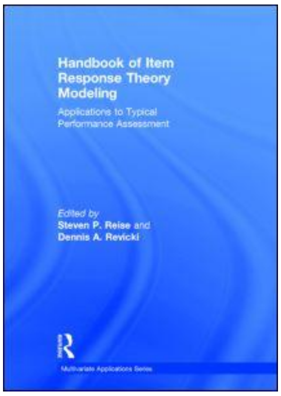

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

An Illustration of the Two-Tier Item Factor Analysis Model

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch10
Wes E. Bonifay
Published online on: 16 Dec 2014

10 An Illustration of the Two-Tier Item Factor Analysis Model

Wes E. Bonifay

Introduction

Multidimensional item response theory (IRT) presumes that the items on a test measure multiple underlying latent constructs. Thanks to the introduction of item factor analysis (IFA; Bock, Gibbons, & Muraki, 1988) and recent advances in efficient estimation of complex latent variable models (e.g., Rijmen, Vansteelandt, & De Boeck, 2008), multidimensional IRT has become increasingly popular in the modeling of item responses. Specifically, the standard (correlated-traits) multidimensional IRT model (see Reckase, 2009) and more sophisticated models such as the item bifactor model (Gibbons & Hedeker, 1992) and the testlet response model (Bradlow, Wainer, & Wang, 1999) have seen widespread use in psychological and educational measurement.

The bifactor IFA model and the testlet response model both utilize a factor structure that includes a single primary (or general) factor and multiple specific factors. The primary factor accounts for the latent dimension underlying all items in a test instrument, while the specific factors are designed to account for the residual dependence that exists after extracting this primary factor. Of course, addressing this residual dependence tends to result in models that provide superior goodness-of-fit to the data. Accordingly, these primary factor approaches to multidimensional item response analysis have seen a rise in popularity in recent years (the bifactor IRT model, in particular, has enjoyed a resurgence as of late; see Reise, 2012).

Although the testlet and bifactor IRT approaches are certainly bolstered by their ability to model residual dependence among test items, permitting just a solitary primary dimension may be an unnecessary restriction. To overcome this limitation, Cai (2010a) proposed the two-tier IFA model, which includes a Thurstonian simple structure among the primary dimensions while retaining the specific factors. The two-tier model positions the latent variables into two classes, or tiers: primary dimensions and specific dimensions. As in the testlet response and (confirmatory) bifactor IRT models, the primary tier dimensions and the specific tier dimensions are not correlated, and all specific dimensions are mutually orthogonal. Certain restrictions are then imposed on the factor pattern such that each item can be influenced by multiple primary factors and a single specific factor. Of course, the feature that distinguishes the two-tier model from previous multidimensional IRT models is the presence of *multiple* primary (or general) dimensions. Moreover, as discussed later in this chapter, the correlated-traits, bifactor IRT, and testlet response models are all subsumed by the more general two-tier model.

Cai (2010a) explores multiple uses for such a structure. First, the inclusion of multiple primary dimensions results in a superior measurement model, both in terms of measurement reliability and goodness-of-fit. This chapter will present a real data analysis that showcases the psychometric advantages of the two-tier IFA model. Second, the model can be adapted, quite easily, for longitudinal modeling (and in fact, the longitudinal bifactor IRT model (Hill, 2006) can be understood as a special case of the two-tier model). In the longitudinal case, the estimated correlation between the primary dimensions in the model represents the degree to which the latent trait is stable over time. The two-tier approach to longitudinal measurement improves upon standard longitudinal factor analytic models (e.g., Tisak & Meredith, 1989), which are not designed for item-level data. Third, the two-tier model can be used to model individual idiosyncrasies in response style by extending the random intercept IFA model introduced by Maydeu-Olivares and Coffman (2006). The longitudinal and response style applications of the two-tier model will not be discussed in this chapter; see Cai (2010a) for further information.

Research Methods

The Two-Tier Structure

As an illustration of the structure of a two-tier IFA model, suppose a test instrument includes 18 items and that the factor structure underlying the test includes three primary dimensions in the first tier and two specific dimensions per primary dimension (for a total of six specific latent factors) in the second tier. For this hypothetical model, the standard factor analytic measurement model $\boldsymbol{y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\epsilon}$ can be written as:

$$
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{18} \end{pmatrix}
=
\left(\begin{array}{ccc|cccccc}
\lambda_{1,1} & & & \lambda_{1,4} & & & & & \\
\lambda_{2,1} & & & \lambda_{2,4} & & & & & \\
\lambda_{3,1} & & \lambda_{3,3} & \lambda_{3,4} & & & & & \\
\lambda_{4,1} & & & & \lambda_{4,5} & & & & \\
\lambda_{5,1} & & & & \lambda_{5,5} & & & & \\
\lambda_{6,1} & & & & \lambda_{6,5} & & & & \\
 & \lambda_{7,2} & & & & \lambda_{7,6} & & & \\
 & \lambda_{8,2} & & & & \lambda_{8,6} & & & \\
 & \lambda_{9,2} & & & & \lambda_{9,6} & & & \\
 & \lambda_{10,2} & & & & & \lambda_{10,7} & & \\
 & \lambda_{11,2} & & & & & \lambda_{11,7} & & \\
 & \lambda_{12,2} & & & & & \lambda_{12,7} & & \\
 & & \lambda_{13,3} & & & & & \lambda_{13,8} & \\
 & & \lambda_{14,3} & & & & & \lambda_{14,8} & \\
 & & \lambda_{15,3} & & & & & \lambda_{15,8} & \\
 & & \lambda_{16,3} & & & & & & \lambda_{16,9} \\
 & \lambda_{17,2} & \lambda_{17,3} & & & & & & \lambda_{17,9} \\
 & & \lambda_{18,3} & & & & & & \lambda_{18,9}
\end{array}\right)
\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ -- \\ \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \\ \xi_6 \end{pmatrix}
+
\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \end{pmatrix}, \quad (10.1)
$$

where the dashed lines in the $\boldsymbol{\Lambda}$ (factor pattern) matrix and the $\boldsymbol{\eta}$ (latent variable) vector are included simply to delineate the first and second tiers. Notice that the two-tier model allows for cross-loadings among the primary (first-tier) dimensions; the $\boldsymbol{\Lambda}$ matrix clearly shows that item y_3 is free to load on primary factors η_1 and η_3, while item y_{17} is free to load on primary factors η_2 and η_3. In fact, the two-tier model permits any primary tier factor pattern (subject to identification) to be specified and tested. This ability to include primary dimension cross-loadings is the central feature that makes the two-tier model more flexible than the item bifactor or testlet response models. The $\boldsymbol{\Lambda}$ matrix also displays each specific factor ξ; as in the bifactor IRT and testlet models, cross-loadings between specific factors are prohibited in the two-tier IFA model.

