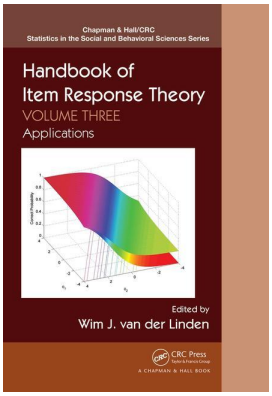


Handbook of Item Response Theory
Volume Three: Applications

Wim J. van der Linden

Dimensionality Analysis

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9781315117430-3
Robert D. Gibbons, Li Cai
Published online on: 15 Dec 2017

3

Dimensionality Analysis

Robert D. Gibbons and Li Cai

CONTENTS

3.1 Introduction

Much of item response theory (IRT) is based on the assumption of unidimensionalty; namely, the associations among the item responses are explained completely by a single underlying latent variable, representing the target construct being measured. While this is often justified in many areas of educational measurement, more recent interest in measuring patient-reported outcomes (Gibbons et al., 2008, 2012) involves items that are drawn from multiple uniquely correlated subdomains violating the usual conditional independence assumption inherent in unidimensional IRT models. As alternatives, both unrestricted item factor analytic models (Bock & Aitkin, 1981) and restricted or confirmatory item factor analytic models (Gibbons & Hedeker, 1992) have been used to accommodate the multidimensionality of constructs for which the unidimensionality assumption is untenable. A concrete example is the measurement of depressive severity, where items are drawn from mood, cognition, and somatic impairment subdomains. While this is a somewhat extreme example, there are many more borderline cases where the choice between a unidimensional model and a multidimensional model is less clear, or the question of how many dimensions is "enough" is of interest. In this chapter, we explore the issue of determining the dimensionality of a particular measurement process. We begin by discussing multidimensional item factor analysis models, and then consider the consequences of incorrectly fitting a unidimensional model to multidimensional data.

We also discuss nonparametric approaches such as DIMTEST (Stout, 1987). We then examine different approaches to testing dimensionality of a given measurement instrument, including approximate or heuristic approaches such as eigenvalue analysis, as well as more statistically rigorous limited-information and full-information alternatives. Finally, we illustrate the use of these various techniques for dimensionality analysis using a relevant example.

3.2 Classical Multiple Factor Analysis of Test Scores

Multiple factor analysis as formulated by Thurstone (1947) assumes that the test scores are continuous measurements standardized to mean zero and standard deviation one in the sample. (Number-right scores on tests with 30 or more items are considered close enough to continuous for practical work.) The Pearson product-moment correlations between all pairs of tests are then sufficient statistics for factor analysis when the population distribution of the scores is multivariate normal. Because the variables are assumed standardized, the mean of the distribution is the null vector and the covariance matrix is a correlation matrix. If the dimensionality of the factor space is D, the assumed statistical model for the pth observed score y is

$$y_p = \alpha_{p1}\theta_1 + \alpha_{p2}\theta_2 + \cdots + \alpha_{pD}\theta_D + \epsilon_j, \tag{3.1}$$

where the underlying vector of latent variables attributable to individual differences is

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D).$$

Just as the observed variables, the latent variables are assumed to follow a standard multivariate normal distribution but are uncorrelated; that is, their covariance matrix is a $D \times D$ identity matrix. The residual term (also called unique factor), ϵ_p, that accounts for all remaining variation in y_p is assumed to be normal with mean 0 and variance $1 - \omega_p^2$, where

$$\omega_p^2 = \sum_{d=1}^{D} \alpha_{pd}^2,$$

which Thurstone called the *communality* of observed variable j. Estimation of the loadings requires the restriction $1 - \omega_p^2 > 0$ to prevent inadmissible solutions (also known as *Heywood* cases). Moreover, the unique factor consists of both systematic variation and error of measurement. Thus, if the reliability of the test is known to be ρ, ω_p^2 cannot be greater than ρ.

On the above assumptions, efficient statistical estimation of the factor loadings from the sample correlation matrix is possible and available in published computer programs. In fact, only the communalities need to be estimated: once the communalities are known, the factor loadings can be calculated directly from the so-called "reduced" correlation matrix via matrix decompositions, in which the diagonal elements of the sample correlation matrix are replaced by the corresponding communalities (Harman, 1967).

3.3 Classical Item Factor Analysis

In item factor analysis, the observed item responses are assigned to one of two-or-more predefined categories. For example, test items marked right or wrong are assigned to dichotomous categories; responses to essay questions may be assigned to ordered polytomous categories (grades) *A, B, C, D* in order of merit; responses in the form of best choice among multiple alternatives may be assigned to nominal polytomous categories.

To adapt the factor analysis model for test scores to the analysis of categorical item responses, we assume that the *y*-variables are also unobservable. We follow Thurstone in referring to these underlying variables as *response processes*. In the dichotomous case, a process gives rise to an observable correct response when y_p exceeds some threshold γ_i specific to item i. On the assumption that y_p is standard normal, γ_i divides the area under the normal curve into two sections corresponding to the probability that a respondent with given value of $\boldsymbol{\theta}$ will respond in the first or second category. Designating the categories 1 and 2, we may express these conditional probabilities given $\boldsymbol{\theta}$ as

$$P_{i1}(\boldsymbol{\theta}) = \Phi(z_p - \gamma_i) \quad \text{and} \quad P_{j2}(\boldsymbol{\theta}) = 1 - \Phi(z_p - \gamma_i),$$

where Φ is the cumulative normal distribution function and

$$z_p = \sum_{d=1}^{D} \alpha_{pd}\theta_d.$$

The *unconditional* response probabilities, on the other hand, are the areas under the standard normal curve above and below $-\gamma_i$ in the population from which the sample of respondents is drawn. The area above this threshold is the classical *item difficulty*, p_i, and the standard normal deviate at p_i is a large sample estimator of $-\gamma_i$ (Lord & Novick, 1968, Chapter 16).

These relationships generalize easily to ordered polytomous categories. Suppose item j has m_j ordered categories; we then replace the single threshold of the dichotomous case with m_{i-1} thresholds, say, $\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{i,m_i-1}$. The category response probabilities conditional on $\boldsymbol{\theta}$ are the m_j areas under the normal curve corresponding to the intervals from minus to plus infinity bounded by the successive thresholds:

$$P_{ih}(\boldsymbol{\theta}) = \Phi(z_p - \gamma_{ih}) - \Phi(z_p - \gamma_{i,h-1}),$$

where $\Phi(z_p - \gamma_{i0}) = 0$ and $\Phi(z_p - \gamma_{im_i}) = 1 - \Phi(z_p - \gamma_{i,m_i-1})$.

