LOGISTIC APPROXIMATIONS OF MARGINAL TRACE LINES FOR BIFACTOR ITEM RESPONSE THEORY MODELS

Brian Dale Stucky

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2011

Approved by:
David Thissen, Ph.D.
Patrick J. Curran, Ph.D.
Robert C. MacCallum, Ph.D.
A.T. Panter, Ph.D.
Eric Youngstrom, Ph.D.

ABSTRACT

BRIAN STUCKY: Logistic Approximations of Marginal Trace Lines for Bifactor Item Response Theory Models
(Under the direction of David Thissen, Ph.D.)

Bifactor item response theory models are useful when item responses are best represented by a general, or primary, dimension and one or more secondary dimensions that account for relationships among subsets of items. Understanding slope parameter estimates in multidimensional item response theory models is often challenging because interpretation of a given slope parameter must be made conditional on the item's other parameters. The present work provides a method of computing marginal trace lines for an item loading on more than one dimension. The marginal trace line provides the relationship between the item response and the primary dimension, after accounting for all other dimensions. Findings suggest that a logistic function, common in many applications of item response theory, closely approximates the marginal trace line in a variety of model related conditions. Additionally, a method of IRT-based scoring is proposed that uses the logistic approximation marginal trace lines in a unidimensional fashion to compute scaled scores and standard deviation estimates for the primary dimension.

The utility of the logistic approximation for marginal trace lines is considered across a wide range of varying bifactor parameter estimates, and under each condition the marginal is closely approximated by a logistic function. In addition, it is shown that use of the logistic approximations to conduct item response theory-based scoring should be restricted to selecting a single item from each secondary dimension in order to control for local

dependence. Under this restriction, scaled scores and posterior standard deviations are nearly equivalent to other MIRT-based scoring procedures. Finally, a real-data application is provided which illustrates the utility of logistic approximations of marginal trace lines in item selection and scale development scenarios.

ACKNOWLEDGEMENTS

I am indebted to my adviser Dr. David Thissen, and the faculty and students of the L.L. Thurstone Psychometric Laboratory. I also wish to thank Randy Schuler.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

AN OVERVIEW OF MULTIDIMENSIONAL ITEM RESPONSE THEORY

Item response theory (IRT) is a useful technique for item analysis and scoring which is becoming increasingly common in educational measurement, health outcomes research, and psychology. IRT models propose that the probability of response to an item is a function of the characteristics of the item (i.e., item parameters) and the individual's location on the latent trait(s) (i.e., person parameters). This item response function, or trace line, conveys all information available from the item that can be used to estimate an individual's latent trait. When used in combination with multiple items, the trace lines form the likelihood, from which one can determine the location on the latent variable where the trait level is most likely.

IRT score estimates may be computed for either unidimensional (UIRT) or multidimensional (MIRT) models. UIRT scores are appropriate when the relationships among the items, given an individual's trait level, can be accounted for by a single latent variable. When no additional latent variables are needed to account for response covariation beyond the single dimension, the item set satisfies the assumptions of unidimensionality and local independence. However, if fitting the item response data requires multiple latent variables, then MIRT models are needed to achieve local independence.

Often there are situations in which unidimensional scores are desired, but fitting a UIRT model suggests local dependence (LD) between small numbers of items. LD refers to relationships among item responses that are not accounted for by a single dimension. In

these situations it is often useful to account for LD among a subset of items by estimating an additional latent variable (e.g., in a bifactor model). As a special class of MIRT models, bifactor models account for the shared relations among all the items through a general, or primary, dimension and one or more secondary dimensions, orthogonal to the primary dimension, which contain loadings only for those locally dependent items.

Traditionally, bifactor models have been employed only to identify LD (i.e., multidimensionality). In order to provide unidimensional scores for such models, the most common approach has been to set items aside from secondary dimensions, eliminating the dependence, and then to use the remaining items to compute scores with a unidimensional model. The present research aims to develop a method in which violations of unidimensionality can be accounted for in a bifactor model while still producing unidimensional scores. In other words, the model is allowed the flexibility to account for multiple dimensions, while the scores reflect the individual's location on the general factor.

The Effects of Ignoring Local Dependence

When a set of items is best represented by a single dimension it is referred to as unidimensional. Unidimensionality implies local independence, which indicates that all the relationships among the data are accounted for by the underlying latent variable. Consider a pair of items *i* and *j* with trace lines T_i and T_j which "trace" the probability of response given the latent variable (θ). If the response model is defined by a single dimension, then the probability of an individual correctly responding to both items is equal to the product of the individual trace lines given the latent variable:

$$T(u_i = 1, u_j = 1 \mid \theta) = T(u_i = 1 \mid \theta) T(u_j = 1 \mid \theta). \quad (1)$$

In other words, if local independence holds, then the joint likelihood of a particular response pattern is properly represented by the product of the separate probabilities of item responses. Of course, this should also hold for all the items in a test conditional on θ.

In the 1980's, prior to usable implementations of MIRT models, researchers struggled to determine how robust IRT models were to violations of unidimensionality. The majority of this work involved generating multidimensional data from simple structure factor analysis models with varying degrees of correlation between factors, and then comparing parameter and individual trait estimates after fitting UIRT models. To briefly summarize, numerous authors suggest that when separate dimensions are correlated greater than about r = .60, a single factor may adequately represent the factor structure (Folk & Green, 1989; Drasgow & Parsons, 1983; Ackerman, 1989; Harrison, 1986; Reckase, 1979). Additionally, trait recovery is improved when the general factor is strongly unidimensional and contains a large number of items with a high degree of information (Harrison, 1986).