The two-tier model also requires certain constraints on the factor covariance matrix such that the covariances among the $\boldsymbol{\eta}$ primary dimensions are freely estimated while the covariances among the $\boldsymbol{\xi}$ specific dimensions are fixed at zero. These specifications result in the following factor covariance matrix structure:

$$\begin{pmatrix} \Sigma & \\ 0 & \operatorname{diag}(\tau) \end{pmatrix}, \tag{10.2}$$

where $\boldsymbol{\Sigma}$ is the portion of the covariance matrix that is associated with the first tier dimensions and diag(τ) is the portion associated with the second tier dimensions. $\boldsymbol{\Sigma}$ can be of any type (subject to identification), thereby allowing any pattern of correlations among the primary factors, while the τ quadrant is necessarily diagonal, thereby ensuring that all specific factors are mutually orthogonal. Further, both the primary dimensions and the specific dimensions have unit variances.

To continue our illustration, let us consider the factor variance/covariance matrix of the hypothetical model:

$$\left(\begin{array}{ccc|cccccc} \sigma_{11} & & & & & & & & \\ \sigma_{21} & \sigma_{22} & & & & & & & \\ \sigma_{31} & \sigma_{32} & \sigma_{32} & & & & & & \\ \hdashline & & & \tau_{44} & & & & & \\ & & & & \tau_{55} & & & & \\ & & & & & \tau_{66} & & & \\ & & & & & & \tau_{77} & & \\ & & & & & & & \tau_{88} & \\ & & & & & & & & \tau_{99} \end{array}\right) \tag{10.3}$$

Again, the dashed lines are included in order to differentiate between the (correlated) first-tier dimensions and the (orthogonal) second-tier dimensions. The off-diagonal σ_{ij} elements in the upper-left quadrant indicate that all three primary dimensions in our example model are allowed to mutually covary. However, the off-diagonal τ_{ij} elements in the lower-right quadrant are constrained to zero because, just as in the bifactor and testlet models, all specific dimensions are mutually orthogonal in the two-tier model.

Figure 10.1 is a visual representation of the example model. By using the diagrammatic conventions of square indicators and circular latent variables, Figure 10.1 clearly displays the two-tier structure of the model. The first tier is represented by the three primary latent variables in the top portion of the figure, while the second tier is represented by the six specific latent factors in the bottom portion. The cross-loadings between primary dimensions are also evident in this figure. The covariances between the primary dimensions are, of course, represented by the double-headed arrows between the first tier factors.

Figure 10.1 also enables one to visualize the relationships between the two-tier model and the correlated-traits, bifactor IRT, and testlet response models. Specifically, a standard

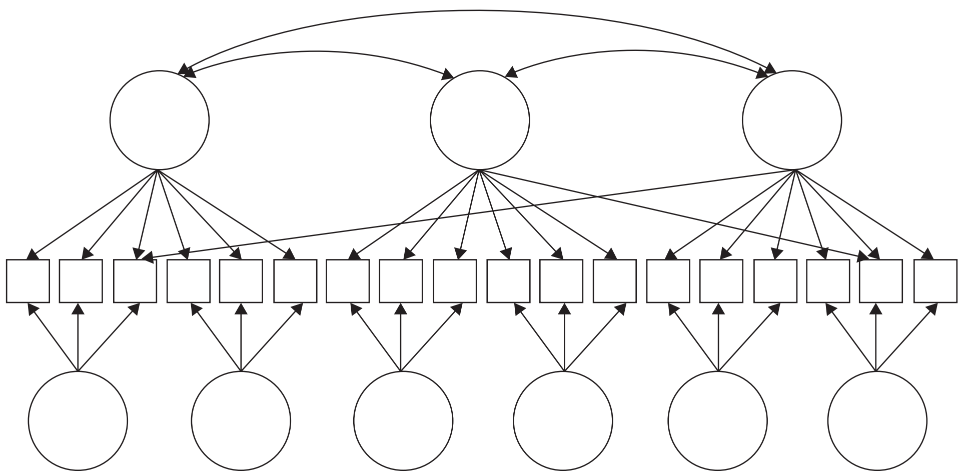

Figure 10.1 The two-tier factor structure of the hypothetical model.

correlated-traits IRT model, which does not account for residual dependence, would be produced by completely eliminating the second tier latent dimensions and all associated paths. Removing the covariances between the primary dimensions in the first tier would result in three separate bifactor IRT models. The testlet response model would be found by removing the primary dimension covariances, specifying equality constraints between the primary and specific slopes of each item, and freely estimating the specific factor variances. In sum, the two-tier IFA model incorporates the most common multidimensional IRT models and is therefore a more general and flexible structure.

The Two-Tier Model

The data analysis presented later in this chapter is based on a psychiatric symptom checklist of binary response (*presence* vs. *absence*) items. Accordingly, we will briefly discuss the two-tier formulation of the classical two-parameter logistic (2PL) IRT model. Although the 2PL model shown in Equation (10.4) is strictly for use with dichotomously scored items, the two-tier IFA model can also be implemented using polytomous items. See Cai (2010a) for two-tier adaptations of the graded response model (Samejima, 1969, 1997) and the nominal categories model (Thissen, Cai, & Bock, 2010).

Let there be p underlying primary dimensions and S specific dimensions. The conditional probability of correct/endorsement response for a dichotomously scored item y_{ij} is then defined by:

$$P_j\left(y_{ij} = 1 \mid \eta_i, \xi_{is}, \theta\right) = \frac{1}{1 + \exp\left\{-\left[\alpha_i(\theta) + \left[\beta_i(\theta)\right]'\eta_i + \beta_{is}(\theta)\xi_{is}\right]\right\}}, \tag{10.4}$$

where α_i is the intercept, $\boldsymbol{\beta}_i$ is the $p \times 1$ vector of item slopes on the primary factor vector $\boldsymbol{\eta}_i$, β_{is} is the item slope on specific factor ξ_{is}, and $\boldsymbol{\theta}$ is a vector of all estimable and/or structural parameters in the two-tier model. Note that $\boldsymbol{\theta}$ is not interpreted according to traditional IRT conventions; rather, Equation (10.4) adopts the standard statistical usage of $\boldsymbol{\theta}$ as the parameter vector, which is necessary because of the confirmatory quality of the (potentially complex) two-tier model. The logit term in the denominator clearly highlights the dependence of the item parameters on the parameter vector $\boldsymbol{\theta}$. Finally, the conditional probability for the incorrect/non-endorsement response is simply $P_j(y_{ij} = 0 \mid \boldsymbol{\eta}_i, \xi_{is}, \boldsymbol{\theta}) = 1 - P_j(y_{ij} = 1 \mid \boldsymbol{\eta}_i, \xi_{is}, \boldsymbol{\theta})$.