Because the response processes are unobserved, their product-moment correlation matrix cannot be calculated directly. Classical methods of multiple factor analysis do not apply directly to item response data. Full maximum likelihood estimation of the item correlation matrix requires calculating the normal orthant probabilities involving integrals over as many dimensions as there are items. While it is theoretically possible, actual computation remains difficult even with modern estimation approaches (Song & Lee, 2003).

However, an approximation to the item correlations can be inferred from the category joint-occurrence frequencies tallied over the responses in the sample. Assuming in the two-dimensional case that the marginal normal distribution of the processes is standard

bivariate normal, the correlation value that best accounts for the observed joint frequencies can be estimated, for example, using pairwise maximum likelihood. If both items are scored dichotomously, the result is the well-known tetrachoric correlation coefficient, an approximation for which was given by Divgi (1979). If one or both items are scored polytomously, the result is the less common polychoric correlation (Jöreskog, 2002). The correlations for all distinct pairs of items can then be assembled into a correlation matrix and unities inserted in the diagonal to obtain an approximation to the item correlation matrix. Because the calculation of tetrachoric and polychoric correlations breaks down if there is a vacant cell in the joint-occurrence table, a small positive value such as 0.5 (i.e., a continuity correction) may be added to each cell of the joint frequency table.

For the purpose of determining dimensionality, the correlation matrix described above can be subjected to the method of principal components or principal factor analysis with iterated communalities (Harman, 1967, p. 87). Classical principal factor analysis of item responses can be useful in its own right, or as a preliminary to more exact and more computationally intensive IRT procedures such as maximum marginal likelihood item factor analysis. In the latter role, the classical method provides a quick way of giving an upper bound on a plausible number of factors in terms of the total amount of association accounted for. It also gives good starting values for the iterative procedures discussed in the following section.

3.4 Item Factor Analysis Based on IRT

IRT-based item factor analysis makes use of all information in the original categorical responses and does not depend on pairwise indices of association such as tetrachoric or polychoric correlation coefficients. For that reason, it is referred to as *full-information* item factor analysis. It works directly with item response models giving the probability of the observed categorical responses as a function of latent variables descriptive of the respondents and parameters descriptive of the individual items. It differs from the classical formulation in its scaling, however, because it does not assume that the response process has unit standard deviation and zero mean; rather, it assumes that the *residual* term has unit standard deviation and zero mean. The latter assumption implies that the response processes have zero mean and standard deviation equal to

$$\sigma_{y_i} = \sqrt{1 + \Sigma_d^D \alpha_{id}^2}.$$

Inasmuch as the scale of the model affects the relative size of the factor loadings and thresholds, we rewrite the model for dichotomous responses in a form in which the factor loadings are replaced by factor slopes, a_{id}, and the threshold is absorbed in the intercept, c_i:

$$y_i = \sum_{d=1}^{D} a_{id}\theta_d + c_i + \epsilon_i.$$

To convert factor slopes into loadings, we divide by the above standard deviation and similarly convert the intercepts to thresholds:

$$\alpha_{id} = a_{id}/\sigma_{y_i} \quad \text{and} \quad \gamma_i = -c_i/\sigma_{y_i}.$$

Conversely, to convert to factor analysis units, we change the standard deviation of the residual from one to

$$\sigma^*_{\epsilon_i} = \sqrt{1 - \Sigma^D_d \alpha^2_{id}},$$

and change the scale of the slopes and intercept accordingly:

$$a_{id} = \alpha_{id}/\sigma_{\epsilon^*_i} \quad \text{and} \quad c_i = -\gamma_i/\sigma^*_{\epsilon_i}.$$

For polytomous responses, the model generalizes as

$$z_i = \sum_{d=1}^{D} a_{id}\theta_d,$$

$$P_{ih}(\theta) = \Phi(z_i + c_{ih}) - \Phi(z_i + c_{i,h-1}),$$

where $\Phi(z_i + c_{i0}) = 0$ and $\Phi(z_i + c_{im_i}) = 1 - \Phi(z_i + c_{i,m_i-1})$ as previously.

In the context of item factor analysis, this is the multidimensional generalization of the *graded* model introduced by Samejima (1969). Similarly, the *rating scale* model of Andrich (1978), in which all items have the same number of categories and the thresholds are assumed to have the same spacing but may differ in overall location, can be generalized by setting the above linear form to $z_i + e_i + c_h$, where e_i is the location intercept.

3.5 Maximum Likelihood Estimation of Item Slopes and Intercepts

There is a long history, going back to Fechner (1860), of methods for estimating the slope and intercept parameters of models similar to the above—that is, models in which the response process is normally distributed and the deviate is a linear form. These so-called *normal transform* models differ importantly from the IRT models, however, in assuming that the $\boldsymbol{\theta}$ variables are manifest measurements of either observed or experimentally manipulated variables. In Fechner's classic study of the sensory discrimination thresholds for lifted weights, the subjects were required to lift successively each of a series of two small, identical appearing weights differing by fixed amounts and say which feels heavier. Fechner fitted graphically the inverse normal transforms of the proportion of subjects who answered correctly and used the slope of the fitted line to estimate the standard deviation as a measure of sensory discrimination. Much later, R. A. Fisher (Bliss, 1935) provided a maximum likelihood method of fitting similar functions used in the field of toxicology to determine the so-called 50% lethal dose of pesticides. This method eventually became known as probit analysis (Bock & Jones, 1968; for behavioral applications, see Finney, 1952).

To apply Fisher's method of analysis to item factor analysis, one must find a way around the difficulty that the variable values (i.e., the θs) in the linear predictor are unobservable. The key to solving this problem lies in assuming that the values have a specifiable distribution in the population from which the respondents are drawn (Bock & Lieberman, 1970). This allows us to integrate over that distribution numerically to estimate the expected numbers of respondents located at given points in the latent space who respond in each of the categories. These expected values can then be subjected to a multidimensional version of probit analysis. The so-called EM method of solving this type of estimation

problem (Aitkin, Volume Two, Chapter 12; Bock & Aitkin, 1981) is an iterative procedure starting from given initial values. It involves calculating expectations (the E-step) that depend on both the parameters and the observations, followed by likelihood maximization (the M-step) that depends on the expectations. These iterations can be shown to converge on the maximum likelihood estimates under very general conditions (Dempster et al., 1977). In IRT and similar applications, this approach is called *maximum marginal likelihood* estimation because it works with the marginal probabilities of response rather than the conditional probabilities (Glas, Volume Two, Chapter 11). Details in the context of item factor analysis are given in Bock and Aitkin (1981) and Bock and Gibbons (2010, Appendix).

