous item responses*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Conditionally accepted for publication in *Psychometrika*.

Pilkonis, P.A., Choi, S.W., Reise, S.P., Stover, A.M., Riley, W.T., Cella, D., & PROMIS Cooperative Group (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, Anxiety, and Anger, *18*, 263–283.

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

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., . . . Cella, D (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22–31.

Reise, S.P. (2012). The rediscovery of the bifactor measurement model. *Multivariate Behavioral Research, 47*, 667–696.

Reise, S.P., Moore, T.M., & Haviland, M.G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544–559.

Reise, S.P., Morizot, J., & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19–31.

Reise, S.P., Scheines, R., Widaman, K.F., & Haviland, M.G. (2013). The effects of multidimensionality on structural coefficients in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*, 5–26.

Revelle, W. (2013). *Package "psych."* Retrieved from https://personality-project.org/r/psych.manual.pdf.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

Schilling, S., & Bock, R.D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.

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

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371–384.

Steinberg, L. (1994). Context and serial order effects in personality measurement: Limits on the generality of "measuring changes the measure." *Journal of Personality and Social Psychology, 66*, 341–349.

Stucky, B. D., Edelen, M. O., Vaughan, C. A., Tucker, J. S., & Butler, J. (2014). The psychometric development and initial validation of the DCI-A Short Form for adolescent therapeutic community treatment process. *Journal of Substance Abuse and Treatment, 46*, 516–52.

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37*, 23–39.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

ten Berge, J. M. F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613–625.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Thissen, D., & Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 123–144). Washington, DC: American Psychological Association.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace Lines for Testlets: A use of multiple-categorical-response models. *Journal for Educational Measurement*, *26*, 247–260.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181–195.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory. An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Kluwer-Nijhoff, 245–270.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *ETS Research Report 02–02*.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. A., Strauss, M. E., & McCormick, R. A. (1995a). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptoms. *Journal of Abnormal Psychology, 104*, 3–14.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995b). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology, 104*, 15–25.

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.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213–249.