Application

We turn now to a real data application of the two-tier IFA model. The analysis discussed in this chapter is based on the item responses of 3,999 individuals who participated in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D; Fava et al., 2003; Rush et al., 2004) trial.[1] All participants satisfied DSM-IV criteria for nonpsychotic major depressive disorder, as indicated by a Hamilton Rating Scale for Depression (HAM-D) score of at least 14. The STAR*D studies were sponsored by the National Institute of Mental Health and conducted at 41 clinical outpatient facilities throughout the United States. Further information on the STAR*D trial and several publically available data sets can be found at www.star-d.org.

The Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001) is a 139-item[2] self-report scale that was designed to measure the 15 most common Axis I psychiatric disorders that are encountered in outpatient mental health settings: major depressive disorder (MDD), dysthymia (DYS), post-traumatic stress disorder (PTS), bulimia (BUL), obsessive-compulsive disorder (OCD), panic disorder (PAN), mania (MAN), psychosis (PSY), agoraphobia (AGO), social phobia (SOC), alcohol abuse (ALC), drug abuse (DRUG), generalized anxiety disorder (GAD), somatoform disorder (SOM), and hypochondria (HYPO). Zimmerman and Mattia (2001) reported that the PDSQ subscales had an average alpha coefficient of 0.86, with 14 of the 15 subdomains producing alpha values greater than 0.80. The authors also found an average test-retest reliability coefficient of 0.83, with nine of the subdomains producing test-retest coefficients that exceeded 0.80. Further, tests of convergent and discriminant validity indicated that each subscale was more highly correlated with other measures of the same constructs than with other measures of different symptoms (e.g., scores from the MDD subscale of the PDSQ were more highly correlated with scores from an alternate measure of depression than they were with scores from a measure of panic disorder). The developers of this instrument concluded that the strong diagnostic performance of these subscales allows for accurate classification of patients (Zimmerman & Mattia, 2001). Accordingly, the PDSQ is intended to be used as a screening instrument that incoming patients should complete prior to receiving a formal diagnostic evaluation.

When administering such a screening questionnaire, it is important to note that the factor structure of an instrument can impact its diagnostic properties. If, for example, the PDSQ has a unidimensional structure, then each of the 15 subscales can be thought of as representing a different threshold along a single underlying continuum. At the opposite extreme, the PDSQ may be a composite of 15 qualitatively distinct, psychometrically unrelated subdomains. Between these two alternatives exists the item bifactor model (Gibbons & Hedeker, 1992), wherein the structure of an instrument includes an overall

1 I would like to thank Waguih IsHak, Ph.D., for providing the PDSQ data set.

2 Two items were removed from the analyses in this chapter. Items 5 and 6 asked conflicting questions about appetite decrease and increase, respectively, while items 7 and 8 asked conflicting questions about sleep excess and deprivation, respectively. These contradictory items were eliminated, resulting in 137 items.

latent dimension (the "general factor") as well as several smaller, distinct dimensions (or "specific factors"). In the case of the PDSQ, a bifactor model would include an overall "psychiatric impairment" factor as well as 15 orthogonal specific factors, each of which represents one of the psychiatric disorders listed earlier. A practitioner could then ascertain, simultaneously, a patient's overall psychiatric impairment as well as the extent to which the patient conveys the symptoms of MDD, bulimia, and so on. Such a structure would thereby give added support to the practice of making diagnostic inferences based on PDSQ responses.

Gibbons, Rush, and Immekus (2009) fit the PDSQ data to several IRT models, including a bifactor model. The authors presented their bifactor results in terms of item factor loadings and showed that most of the items had acceptable loadings on both a general factor as well as a domain-specific factor. However, the MDD items produced relatively lower item factor loadings on the primary dimension (ranging from –0.04 to 0.34, with an average loading of 0.23) and quite heterogeneous specific factor loadings (ranging from 0.03 to 0.90). Non-MDD items, on the other hand, had higher average loadings on the primary dimension and more homogeneous loadings within each domain (see Table 1 in Gibbons, Rush, & Immekus, 2009).

A replication of this analysis (in terms of standard IRT parameters rather than item factor loadings) revealed that, although the bifactor model does explain the multidimensionality inherent in the PDSQ, many of the parameter estimates related to the MDD items were inadequate relative to the estimates of the other (non-MDD) domains. Regarding the IRT parameter estimates of the primary dimension, we found that the slope estimates of the MDD items ranged from 0.31 to 0.89 ($M = 0.57$) while the slope estimates of the non-MDD items ranged from 0.53 to 2.83 ($M = 1.60$). As for the domain-specific factors, the mean MDD slope estimate was 1.15 while the average slope estimate of each non-MDD domain ranged from 1.33 to 2.49 (with an overall mean of 1.88). In sum, the MDD items had lower primary and specific dimension discrimination parameter estimates than the other 14 domains on the PDSQ. Clearly, this discrepancy suggests that, in a bifactor structure, the MDD items are not as aligned with the general factor as well as the non-MDD items.

Two implications can be drawn from these results. First, the broad range of item factor loadings on the MDD specific factor suggests the presence of multidimensionality within this domain. Exploratory factor analyses (EFA) revealed that all of the PDSQ subdomains except MDD were sufficiently unidimensional. MDD was found to be composed of seven distinct factors, which aligned to item content related to general sadness, lack of interest, physical effects of depression, self-loathing, difficulty concentrating, suicidal ideation, and suicidal intent. However, because all 19 items were certainly indicative of an overall depression dimension, it was logical to fit a bifactor model to the MDD items. Thus, we conducted a confirmatory bifactor item factor analysis based on the optimal (7-factor) EFA solution.[3] The results of this analysis are displayed in Table 10.1. The pattern of loadings supports the multidimensionality of the MDD subscale—a modeling concern that was not addressed by Gibbons, Rush, and Immekus (2009) in their bifactor model of the PDSQ.

The second implication of the bifactor results is that the MDD items, relative to the other 14 PDSQ domains, appear to have a distinct relationship with the overall

3 MDD items 17 and 21 were causing estimation problems, so they were removed from this and all subsequent analyses.

Table 10.1 Confirmatory Bifactor Item Factor Analysis of the Major Depressive Disorder (MDD) Subdomain Items of the Psychiatric Diagnostic Screening Questionnaire

	General Factor	*Specific Factors*						
		1	2	3	4	5	6	7
MDD_01	.71	.46						
MDD_02	.66	.49						
MDD_03	.56		.67					
MDD_04	.55		.67					
MDD_10	.41		.28					
MDD_05	.25			.45				
MDD_07	.22			.63				
MDD_09	.30			.39				
MDD_11	.48				.43			
MDD_12	.64				.69			
MDD_13	.71				.47			
MDD_14	.48					.66		
MDD_15	.54					.64		
MDD_16	.67						.54	
MDD_18	.66						.54	
MDD_19	.62							.62
MDD_20	.60							.64

Note: N = 3999. RMSEA based on the M2 statistic = 0.05.