3.6 Confirmatory Item Factor Analysis and the Bifactor Pattern

There are two major limitations of the unrestricted or exploratory factor analysis model described above. First, interpretation of the final solution depends on selecting the appropriate rotation of the final solution (e.g., varimax, quartimin, etc.; for a review, see Browne, 2001). Second, modern simulation-based estimation approaches notwithstanding (e.g., Cai, 2010a,b), the full-information IRT approach remains demanding in terms of the number of dimensions that can be evaluated because the computational complexity associated with the integrals in the likelihood equations is exponentially increasing in the number of factors. In confirmatory factor analysis, the first limitation (indeterminacy due to rotation) is resolved by assigning arbitrary fixed values to certain loadings of each factor during maximum likelihood estimation. In general, fixing of loadings will imply nonzero correlations of the latent variables, but this does not invalidate the analysis. The correlations may also be estimated if desired by selecting an oblique rotation criterion. An important example of confirmatory item factor analysis—which resolves the second problem of limitation of the number of dimensions that can be numerically evaluated—is the bifactor patterns for general and group factors, which applies to tests and scales with item content drawn from several well-defined subareas of the domain in question. Two prominent examples are tests of educational achievement consisting of reading, mathematics and science areas, and self-reports of health status covering both physical and emotional impairment. The main objective in the use of such instruments is to estimate a single score measuring, in these examples, general educational achievement or overall health status.

To analyze these kinds of structures for dichotomously scored item responses, Gibbons and Hedeker (1992) developed full-information item bifactor analysis for binary item responses, and Gibbons extended it to the polytomous case (Gibbons et al., 2007). Cai et al. (2011) further generalized the model to handle multiple groups. To illustrate, consider a set of n test items for which a D-factor solution exists with one general factor and $D-1$ group or method-related factors. The bifactor solution constrains each item j to a nonzero loading α_{i1} on the primary dimension and a second loading ($\alpha_{id}, d = 2, \ldots, D$) on not more than one of the $D-1$ group factors. For four items, the bifactor pattern matrix might be

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}.$$

This structure, which Holzinger and Swineford (1937) termed the "bifactor" pattern, also appears in the inter-battery factor analysis of Tucker (1958) and is one of the confirmatory factor analysis models considered by Jöreskog (1969). In the latter case, the model is restricted to test scores assumed to be continuously distributed. However, the bifactor pattern might also arise at the item level (Muthén, 1989). Gibbons and Hedeker (1992) showed that paragraph comprehension tests, where the primary dimension represents the targeted process skill and additional factors describe content area knowledge within paragraphs, were described well by the bifactor model. In this context, they showed that items were conditionally independent between paragraphs, but conditionally dependent within paragraphs. More recently, the bifactor model has been applied to problems in patient-reported outcomes in physical and mental health measurement (Gibbons et al., 2008, 2012, 2014). As shown by Gibbons and Hedeker (1992), the bifactor model always reduces the dimensionality of the likelihood equation to two, regardless of the number of secondary factors.

In the bifactor case, the graded response model (Gibbons et al., 2007) is

$$z_{ih}(\boldsymbol{\theta}) = \sum_{d=1}^{D} a_{id}\theta_d + c_{ih}, \tag{3.2}$$

where only one of the $d = 2, \ldots, d$ values of a_{id} is nonzero in addition to a_{i1}. Assuming independence of the $\boldsymbol{\theta}$, in the unrestricted case, the multidimensional model above would require a d-fold integral in order to compute the unconditional probability for response pattern $\boldsymbol{U}$, that is,

$$P\left(\boldsymbol{U} = \bigcup_p\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L_i(\boldsymbol{\theta})g(\theta_1)g(\theta_2)\ldots g(\theta_D)d\theta_1 d\theta_2 \ldots d\theta_D, \tag{3.3}$$

where $L_p(\boldsymbol{\theta})$ is the likelihood of $\bigcup_p$ conditional on $\boldsymbol{\theta}$. The corresponding unconditional or marginal probability for the bifactor model reduces to

$$P\left(\boldsymbol{U} = \bigcup_p\right) = \int_{-\infty}^{\infty}\left\{\prod_{d=2}^{D}\int_{-\infty}^{\infty}\left[\prod_{i=1}^{n}\prod_{h=1}^{m_i}[\Phi_{ih}(\theta_1,\theta_d) - \Phi_{ih-1}(\theta_1,\theta_d)]^{u_{ijh}}\right]g(\theta_d)d\theta_d\right\}g(\theta_1)d\theta_1, \tag{3.4}$$

which can be approximated to any degree of practical accuracy using two-dimensional Gauss–Hermite quadrature, since for both the binary and graded bifactor response models, the dimensionality of the integral is two regardless of the number of subdomains (i.e., $D - 1$) that comprised the scale.

3.7 Unidimensional Models and Multidimensional Data

A natural question is whether there is any adverse consequence of applying unidimensional IRT models to multidimensional data. To answer this question, Stout and

coworkers (e.g., Stout, 1987; Zhang & Stout, 1999) took a distinctly nonparametric approach to characterize the specific conditions under which multidimensional data may be reasonably well represented by a unidimensional latent variable. They emphasized a core concept that subsequently became the basis of a family of theoretical and practice devices for studying dimensionality, namely, local independence as expressed using conditional covariances. To begin, given n items in a test, the strong form of local independence states that the conditional response pattern probability factors into a product of conditional item response probabilities, that is,

$$P\left(\boldsymbol{U} = \bigcup | \boldsymbol{\theta}\right) = \prod_{i=1}^{n} P(U_i = u_i | \boldsymbol{\theta}).$$

Correspondingly, a test is weakly locally independent with respect to $\boldsymbol{\theta}$ if the conditional covariance is zero for all item pairs i and i′:

$$Cov(U_i, U_{i'} | \boldsymbol{\theta}) = 0.$$

The conditional covariances provide a convenient mechanism to formalize the notion of an *essentially unidimensional* test that possesses one essential dimension and (possibly) a number of nuisance dimensions. A test is said to be essentially independent (Stout, 1990) with respect to $\boldsymbol{\theta}$ if

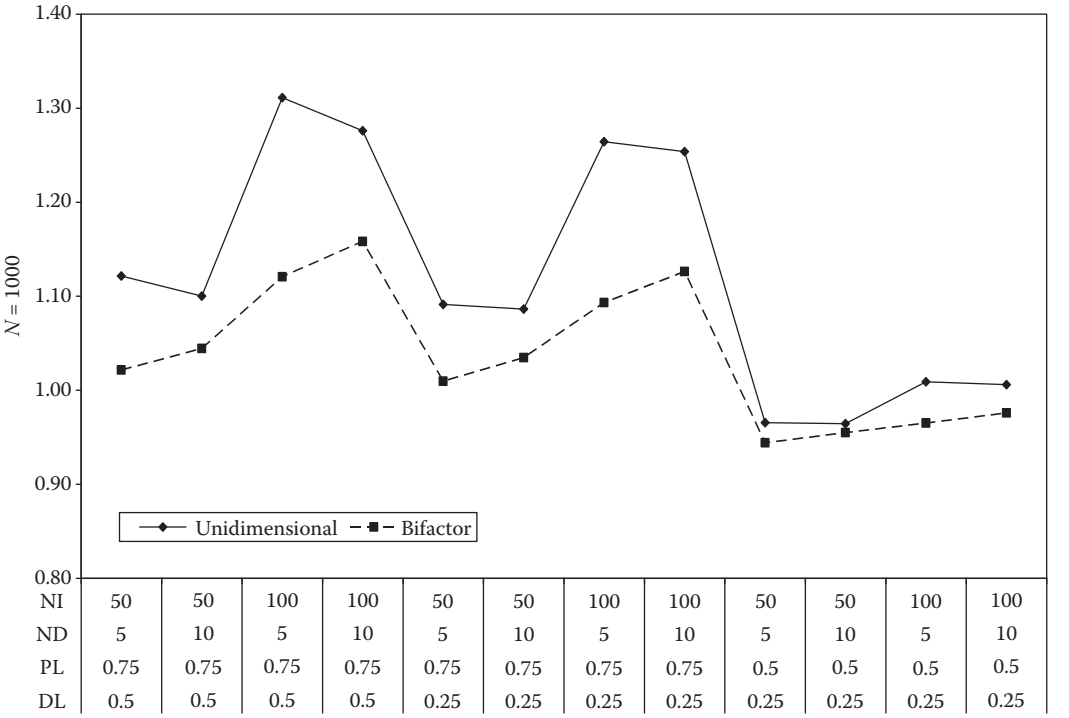
$$D_n(\boldsymbol{\theta}) = \frac{\sum_{1 \le i < i' \le n} |Cov(U_i, U_{i'} | \boldsymbol{\theta})|}{\binom{n}{2}} \to 0, \text{ as } n \to \infty.$$

In other words, essential independence states that the average value of conditional covariances across all item pairs is small as the test length increases. If the minimal dimensionality of $\boldsymbol{\theta}$ necessary for an item pool to satisfy essentially independence is equal one, then the test is said to be essentially unidimensional.

The mathematical condition above suggests a statistical procedure for testing essential unidimensionality (Stout, 1987). In brief, the test is split into two subsets called assessment tests (AT1 and AT2), and a longer subset called the partitioning test (PT). The items for AT1 are chosen to be saturated with the same dominant latent trait, but are as dimensionally different as possible from the items in the PT. Then AT2 is selected such that the items have similar difficulty as AT1. Each test taker's total score on the PT is used to group the test takers into several homogeneous subgroups. The PT total score becomes the conditioning score (effectively as a surrogate of θ) to calculate required conditional covariances for statistical hypothesis testing using the AT1 and AT2 item responses. The procedure as formalized by Nandakumar and Stout (1993) is referred to as DIMTEST.

Gibbons et al. (2007) studied the consequences of fitting unidimensional models to multidimensional data empirically. The question they asked was slightly different. To the extent that the primary dimension of interest can be preserved in a unidimensional model and in the primary factor of a bifactor model or possibly in an exploratory item factor analysis model, does the specific model used make a difference in the results?

They conducted a simulation study to investigate the effects of applying Samejima's (1969) graded response model in unidimensional and bifactor form to multidimensional data. Conditions studied were (a) test length, 50 items or 100 items; (b) number of dimensions, 5 or 10; (c) primary loadings, 0.50 or 0.75; and (d) domain loadings, 0.25 or 0.50. Outcome results include standard deviation of expected *a posteriori* (EAP) estimates



NI	50	50	100	100	50	50	100	100	50	50	100	100
ND	5	10	5	10	5	10	5	10	5	10	5	10
PL	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.5	0.5	0.5	0.5
DL	0.5	0.5	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

FIGURE 3.1
Mean standard deviations of θ of the unidimensional and bifactor models based on 1000 replications per condition (number of items [NI] = 50 or 100, number of dimensions [ND] = 5 or 10, primary loadings [PL] = 0.50 or 0.75, domain loadings [DL] = 0.25 or 0.50).

of θ, posterior standard deviations (PSDs, or standard errors) of Bayes EAP scores, log-likelihood (model fit), differences between EAP and actual θ, and percentage change between unidimensional and bifactor models of these variables. The generated data were based on a four-point categorical scale, and the examinee distribution was assumed to be normal, N(0, 1), based on 1000 replications. In the following, we summarize the key findings of this study.

Figure 3.1 reports the standard deviations of the θ estimates for the unidimensional and bifactor models across the 12 simulated conditions. Inspection of the figure indicates that the EAP estimates based on the unidimensional model were more varied across all conditions. The magnitude of the difference decreased when the primary and secondary loadings decreased, leading to a more unidimensional solution. As shown, as the number of items increased from 50 to 100, the EAP estimates from both models became more varied, but not as severe for the bifactor model.

Figure 3.2 reports the mean PSD of the Bayes EAP estimates. As shown, the differences in the PSD between the models can be dissected in terms of the dimensionality of the underlying data. Specifically, in the conditions in which the primary loadings are 0.75 and the domain loadings are 0.50, the PSD of the unidimensional model substantially underestimates the PSDs from the bifactor model. As shown, the largest PSD for the unidimensional model occurs with 100 items and 5 dimensions. The PSD estimated by the bifactor model remains fairly consistent across the conditions in which the underlying structure can be regarded as strongly multidimensional (i.e., primary loadings = 0.75, domain