"psychiatric impairment" dimension. That is, the lower discrimination parameter estimates of the MDD subscale indicate that the depression items are not sufficiently explained by the general factor. One solution would be to model two separate bifactor solutions—one for the MDD items and one for the non-MDD items. Although this may suffice, Cai's (2010a) two-tier approach would allow all PDSQ items to be analyzed in a single model that would account for the multidimensionality of the MDD subscale while permitting the MDD items to load on a separate primary dimension from the other 14 subdomains. Moreover, the two-tier model draws on the covariance between the primary dimensions to provide more precise parameter estimates (i.e., lower standard errors) (Cai, 2010a).

The proposed two-tier item factor analysis model of the PDSQ is visually represented in Figure 10.2. Note that the alcohol abuse and drug abuse domains were not included in the model; as Gibbons, Rush, and Immekus noted, "ALC and DRUG domains were relatively independent of the primary dimension that the PDSQ measures, and represent 'independent' factors" (Gibbons, Rush, & Immekus, 2009, p. 408). Further, two MDD items (MDD_17 and MDD_21), several bulimia items (BUL_01 through BUL_04, BUL_06,

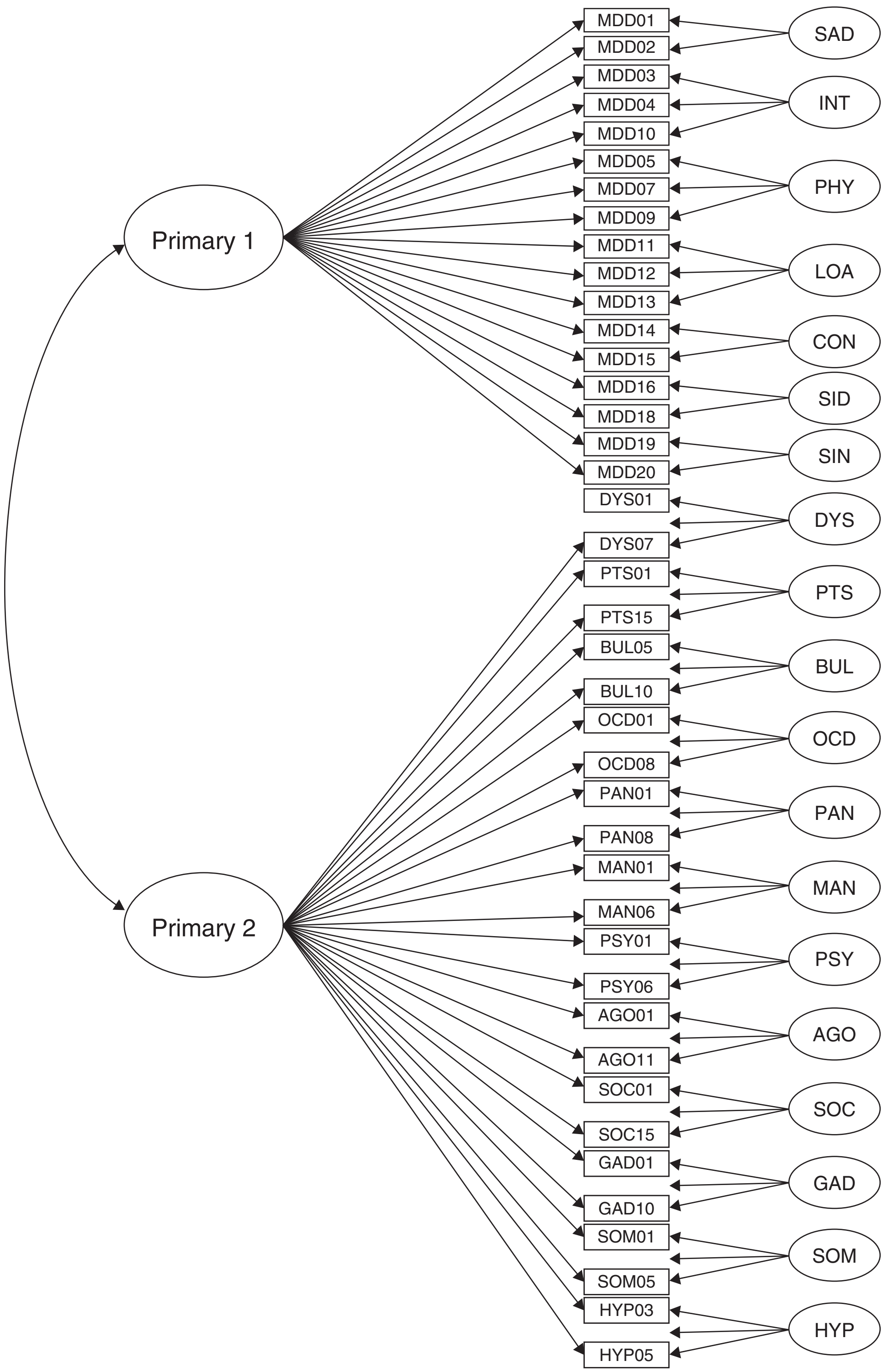

Figure 10.2 A two-tier model of the Psychiatric Diagnostic Screening Questionnaire.

and BUL_07), and a pair of hypochondria items (HYP_01 and HYP_02) were eliminated because of estimation problems. The final model included two correlated primary dimensions (MDD and non-MDD domains) in the first tier, 19 orthogonal specific dimensions (7 MDD specific factors + 12 non-MDD specific factors) in the second tier, and a total of 115 items (17 MDD items + 98 non-MDD items).

Results

Cai (2010a) proposed the two-tier full-information item factor analysis model to address data structures wherein there are multiple correlated primary factors rather than the single underlying dimension that characterizes the bifactor IRT and testlet response models. Rather than fitting separate bifactor models for each primary dimension, the two-tier model utilizes the correlation between primary dimensions to improve the accuracy of the parameter estimates. A two-tier model was therefore fit to the PDSQ data by specifying that the first primary dimension accounted for the MDD items and the second primary dimension accounted for the remaining items/subdomains.

This model was evaluated using the flexMIRT (version 2.00; Cai, 2013) multidimensional item analysis software. The PDSQ utilizes a binary response scale, so each item was fit to the two-parameter logistic (2PL) model shown earlier (Equation (10.4)). Parameter estimates were computed using the Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010b) algorithm. The MH-RM estimation algorithm is ideal for the type of high-dimensionality factor structure that is present in the PDSQ, especially when computational burden is a concern. To further aid in the estimation process, a beta prior of (2,1) was specified for each of the 115 items. Finally, for identification purposes, an equality constraint was set for each of the specific dimension doublets among the MDD items.