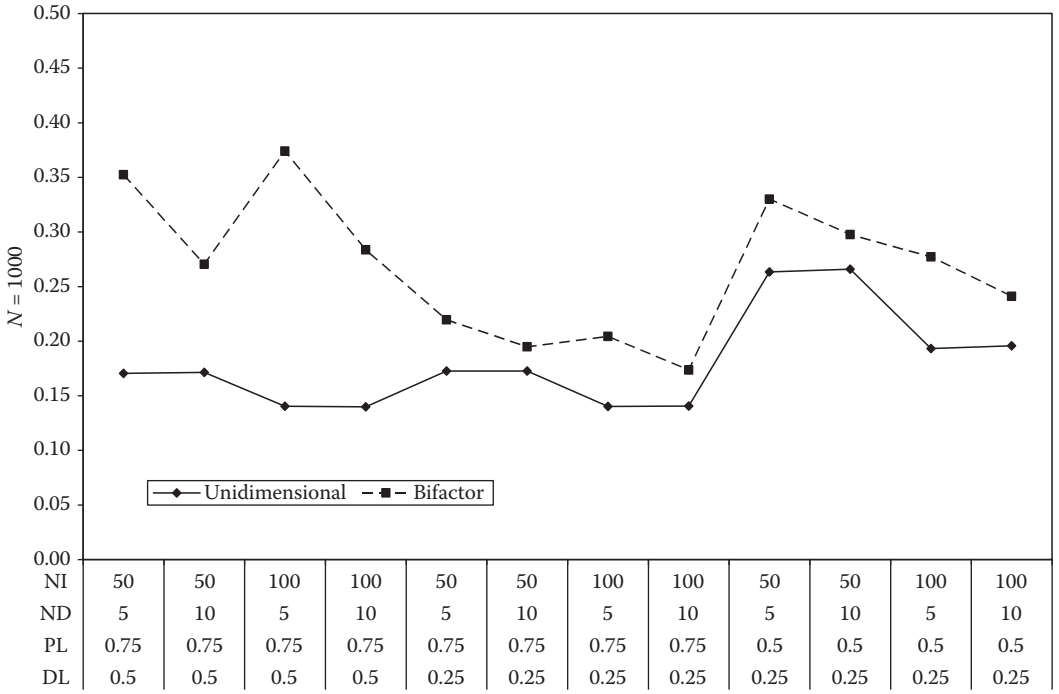


FIGURE 3.2
Mean PSDs of Bayes expected a posterior scores of the unidimensional and bifactor models based on 1000 replications per condition (number of items [NI] = 50 or 100, number of dimensions [ND] = 5 or 10, primary loadings [PL] = 0.50 or 0.75, domain loadings [DL] = 0.25 or 0.50).

loadings = 0.50). For the conditions in which the primary loadings are 0.75 and the domain loadings are 0.25, the PSD for the unidimensional approaches that for the bifactor model but, nevertheless, continues to underestimate the bifactor result, which is the correct value in this case. The largest discrepancies between the PSD of these models occur when the number of dimensions is 5 and the number of items is 50 and 100. The smallest difference between the mean PSDs for the unidimensional and bifactor models occurs when the number of dimensions is 10 with 50 items. For the bifactor model, the PSD decreases slightly when the number of items increases from 50 to 100. However, the number of dimensions does not seem to significantly influence the PSD of the bifactor model.

The results of this study illustrate the consequences attached to applying a unidimensional IRT model to data with varying degrees of multidimensionality compared to the bifactor model. The first set of results addressed the variability in estimated θ values, or examinees' standing on the latent trait. Compared to the unidimensional model, the bifactor model yielded θ estimates that were more homogeneous across simulated data structures. As a consequence, studies that are designed to evaluate educational or clinical interventions will have increased statistical power to detect meaningful effects when scores are based on a bifactor model and the underlying data are the result of a multidimensional response process.

PSD estimates were found to be underestimated across all conditions for the unidimensional model. For the bifactor model, PSD values were consistently below 0.20 across conditions, except when the total test length was 50 and the primary loadings were 0.50

and the domain loadings were 0.25. One setting in which the underestimation of PSDs could affect test scores is in computer adaptive testing, in which each item is intentionally selected to provide the most information for estimating examinee ability in the sense of greatest reduction of PSD. Using PSD estimates based on the unidimensional model may therefore lead to suboptimal estimates of examinee ability. Used as measurement error variance, the inverse squared unidimensional PSDs are not valid for weighting observations in statistical analyses using the scores as data.

3.8 Limited-Information Goodness-of-Fit Tests

The nonparametric indices based on conditional covariances such as DIMTEST do not explicitly specify a distribution of the θs. Hence, they require the use of external conditioning subscores such as the partitioning total score. When an item factor analysis model is fitted using standard estimation methods such as maximum marginal likelihood, population distributions of θ are routinely assumed. Therefore, upon finding the maximum likelihood solution, the model yields expected probabilities for each single item, as well as joint probabilities for item pairs, triplets, quadruplets, etc. When contrasted against the observed probabilities, the residuals may be used to derive goodness-of-fit statistics. Most of the time, univariate and bivariate association information is used.

In the context of IRT, statistics based on (mostly) univariate and bivariate subtables are referred to as limited-information goodness-of-fit statistics, in contrast to full-information statistics (e.g., the Pearson's chi-square statistic) that are based on residuals of the full item by item-by-item cross-classifications. Despite the apparent loss of information due to collapsing the full contingency table into a series of first- and second-order association tables, limited-information test statistics have been suggested as a potential solution to the Achilles' heel of full-information statistics, namely, the sparseness of the underlying multiway contingency table upon which the IRT model is defined (Bartholomew & Tzamourani, 1999). The number of cells in the table is exponentially increasing in the number of items, and for tests of realistic length, the table will become extremely sparse for any conceivable sample size. The sparseness invalidates the usual asymptotic chi-square approximations to the distribution of Pearson's statistic or the likelihood ratio statistic, making model fit testing decisions based on full-information statistics untrustworthy in practical situations. On the other hand, test statistics based on univariate and bivariate subtables maintain Type I error rate control and have adequate power (see, e.g., Cai et al., 2006). In particular, Maydeu-Olivares and Joe's (2005) M_2 family of test statistics has witnessed increasing popularity.

In the context of multidimensional IRT, Cai and Hansen (2012) extended the dimension reduction technique, already used in parameter estimation of bifactor models, to limited-information goodness-of-fit testing. For example, for a bifactor model, the probabilities and derivatives for computing limited-information test statistics require at most two-dimensional numerical integration, regardless of the number of factors in the model, making it feasible to test much larger models with many latent variables. In addition, Cai and Hansen (2012) developed a new quadratic form test statistic, which they call M_2^*, that is based on the general limited-information testing principles proposed by Joe and Maydeu-Olivares (2010). The statistic is best understood as a further reduction (or concentration) of the univariate and bivariate subtables. When the item responses are polytomous, this new

statistic can be substantially better calibrated and more powerful than M_2. In addition, the chi-square distributed test statistics can be used to calculate fit measures such as the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993) that are free from the influence of sample size.

The details of limited-information goodness-of-fit testing are more substantial than those can be covered in this chapter. In brief, the development begins with the realization that the IRT model can be written as a function of the (marginal) response pattern probability $\pi_{\bigcup}(\boldsymbol{\gamma})$ for pattern $\bigcup$, where $\boldsymbol{\gamma}$ is a notational shorthand for the collection of free and estimable parameters in the model. Suppose there are C possible response patterns. Let us define the $C \times 1$ vector of modeled probabilities as $\boldsymbol{\pi}(\boldsymbol{\gamma})$ and the corresponding $C \times 1$ vector of observed proportions as $\boldsymbol{p}$. Let the $C \times 1$ population cell probabilities be $\boldsymbol{\pi}$. The null hypothesis being evaluated in the goodness-of-fit testing situation is $H_0{:}\boldsymbol{\pi}(\boldsymbol{\gamma}) = \boldsymbol{\pi}$, for some $\boldsymbol{\gamma}$, versus the alternative $H_A{:}\boldsymbol{\pi}(\boldsymbol{\gamma}) \neq \boldsymbol{\pi}$, for any $\boldsymbol{\gamma}$.

Suppose the total sample size is N. Treating $\boldsymbol{p}$ as the fixed observed data, maximizing (e.g., using the EM algorithm) the multinomial likelihood with cell probabilities given by $\boldsymbol{\pi}(\boldsymbol{\gamma})$ leads to the maximum marginal likelihood estimator $\hat{\boldsymbol{\gamma}}$. Let the fitted cell probabilities be $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\boldsymbol{\gamma})$. The cell residuals are $\boldsymbol{e} = \boldsymbol{p} - \hat{\boldsymbol{\pi}}$. Standard discrete multivariate analysis results (Rao, 1973) suggest that the cell residuals are asymptotically C-variate normal under the null hypothesis:

$$\sqrt{N}\boldsymbol{e} = \sqrt{N}(\boldsymbol{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Xi}),$$

where $\boldsymbol{\Xi} = \boldsymbol{D} - \boldsymbol{\pi}\boldsymbol{\pi}' - \boldsymbol{\Delta}(\boldsymbol{\gamma})\boldsymbol{\mathcal{F}}(\boldsymbol{\gamma})^{-1}\boldsymbol{\Delta}(\boldsymbol{\gamma})'$, $\mathbf{D} = \mathrm{diag}(\boldsymbol{\pi})$,

$$\boldsymbol{\Delta}(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}$$

is the Jacobian of the model, and $\mathcal{F} = \boldsymbol{\Delta}(\boldsymbol{\gamma})'\boldsymbol{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\gamma})$ is the Fisher information matrix.

Subtable probabilities such as the univariate and bivariate probabilities are linear functions of the cell probabilities (Cai et al., 2006). The relationship can be conveniently expressed using reduction operator matrices (Joe & Maydeu-Olivares, 2010). Let $\boldsymbol{T}$ be a particular fixed $q \times C$ matrix with full row rank that achieves the reduction of $\boldsymbol{\pi}$ into lower-order probabilities. The new vector of residuals retains asymptotic normality

$$\sqrt{N}\boldsymbol{r} = \sqrt{N}\boldsymbol{T}\boldsymbol{e} = \sqrt{N}\boldsymbol{T}(\boldsymbol{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Xi}\boldsymbol{T}' = \bar{\boldsymbol{D}} - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}' - \bar{\boldsymbol{\Delta}}(\boldsymbol{\gamma})\mathcal{F}(\boldsymbol{\gamma})^{-1}\bar{\boldsymbol{\Delta}}(\boldsymbol{\gamma})'$, and $\bar{\boldsymbol{D}} = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}'$, $\bar{\boldsymbol{\pi}} = \boldsymbol{T}\boldsymbol{\pi}$, with the $q \times \dim(\boldsymbol{\gamma})$ (local) Jacobian matrix given by

$$\bar{\boldsymbol{\Delta}}(\boldsymbol{\gamma}) = \boldsymbol{T}\boldsymbol{\Delta}(\boldsymbol{\gamma}) = \boldsymbol{T}\frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} = \frac{\partial \bar{\boldsymbol{\pi}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}.$$

If the IRT model is locally identified, that is, $\bar{\boldsymbol{\Delta}}(\boldsymbol{\gamma})$ has full column rank, then there exists a $q \times [q - \dim(\boldsymbol{\gamma})]$ orthogonal complement matrix $\bar{\boldsymbol{\Delta}}_c$ such that $\bar{\boldsymbol{\Delta}}_c'\bar{\boldsymbol{\Delta}}(\boldsymbol{\gamma})$ is a null matrix. This implies

$$\sqrt{N}\bar{\boldsymbol{\Delta}}_c'\boldsymbol{r} \xrightarrow{D} \mathcal{N}_{q-\dim(\boldsymbol{\gamma})}(\mathbf{0}, \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega} = \bar{\boldsymbol{\Delta}}_c'(\bar{\boldsymbol{D}} - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}')\bar{\boldsymbol{\Delta}}_c'$. Evaluating the model-implied probabilities and the Jacobian elements at the maximum likelihood estimate, the following limited-information statistic is asymptotically centrally chi-square distributed with $q - \dim(\boldsymbol{\gamma})$ degrees of freedom:

$$M = N\boldsymbol{r}'\bar{\boldsymbol{\Delta}}_c\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{\Delta}}_c'\boldsymbol{r}.$$

3.9 Example

As an illustration, we analyze data obtained with the *Quality of Life Interview for the Chronically Mentally Ill* (Lehman, 1988) from 586 chronically mentally ill patients. The instrument consists of one global life-satisfaction item followed by 34 items in seven subdomains, namely, Family, Finance, Health, Leisure, Living, Safety, and Social, with four, four, six, six, five, five, and four items, respectively. Respondents were instructed to rate each item in turn on a seven-point scale consisting of ordered response categories: *terrible, unhappy, mostly dissatisfied, mixed, about equally satisfied and dissatisfied, mostly satisfied, pleased, delighted*. Both the multiple content areas of the subdomains and their labeling as such encourage responses to the set of items as a whole rather than considered responses to the individual items. This effect creates dependencies between responses within the sets that violate the assumption of conditional independence of response required for conventional one-dimensional IRT analysis.