Item Analysis

Table 10.2 displays the item parameter estimates of the first primary dimension (MDD) and Table 10.3 presents the estimates of the second primary dimension (non-MDD subdomains). These results are presented in separate tables for the sake of clarity, but the parameter estimates were produced by the same (two-tier) model. It is important to note that the covariance between the two primary dimensions was estimated at 0.45. This value indicates that the MDD subdomain is not completely unrelated to the other subdomains, nor are the two primary dimensions so strongly correlated that they can be thought of as measuring the exact same construct. In other words, the two primary dimensions in this model are related, but not identical.

Although the results in Tables 10.2 and 10.3 are too extensive to discuss in detail here, there are a few broad conclusions that can be drawn. First, the vast majority of the items were quite discriminating with regard to their respective primary and specific dimensions. Of the 230 estimated slope parameters (115 primary dimension estimates + 115 specific dimension estimates), 196 were at least 1.0 and the lowest slope overall was estimated at 0.49. On the first primary dimension, the least discriminating item was MDD_05, which asked about a decrease in appetite, and the most discriminating item was MDD_12, which addressed feeling negatively about oneself. On the second primary dimension, the least discriminating item was MAN_01 ("During the past six months, did you feel excessively cheerful and happy, much more than usual, and the good mood lasted most of the day for at least several days?"); the most discriminating item was AGO_11 ("Did you avoid [certain] situations because they make you feel anxious or fearful?"). Overall, the most discriminating subdomains of the second primary dimension were agoraphobia, panic disorder, and post-traumatic stress disorder.

Table 10.2 Slope and Intercept Parameter Estimates of the First Primary Dimension

	Item	*Primary Slope*	*Specific Slope*	*Threshold*
General Sadness	MDD_01	2.18	1.60	–2.55
	MDD_02	2.03	1.60	–.87
Lack of Interest	MDD_03	1.96	2.68	–1.27
	MDD_04	1.71	2.17	–1.33
	MDD_10	.79	.56	.32
Physical Effects	MDD_05	.53	.88	–.23
	MDD_07	.58	1.43	.26
	MDD_09	.70	.70	–2.34
Self-Loathing	MDD_11	1.16	.87	–.91
	MDD_12	3.73	4.18	–.71
	MDD_13	2.29	1.46	–.45
Poor Concentration	MDD_14	1.51	1.90	–1.02
	MDD_15	1.74	1.90	–.62
Suicidal Ideation	MDD_16	2.15	1.86	.25
	MDD_18	2.09	1.86	.38
Suicidal Intent	MDD_19	2.13	2.33	.30
	MDD_20	2.04	2.33	1.41

Note: $N = 3999$. All estimates are significant at $p < 0.05$.

Tables 10.2 and 10.3 also present the thresholds for each item. These values were found by taking the intercept parameter estimates provided by flexMIRT and transforming them to threshold values using the equation:

$$B_i = \frac{-c_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}, \tag{10.5}$$

Table 10.3 Slope and Intercept Parameter Estimates of the Second Primary Dimension

	Item	*Primary Slope*	*Specific Slope*	*Threshold*		*Item*	*Primary Slope*	*Specific Slope*	*Threshold*
Dysthymia	DYS_01	1.28	2.79	–.42	*Psychosis*	PSY_01	1.35	1.20	1.60
	DYS_02	.87	1.38	–.08		PSY_02	1.77	1.63	1.01
	DYS_03	.99	1.88	–.66		PSY_03	1.79	1.27	1.91
	DYS_04	1.18	2.85	–.57		PSY_04	1.24	1.26	2.82
	DYS_05	1.27	2.24	–.24		PSY_05	1.63	1.40	2.43
	DYS_06	1.19	1.88	–.69		PSY_06	1.39	1.17	2.22
	DYS_07	1.23	1.75	–.33					
					Agoraphobia	AGO_01	2.00	.73	.82
Post-traumatic Stress	PTS_01	.69	1.09	.12		AGO_02	2.05	1.38	1.21
	PTS_02	.64	.85	.64		AGO_03	2.80	2.13	.55
	PTS_03	1.81	3.00	.20		AGO_04	2.40	1.58	.87
	PTS_04	2.00	3.21	.39		AGO_05	1.49	1.17	1.44
	PTS_05	2.39	3.77	.24		AGO_06	1.57	1.44	1.34
	PTS_06	2.47	3.59	.30		AGO_07	1.56	1.11	1.44
	PTS_07	2.01	3.09	.13		AGO_08	1.30	.49	1.52
	PTS_08	1.85	2.37	.48		AGO_09	2.19	1.45	1.93
	PTS_09	1.91	2.27	.66		AGO_10	2.78	2.18	.62
	PTS_10	2.17	2.16	.73		AGO_11	2.81	2.27	.45
	PTS_11	2.01	2.98	.53					
	PTS_12	1.61	2.39	.57	*Social Phobia*	SOC_01	1.84	2.15	.08
	PTS_13	1.49	1.90	.88		SOC_02	1.77	2.25	–.03
	PTS_14	1.70	2.42	.26		SOC_03	1.65	2.27	–.14
	PTS_15	1.92	2.15	.72		SOC_04	1.92	2.19	.11
						SOC_05	2.30	2.45	.36
Bulimia	BUL_05	.61	1.54	1.28		SOC_06	1.12	1.73	.14
	BUL_08	.87	2.05	1.53		SOC_07	1.27	1.27	1.12
	BUL_09	1.00	2.34	1.95		SOC_08	1.16	.83	1.95
	BUL_10	.58	1.92	.19		SOC_09	1.07	.95	1.64
						SOC_10	2.14	3.38	.15
Obsessive-Compulsive	OCD_01	1.23	1.03	1.80		SOC_11	1.63	2.53	.29
	OCD_02	1.46	.97	.72		SOC_12	.96	1.47	.97
	OCD_03	1.52	.70	.76		SOC_13	1.86	2.78	.40
	OCD_04	1.79	2.19	1.39		SOC_14	2.01	2.30	.25
	OCD_05	1.59	2.04	1.16		SOC_15	1.92	1.99	.22
	OCD_06	1.47	1.68	1.52					
	OCD_07	1.92	2.27	1.03		GAD_01	1.52	1.03	–.07
	OCD_08	1.34	1.36	1.74		GAD_02	1.53	.84	–.11

(Continued)

Table 10.3 (*Continued*)