3.9.1 Exploratory Item Factor Analysis

Given that the items are clustered within seven content domains, for the purpose of dimensionality assessment, we considered models containing one through eight factors, to determine if an additional factor explained any significant additional variation in item responses over the seven specified subdomains. Chi-square statistics for the addition of each successively added factor are shown in Table 3.1. Very roughly, a chi-square value is significant if it is at least twice as large as its degrees of freedom. By this rule, even the addition of an eighth orthogonal factor shows no danger of over-factoring, although

TABLE 3.1

Quality of Life Data ($N = 586$)

Decrease of −2 log L of Solutions with 1–8 Factors			
Solution	**−2 log L**	**Decrease**	**DF**
1	66837.1		
2	66045.0	792.1	34
3	65089.5	955.5	33
4	64118.4	971.5	32
5	63509.1	609.3	31
6	63063.7	445.4	30
7	62677.5	386.2	29
8	62370.5	307.4	28

its contribution to improved goodness of fit is the smallest of any factor. Notice that the decreases are not monotonic; unlike traditional factor analysis of product-moment correlations, the marginal probabilities of the response patterns (which determined the marginal likelihood) reflect changes in all parameters jointly, including the category parameters and not just the factor loadings. Because our inspection of the signs of the loadings of the first seven factors showed relationships to the item groups and the eighth factor did not (see Table 3.2), the seven-factor model is likely the most parsimonious choice.

As expected, all first principal factor loadings are positive, clearly identifying the factor with the overall Quality-of-Life variable (see Table 3.2). In fact, the largest loading is that of item number one, which asks for the respondent's rating of overall quality of life. Only one item, the last, has a loading less than 0.5. As for the six bipolar factors, the significant feature is that the sign patterns of loadings of appreciable size conform to the item groups. Factor 2 strongly contrasts the Finance group with Family Living and Safety; to as lesser extent, Leisure is working in the same direction as Finance. In other words, persons who tend to report better financial positions and quality of leisure are distinguished by this factor from those who report better family relationships and safety. Factor 3 then combines Living and Finance and contrasts them primarily with a combination of Health and Safety. Factor 4 contrasts a combination of Family and Social with Finance, Health, and Safety. Factor 5 combines Social, Living, and Safety versus Family, Finance, and Safety. Factor 6 primarily contrasts Health with Safety. Finally, Factor 7 contrasts Social verses Leisure. The fact that the seven-factor solution has the expected all positive first factor and patterns for the remaining bipolar factors that contrast item groups rather than items within groups clearly supports a bifactor model for these data. Estimated thresholds for this analysis are reported by Bock and Gibbons (2010).

3.9.2 Confirmatory Item Bifactor Analysis

The bifactor model produced a value of $-2\log \mathrm{L} = 64233.3$, which is similar to that obtained for a four-factor model. While the seven-factor unrestricted model provides significant improvement in fit, inspection of the estimated factor loading in Table 3.3 shows that the bifactor model provides the most parsimonious and easily interpretable results. Because the group factor loadings are not constrained to orthogonality with those of the general factor, they are all positive and their magnitudes indicate the strength of the effect of items belonging to common domains. The effects of Family and Finance, for example, are stronger than those of Health and Leisure. It is interesting to note that the empirical reliability for the primary dimension for the bifactor model is 0.9, but is overestimated as 0.95 for a unidimensional model applied to these same data (standard errors of 0.322 and 0.232, respectively). As expected, reliability is overestimated and uncertainty in estimated scale scores is underestimated when the conditional dependencies are ignored. Avoiding this type of bias is a major motivation for item bifactor analysis.

Limited-information goodness-of-fit testing lends additional support for the appropriateness of the bifactor solution. Take the unidimensional model for instance; the Cai-Hansen modified M_2^* statistic is equal to 2674.85 on 385° of freedom, $p < 0.0001$. The statistic uses both univariate and bivariate residual tables. Since there are 35 items, there are $35 + 35 \times (35 - 1)/2 = 630$ residuals available for testing model fit. The unidimensional graded model contains $35 \times 7 = 245$ free item parameters, resulting in $630 - 245 = 385°$ of freedom. The null hypothesis of exact unidimensionality is rejected and the unidimensional model is untenable for this dataset. We may compute the RMSEA index, which is a widely used measure of fit in factor analysis and structural equation modeling, and it

TABLE 3.2

Item Principal Factor Loadings

		Factors						
Item Group	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1	0.769	0.021	0.082	−0.054	0.054	−0.002	0.097
Family								
1	2	0.614	−0.269	0.044	−0.461	0.272	−0.081	0.044
	3	0.687	−0.181	−0.007	−0.380	0.159	0.007	−0.058
	4	0.703	−0.245	−0.045	−0.522	0.257	0.029	0.008
	5	0.729	−0.214	−0.004	−0.505	0.279	0.046	0.019
Finance								
2	6	0.606	0.468	−0.391	0.116	0.284	0.003	−0.101
	7	0.515	0.405	−0.300	0.097	0.220	0.021	−0.032
	8	0.647	0.511	−0.342	0.101	0.276	0.048	−0.092
	9	0.632	0.510	−0.305	0.072	0.242	0.023	−0.069
Health								
3	10	0.568	−0.123	0.132	0.201	0.049	−0.236	0.095
	11	0.644	0.007	0.195	0.128	0.038	−0.443	−0.139
	12	0.627	0.087	0.289	0.074	−0.026	−0.390	−0.167
	13	0.668	−0.052	0.156	0.138	−0.005	−0.383	−0.232
	14	0.678	−0.004	0.154	0.116	0.061	−0.288	0.054
	15	0.701	0.044	0.249	0.045	0.071	−0.154	0.054
Leisure								
4	16	0.741	0.215	0.150	0.030	−0.138	0.156	0.155
	17	0.657	0.149	0.142	−0.017	−0.128	−0.054	0.285
	18	0.721	0.223	0.101	−0.005	−0.173	0.019	0.331
	19	0.749	0.313	0.144	−0.059	−0.199	0.095	0.301
	20	0.670	0.192	0.078	−0.101	−0.162	−0.030	0.295
	21	0.522	−0.002	−0.056	−0.002	−0.099	−0.049	0.042
Living								
5	22	0.664	−0.241	−0.401	0.038	−0.191	0.048	−0.008
	23	0.549	−0.332	−0.325	0.118	−0.140	−0.013	−0.028
	24	0.611	−0.253	−0.529	0.006	−0.190	−0.112	0.042
	25	0.626	−0.347	−0.446	0.079	−0.285	−0.127	0.030
	26	0.568	−0.213	−0.439	0.066	−0.177	−0.018	0.034
Safety								
6	27	0.679	−0.241	0.221	0.299	0.232	0.341	−0.004
	28	0.688	−0.387	0.051	0.317	0.141	0.250	−0.040
	29	0.594	−0.065	0.231	0.145	0.109	0.123	−0.044
	30	0.670	−0.253	0.181	0.276	0.196	0.223	−0.003
	31	0.702	−0.264	0.140	0.336	0.064	0.197	0.006
Social								
7	32	0.688	0.189	0.169	−0.180	−0.399	0.197	−0.375
	33	0.696	0.254	0.099	−0.192	−0.317	0.212	−0.218
	34	0.620	0.203	0.149	−0.118	−0.218	0.161	−0.232
	35	0.494	−0.163	0.122	−0.056	−0.179	0.046	−0.202