						GAD_03	1.32	.95	–.39
	PAN_01	1.81	2.51	.59		GAD_04	1.88	2.21	–.54
	PAN_02	1.95	2.54	.75	*Generalized Anxiety*	GAD_05	2.36	2.51	–.57
	PAN_03	1.90	2.14	.66		GAD_06	1.22	1.12	–.66
	PAN_04	1.87	1.18	.25		GAD_07	1.34	.97	–.81
	PAN_05	2.08	1.13	.49		GAD_08	1.95	1.82	–.84
	PAN_06	2.37	1.97	.34		GAD_09	1.07	.94	–1.12
Panic	PAN_07	2.09	1.07	.83		GAD_10	2.17	2.49	–.57
	PAN_08	1.89	.58	.58					
						SOM_01	.78	.95	.32
	MAN_01	.52	2.84	1.30	*Somatoform*	SOM_02	.99	1.25	–.20
	MAN_02	.57	3.45	1.32		SOM_03	1.18	2.33	1.01
	MAN_03	.64	3.22	1.58		SOM_04	1.01	1.70	1.63
	MAN_04	.96	1.52	1.44		SOM_05	.90	1.04	1.61
Mania	MAN_05	.60	1.48	1.48					
	MAN_06	.77	1.08	1.66	*Hypochondria*	HYP_01	1.62	1.89	1.33
						HYP_02	2.44	2.76	1.18
						HYP_03	1.85	2.74	1.39

Note: N = 3999. All estimates are significant at $p < 0.05$.

where c_i is the intercept parameter for item i, m is the number of dimensions, and $a_{ik}{}^2$ is the squared slope parameter of item i on dimension k (Reckase, 2009). B_i is a multidimensional analog to the standard b ("difficulty" or "severity") parameter in unidimensional IRT. Whereas b indicates the steepest point on an item's trace line, B indicates the steepest point on a multidimensional surface. The interpretation, however, is identical: items with low B values are "easier" or "less severe," meaning there is a high probability of endorsement even when the vector of θ estimates is low, and items with high B values are "more difficult" or "more severe," meaning there is a high probability of endorsement only when the vector of θ estimates is high.

The threshold values in Table 10.2 reveal that the least severe depression item was MDD_01 (B = –2.55), which dealt with general feelings of sadness, while the most severe, by far, was MDD_20 (B = 1.41), which directly addressed suicidal intent. Regarding the non-MDD primary dimension shown in Table 10.3, the least severe item was GAD_09 (B = –1.12), which asked about stress-induced irritability, and the most severe was PSY_04 (B = 2.85), which inquired about the perception of special powers. Overall, the least severe subdomains of the second primary dimension were dysthymia and generalized anxiety disorder, while the most severe subdomains were psychosis and mania.

Although these results support the application of a two-tier model to the PDSQ data, it is useful to examine this model relative to an alternative structure. By reparameterizing the IRT estimates in factor analytic terms (see Wirth & Edwards, 2007), we can make a more direct comparison of the two-tier model with the bifactor results reported by Gibbons, Rush, and Immekus (2009). Table 10.4 displays the improvement in item factor loadings that is provided by the two-tier model. Clearly, allowing the depression items to load on a separate primary dimension drastically increases the item factor loadings. There is not a single item that loads higher on the primary dimension in the bifactor solution, and many items go from having negligible loadings in the bifactor model to quite strong loadings in the two-tier model. To be more precise, the mean loading of the MDD items on the primary dimension in the bifactor model was just 0.26; the mean loading on the primary dimension in the two-tier model was 0.53. Furthermore, we found that the loadings on the second primary dimension were quite similar to the non-MDD general factor loadings of the bifactor model. Hence, the two-tier model was able to boost the relatively weak loadings of the MDD items while preserving the strong loadings of the non-MDD items.

Regarding the specific dimensions, although some loadings were higher in the bifactor model (particularly the suicide items—MDD_16 through MDD_20), there was a marked lack of uniformity among these loadings; of the 17 MDD items in Table 10.4, six of the specific loadings in the bifactor model were below 0.30 while four were above 0.70

Table 10.4 A Comparison of the Item Factor Loadings of the PDSQ Major Depressive Disorder (MDD) Subscale in a Bifactor Model (as reported by Gibbons, Rush, & Immekus (2009)) and a Two-Tier Item Factor Model

	Primary loadings		*Specific loadings*							
	Bifactor	*Two-tier*	*Bifactor*	*Two-tier*						
				1	2	3	4	5	6	7
MDD_01	.17	.68	.52	.50						
MDD_02	.26	.66	.43	.52						
MDD_03	.17	.53	.39		.72					
MDD_04	.20	.53	.35		.67					
MDD_10	.18	.27	.24		.29					
MDD_05	.19	.25	.13			.44				
MDD_07	.25	.35	.08			.62				
MDD_09	.39	.40	.09			.35				
MDD_11	.36	.52	.33				.39			
MDD_12	.30	.64	.49				.71			
MDD_13	.34	.71	.54				.46			
MDD_14	.34	.51	.22					.64		
MDD_15	.33	.56	.28					.62		
MDD_16	.26	.65	.73						.56	
MDD_18	.22	.64	.85						.57	
MDD_19	.26	.59	.75							.65
MDD_20	.23	.58	.77							.66

Note: N = 3999. Bifactor item factor loadings were reported in Gibbons, Rush, and Immekus (2009).

(range = 0.77). The two-tier model addresses this issue by explicitly modeling the multidimensionality inherent in the MDD subscale. The two-tier specific loadings in Table 10.4 are therefore much more homogeneous: the smallest specific loading (MDD_10) is 0.29 and the largest (MDD_12) is 0.71, resulting in a narrower range of 0.42. Further, the mean specific dimension loading increased from 0.42 in the bifactor model to 0.55 in the two-tier model.

Scoring

In addition to improving the parameter estimates of the MDD items, the two-tier model also provides more accurate IRT-scaled scores. Posterior expectations can be computed for each of the $\boldsymbol{\eta}$ primary dimensions and $\boldsymbol{\xi}$ specific dimensions for every individual item response pattern. Estimation of these expected *a posteriori* (EAP) scores requires numerical integration of each latent variable (see Thissen & Wainer, 2001), and the restrictions of the two-tier model provide greater computational efficiency. The mathematical specifics of two-tier EAP estimation will not be discussed in this chapter; see Appendix B in Cai (2010a) for further details regarding EAP estimation in the two-tier IFA model (and Chapter 15 in this volume for scoring in commonly applied multidimensional models).

Cai (2010a) demonstrated that the EAP scores obtained from a two-tier model are almost identical to scores obtained from fitting two separate bifactor models, and that the two-tier scores are more precise and reliable. Specifically, he fit item responses from a combined math/reading test to both models and then compared just the math IRT-scaled scores from a two-tier model against the IRT-scaled scores found by treating the math items as a separate bifactor structure from the reading items. The EAPs were strongly correlated (r = 0.96), but the average standard error of measurement in the two-tier model was 20 percent lower than in the bifactor model (0.38 versus 0.49, respectively). Further, the standard deviation of the individual standard errors of measurement was lower in the two-tier model (0.05) than in the bifactor model (0.07). This improvement in precision is attributed to the “borrowing of strength” that characterizes the two-tier model; by utilizing information from one primary dimension, the two-tier model is better able to differentiate individuals along the other primary dimension (Cai, 2010a).

For the PDSQ analysis, EAPs were computed for each individual across all 21 dimensions, thereby allowing us to rank individuals according to their relative latent trait levels. Once these IRT-scaled scores have been calculated, it becomes possible to estimate an individual’s EAP from his or her raw summed score. Table 10.5 displays the summed score to IRT-scaled score conversion for the MDD and non-MDD dimensions in the two-tier model. Using this table, a practitioner could simply sum a patient’s score on the (reduced) PDSQ,[4] look up the value in the “Summed Score” column, and locate the corresponding IRT-scaled EAP score on each primary dimension. For example, a summed score of 59 is associated with an EAP score of 0.42 (SD = 0.86) on the first primary dimension and 0.75 (SD = 0.41) on the second primary dimension. The posterior standard deviations shown in Table 10.5 indicate that the EAPs associated with the non-MDD primary dimension are more precise than those associated with the MDD primary dimension (which was expected because of the longer test length of the non-MDD dimension). Although not

4 It should be noted once again that this conversion table is based on our truncated version of the PDSQ. The IRT-scaled scores can only be used if the summed score is calculated *without* including items MDD_06, 08, 17, and 21, BUL_01, 02, 03, 04, 06, and 07, HYP_01 and 02, or any of the ALC and DRUG items.

Table 10.5 Summed Score to IRT-Scaled Score Conversion for the Primary Dimensions

Summed Score	*Primary 1*		*Primary 2*		*Summed Score*	*Primary 1*		*Primary 2*	
	EAP	*SD*	*EAP*	*SD*		*EAP*	*SD*	*EAP*	*SD*
0	–3.09	.63	–2.92	.66	59	.42	.86	.75	.41
1	–2.84	.64	–2.78	.65	60	.44	.86	.79	.41
2	–2.62	.64	–2.65	.64	61	.46	.86	.83	.41
3	–2.43	.64	–2.53	.63	62	.48	.86	.87	.41
4	–2.26	.65	–2.42	.62	63	.50	.86	.92	.41
5	–2.10	.65	–2.33	.62	64	.52	.86	.96	.41
6	–1.95	.66	–2.23	.61	65	.55	.86	1.00	.41
7	–1.81	.67	–2.14	.60	66	.57	.86	1.04	.41
8	–1.68	.68	–2.05	.60	67	.59	.86	1.08	.41
9	–1.56	.69	–1.97	.59	68	.61	.86	1.13	.41
10	–1.45	.71	–1.89	.58	69	.63	.86	1.17	.41
11	–1.34	.72	–1.81	.58	70	.65	.86	1.21	.41
12	–1.24	.74	–1.73	.57	71	.68	.86	1.26	.41
13	–1.15	.75	–1.65	.56	72	.70	.86	1.30	.41
14	–1.06	.77	–1.58	.56	73	.72	.86	1.34	.40
15	–.98	.78	–1.50	.55	74	.75	.85	1.39	.40
16	–.91	.79	–1.43	.54	75	.77	.85	1.43	.40
17	–.84	.80	–1.36	.53	76	.79	.85	1.48	.40
18	–.78	.81	–1.29	.53	77	.82	.85	1.52	.40
19	–.72	.82	–1.22	.52	78	.84	.85	1.57	.40
20	–.67	.82	–1.15	.51	79	.87	.85	1.62	.40
21	–.62	.83	–1.08	.51	80	.90	.85	1.67	.40
22	–.57	.83	–1.02	.50	81	.92	.84	1.71	.40
23	–.52	.84	–.95	.50	82	.95	.84	1.76	.41
24	–.48	.84	–.89	.49	83	.98	.84	1.81	.41
25	–.44	.84	–.83	.49	84	1.01	.84	1.87	.41
26	–.40	.85	–.77	.48	85	1.04	.84	1.92	.41
27	–.36	.85	–.72	.48	86	1.07	.84	1.97	.41
28	–.33	.85	–.66	.48	87	1.10	.83	2.03	.41
29	–.29	.85	–.61	.47	88	1.14	.83	2.08	.41
30	–.26	.86	–.55	.47	89	1.17	.83	2.14	.41
31	–.23	.86	–.50	.47	90	1.20	.83	2.20	.42
32	–.20	.86	–.45	.46	91	1.24	.83	2.26	.42
33	–.17	.86	–.40	.46	92	1.28	.82	2.32	.42
34	–.15	.86	–.35	.46	93	1.31	.82	2.38	.42
35	–.12	.86	–.30	.46	94	1.35	.82	2.44	.43

(Continued)

Table 10.5 (Continued)

Summed Score	*Primary 1*		*Primary 2*		*Summed Score*	*Primary 1*		*Primary 2*	
	EAP	*SD*	*EAP*	*SD*		*EAP*	*SD*	*EAP*	*SD*
36	–.09	.86	–.25	.45	95	1.39	.82	2.51	.43
37	–.07	.86	–.20	.45	96	1.44	.81	2.58	.44
38	–.04	.86	–.16	.45	97	1.48	.81	2.64	.44
39	–.02	.86	–.11	.45	98	1.52	.81	2.71	.44
40	.01	.86	–.06	.44	99	1.57	.81	2.79	.45
41	.03	.87	–.02	.44	100	1.62	.80	2.86	.46
42	.05	.87	.03	.44	101	1.67	.80	2.93	.46
43	.08	.87	.07	.44	102	1.72	.80	3.01	.47
44	.10	.87	.12	.44	103	1.77	.80	3.09	.47
45	.12	.87	.16	.43	104	1.82	.79	3.17	.48
46	.15	.87	.20	.43	105	1.88	.79	3.25	.49
47	.17	.87	.25	.43	106	1.94	.79	3.33	.49
48	.19	.87	.29	.43	107	2.00	.78	3.41	.50
49	.21	.87	.33	.43	108	2.06	.78	3.50	.51
50	.23	.87	.38	.43	109	2.12	.78	3.58	.51
51	.25	.87	.42	.42	110	2.19	.77	3.67	.52
52	.27	.87	.46	.42	111	2.27	.76	3.76	.52
53	.30	.87	.50	.42	112	2.35	.76	3.86	.53
54	.32	.87	.54	.42	113	2.44	.75	3.97	.53
55	.34	.87	.58	.42	114	2.55	.75	4.10	.54
56	.36	.87	.63	.42	115	2.68	.74	4.23	.55
57	.38	.86	.67	.42					

Note: EAP = Expected *a posteriori* IRT-scaled score.

displayed here, the flexMIRT software program also provides summed score-to-EAP conversion tables for every combination of primary and specific dimensions.