TABLE 3.3

Item Bifactor Loadings

		Factors						
Item Group	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1	0.789						
Family								
1	2	0.535	0.620					
	3	0.576	0.509					
	4	0.575	0.586					
	5	0.631	0.547					
Finance								
	6	0.476		0.634				
2	7	0.437		0.553				
	8	0.544		0.617				
	9	0.535		0.622				
Health								
3	10	0.560		0.256				
	11	0.528		0.504				
	12	0.486		0.505				
	13	0.529		0.473				
	14	0.650		0.286				
	15	0.714		0.141				
Leisure								
4	16	0.694			0.285			
	17	0.565			0.413			
	18	0.628			0.451			
	19	0.635			0.506			
	20	0.571			0.473			
	21	0.479			0.208			
Living								
5	22	0.536				0.549		
	23	0.484				0.530		
	24	0.497				0.668		
	25	0.508				0.688		
	26	0.508				0.672		
Safety								
6	27	0.557					0.517	
	28	0.593					0.474	
	29	0.533					0.501	
	30	0.558					0.538	
	31	0.591					0.383	
Social								
7	32	0.545						0.438
	33	0.586						0.351
	34	0.520						0.466
	35	0.446						0.296

is equal to 0.08 and a 90% confidence interval of RMSEA is (0.076,0.084). Adopting established conventions in factor analysis (Browne & Cudeck, 1993), an RMSEA that exceeds 0.05 cannot be taken as an indication of good fit. On the other hand, the M_2^* statistic is equal to 546.88 on 351° of freedom. While it remains significant at the 0.05 level, the RMSEA index for the bifactor model is equal to 0.03 and the 90% confidence interval is (0, 0.035), indicating substantially improved fit.

3.10 Discussion

We have shown that for many applications of IRT, multidimensionality rather than unidimensionality should represent the null hypothesis. There are a variety of limited-information and full-information methods for determining the goodness of fit and the underlying dimensionality of a particular test. As it turns out, the bifactor model produces excellent results for a variety of different IRT applications because it (a) uses expert judgement to define the underlying factor structure and (b) evaluates the likelihood in an always computationally tractable way because it reduces it to a two-dimensional integral, which is relatively easy to evaluate in practice. Traditional methods based on eigenvalues have a tendency to identify factors that are not indicative of the underlying trait of interest. By contrast, goodness-of-fit statistics which compare various nested and nonnested statistical models can be used efficiently to evaluate the dimensionality of a particular test. In general, unrestricted item factor analysis should not be used to evaluate multidimensionality. It is poorly specified and is subject to considerable rotational variance leading to a plethora of different conclusions regarding the latest variables. This is not the case for the bifactor model which provides essentially the same answer regardless of small changes in the model specification. Finally, one should be extremely cautious regarding the fitting of a unidimensional model to what are inherently multidimensional data. The net result is an underestimate of the point at which adaptive testing should terminate (i.e., underestimates of the posterior variance of the latent variable estimate), and increases in the empirical variance of the resulting test score. Neither of these conditions is good. As a consequence, the best possible approaches to determining dimensionality should always be used.

References

Andrich, D. 1978. A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

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

Bliss, C. I. with an Appendix by Fisher, R. A. 1935. The calculation of the dosage mortality curve. *Annals of Applied Biology*, *22*, 134.

Bock, R. D. & Gibbons, R. D. 2010. Factor analysis of categorical item responses. In M. Nering and R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models: Development and Applications*. Florence, KY: Lawrence Erlbaum.

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

Bock, R. D. & Jones, L. V. 1968. *The Measurement and Prediction of Judgment and Choice*. San Francisco, CA: Holden-Day.
Bock, R. D. & Lieberman, M. 1970. Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.
Browne, M. W. 2001. An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.
Browne, M. W. & Cudeck, R. 1993. Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136–162). Beverly Hills, CA: Sage.
Cai, L. 2010a. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.
Cai, L. 2010b. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.
Cai, L. & Hansen, M. 2012. Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245–276.
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.
Cai, L., Yang, J. S., & Hansen, M. 2011. Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.
Divgi, D. R. 1979. Calculation of the tetrachoric correlation coefficent. *Psychometrika*, *44*, 169–172.
Fechner, G. T. 1860. *Elemente der Psychophysik*, Volume 1. Leipzig: Breitkopf and Hartel.
Finney, D. J. 1952. *Probit Analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.
Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D., Bhaumik, D. K., Kupfer, D., Frank, E., Grochocinski, V., & Stover, A. 2007. Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.
Gibbons, R. D., Bock, R. D., & Immekus, J. 2007. *The Added Value of Multidimensional IRT Models*. Final Report Contract 2005-05828-00-00, National Cancer Institute. Available at www.healthstats.org.
Gibbons, R. D. & Hedeker, D. 1992. Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.
Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. 2008. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*, 361–368.
Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. K. 2012. The CAT-DI: A computerized adaptive test for depression. *Archives of General Psychiatry*, *69*, 1104–1112.
Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. 2014. Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry*, *171*, 187–194.
Harman, H. H. 1967. *Modern Factor Analysis* (2nd ed.). Chicago, IL: The University of Chicago Press.
Holzinger, K. J. & Swineford, F. 1937. The bi-factor method. *Psychometrika*, *2*, 41–54.
Joe, H. & Maydeu-Olivares, A. 2010. A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393–419.
Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.
Jöreskog, K. G. 2002. Structural Equation Modeling with Ordinal Variables Using LISREL. http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf.
Lehman, A. F. 1988. A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning*, *11*, 51–62.
Lord, F. M. & Novick, M. R. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

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.

Muthén, B. O. 1989. Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Nandakumar, R. & Stout, W. F. 1993. Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications* (2nd ed.). New York, NY: Wiley.

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

Song, X. Y. & Lee, S. Y. 2003. Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivariate Behavioral Research*, *38*, 57–79.

Stout, W. 1987. A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. 1990. A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293–326.

Thurstone, L. L. 1947. *Multiple-Factor Analysis*. Chicago, IL: University of Chicago Press.

Tucker, L. R. 1958. An inter-battery method of factor analysis. *Psychometrika*, *23*, 111–136.

Zhang, J. & Stout, W. 1999. The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–214.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com