Goodness of Fit

The validity of any measurement model is dependent on the degree to which the model fits the data. Unfortunately, the global goodness-of-fit of a full-information item factor analytic model is difficult to evaluate. Bock and Aitkin (1981) showed that the marginal maximum likelihood estimation of item parameters is based on an underlying contingency table of the fully cross-classified item responses. As the number of test items increases, this contingency table can become sparse (see Bartholomew & Tzamourani, 1999), and this sparseness disallows the use of the asymptotic chi-square approximation on which full-information fit statistics are based. Thus, evaluating full-information goodness-of-fit is especially problematic with regard to a scale as lengthy as the PDSQ.

Table 10.6 Relative Fit Statistics of the Bifactor and Two-Tier Item Factor Analysis Models

	Bifactor model	*Two-tier model*
–2LL	379,155.59	377,568.97
AIC	379,837.59	378,252.97
BIC	381,983.77	380,405.45

Note: Statistics are based on the 115-item revised PDSQ. –2LL = –2loglikelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Fortunately, the sparseness of the underlying contingency table does not invalidate the use of the likelihood ratio difference statistic for assessing the relative fit between nested models with similar dimensionality (see Haberman, 1977; Maydeu-Olivares & Cai, 2006). Gibbons, Rush, and Immekus (2009) investigated the relative fit of the (full 139-item) PDSQ by comparing the –2loglikelihood (–2LL) values derived from three competing IFA models: a unidimensional model, a simple structure model with 15 uncorrelated traits (to account for each of the 15 psychiatric subdomains), and a bifactor model. They then conducted a chi-square difference test (based on the –2LL value) for each pair of models (see Gibbons et al., 2007 for a discussion of model comparison tests). The authors concluded that the bifactor model produced significantly improved fit over both the unidimensional model, which ignored the specific subdomain factors, and the simple structure model, which ignored the primary psychiatric impairment factor.

We cannot directly compare our results with those reported by Gibbons, Rush, and Immekus (2009) because the MH-RM estimation of the two-tier model required the deletion of several items. However, fitting a bifactor model to the truncated (115-item) PDSQ data would facilitate a relative fit comparison with the two-tier model. The results of this comparison are displayed in Table 10.6, which reports the –2LL value as well as the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). For all three statistics, lower values indicate a better-fitting model. Not only does the –2LL statistic indicate that the two-tier model fits the data better than the bifactor model, but the AIC and BIC, which are specifically designed to penalize for model complexity, also indicate that the two-tier model achieves better fit.

The difference in degrees of freedom between a bifactor model and a two-tier model is equal to the number of covariances between primary dimensions. The proposed PDSQ two-tier structure includes two primary dimensions, and thus a single covariance; the likelihood ratio difference test between the bifactor and two-tier representations of the PDSQ was therefore based on a single degree of freedom. In terms of relative fit, the two-tier model provided a significant improvement over the bifactor model, $\chi^2_{LR}(1) = 379{,}155.59 - 377{,}568.97 = 1{,}586.62$, $p < 0.001$.

Although full-information fit measures have not yet been developed for two-tier models, Cai (2010a) notes that recent advances in limited-information goodness-of-fit assessment (e.g., Cai et al., 2006; Maydeu-Olivares, Chapter 6 of this volume; Maydeu-Olivares & Joe, 2005, 2006; Orlando & Thissen, 2000) can be used in the development of new global fit statistics for the two-tier model. Further, model misfit diagnostic statistics such as Chen and Thissen's (1997) local dependence index and the $S - X^2$ item fit statistic (Orlando & Thissen, 2000) may be extended for use with the two-tier model. Until such tools are introduced, applications of the two-tier IFA model must rely on relative, rather than absolute, fit statistics.

Summary

The two-tier full-information item factor analysis model is a notable development in the modeling of item response patterns for a number of reasons. First, the two-tier IFA structure encompasses the standard correlated-traits multidimensional IRT model, the testlet response model, and the item bifactor model. Second, the two-tier model is flexible; the ability to specify correlated primary dimensions results in more accurate measurement models (as demonstrated by the PDSQ analysis) and allows for longitudinal IRT modeling and analyses of idiosyncratic response style. Third, the two-tier model provides EAP scores that are more precise and reliable than those computed from a bifactor IRT model. Further, user specification of the two-tier model constraints is relatively straightforward in statistical software programs such as flexMIRT (version 2.00). In fact, the seemingly complex two-tier model of the PDSQ only differed from a bifactor IRT model in two basic ways; (1) by specifying a second primary dimension, and (2) by estimating the covariance between the primary dimensions. Overall, the data analysis presented in this chapter demonstrates that the two-tier full-information item factor analysis model is a powerful and flexible model that is particularly well equipped to handle complex data.[5]

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Akademinai Kiado.

Bartholomew, D.J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research*, *27*(4), 525–546.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*(1), 33–57.

Cai, L. (2013). flexMIRT® version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring. [Computer software]. Chapel Hill, NC: Psychometric Group.

Cai, L., Maydeu-Olivares, A., Coffman, D.L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.

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.

Fava, M., Rush, A.J., Trivedi, M.H., Nierenberg, A.A., Thase, M.E., Sackeim, H.A., . . . Kupfer, D.J. (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR* D) study. *Psychiatric Clinics of North America*, *26*(2), 457–494.

5 Acknowledgments: I would like to thank Scott Monroe for assistance with tuning the Metropolis-Hastings Robbins-Monro estimation algorithm and Li Cai for his valuable input and advice. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B080016 to the University of California, Los Angeles. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

Gibbons, R.D., Bock, R.D., Hedeker, D., Weiss, D.J., Segawa, E., Bhaumik, D.K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.

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*(4), 401–410.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, 1148–1169.

Hill, C.D. (2006). *Two models for longitudinal item response data*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*(1), 55–64.

Maydeu-Olivares, A., & Coffman, D.L. (2006). Random intercept item factor analysis. *Psychological methods*, *11*(4), 344.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Reckase, M. (2009). *Multidimensional item response theory*. Springer.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*(2), 167–182.

Rush, A.J., Fava, M., Wisniewski, S.R., Lavori, P.W., Trivedi, M.H., Sackeim, H.A., . . . Niederehe, G. (2004). Sequenced treatment alternatives to relieve depression (STAR* D): Rationale and design. *Controlled Clinical Trials*, *25*(1), 119–142.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Thissen, D., Cai, L., & Bock, R.D. (2010). The nominal categories item response model. In M.L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43–75). New York: Routledge.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. New York: Routledge.

Tisak, J., & Meredith, W. (1989). Exploratory longitudinal factor analysis in multiple populations. *Psychometrika*, *54*(2), 261–281.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58.

Zimmerman, M., & Mattia, J.I. (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry*, *58*(8), 787.