Though prior research attempted to validate the use of fitting UIRT models to multidimensional data, the costs can be great, including "θ-theft" (i.e., when a small number of locally dependent items define the dimension; Thissen & Steinberg, 2010, p. 131), over-estimating score reliability (Thissen, Steinberg, & Mooney, 1989), and in misrepresenting the data. In practice θ-estimates often give the appearance of robustness to violations of unidimensionality, or as Demars (2006) says, "if the focus is on estimated theta and not the item parameters, any of the models will perform satisfactorily…" (p. 165). Importantly, it is the factor structure, or item parameter interpretations, that are most often misrepresented in unidimensional representations of multidimensional data. So, while differences in score

precision may be appear slight, interpretation of the latent trait based on the dimensions' parameters is often what is most affected.

Much of the past research on the robustness of UIRT models to violations of local independence was conducted prior to the availability of usable implementations of MIRT models. Hence, previous research was motivated in large part by a desire to fit UIRT models, because MIRT models were not a viable alternative. Past investigations, though important in understanding under what conditions essential unidimensionality may be sufficient, are of less relevance now that well established procedures are in place for fitting MIRT models (Reckase, 2009).

Compensatory MIRT Models

When fitting item responses requires more than one latent trait, MIRT models may be appropriate (McKinley & Reckase, 1983; Reckase, 1985). The most widely used are compensatory[1] MIRT models, which model the probability of a response with a linear combination of latent variables (θ-coordinates). In other words, if an individual's location is low on a particular trait, the linear combination may compensate with a high score on another trait.

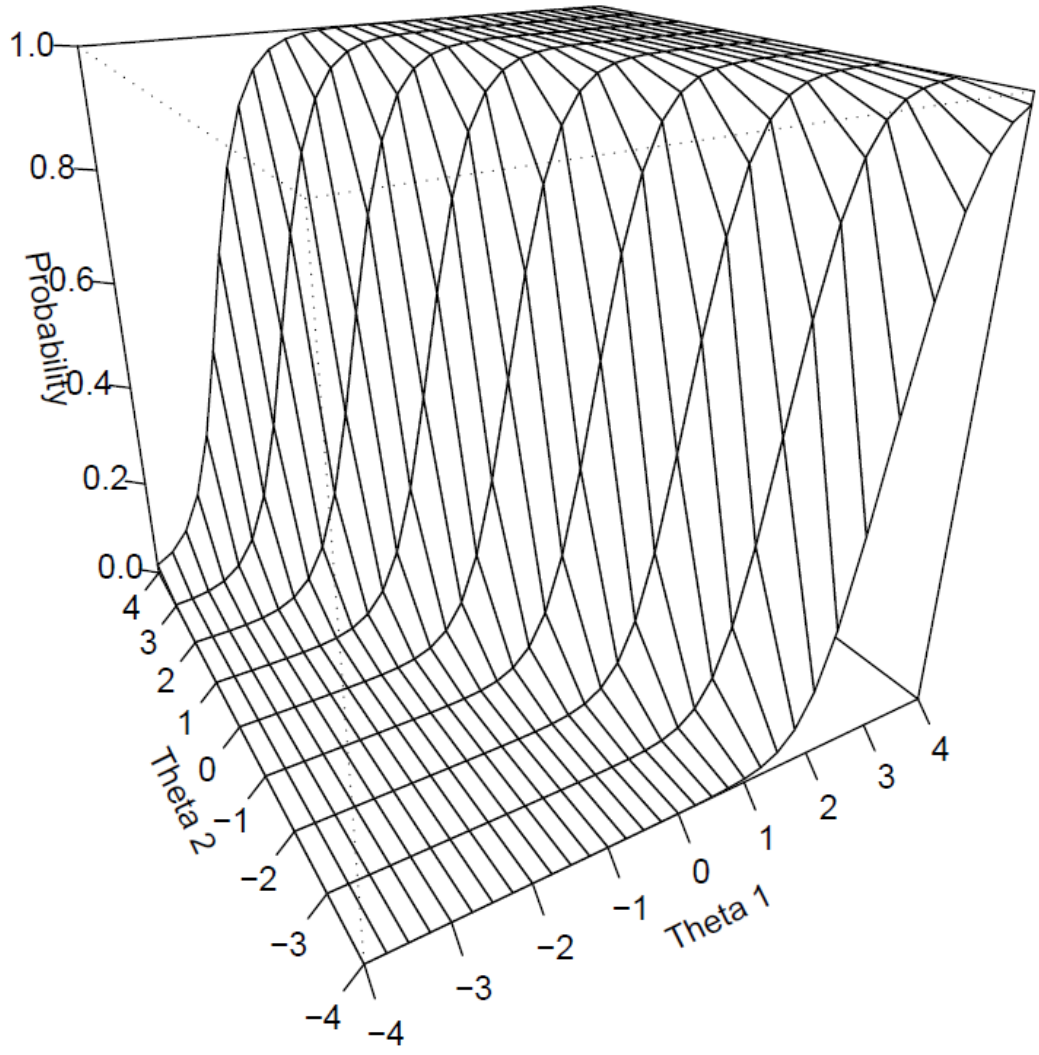
For simplicity, consider the 2PL compensatory MIRT model as an extension of the 2PL univariate IRT model (for multivariate extensions of Samejima's (1969) graded response model see Muraki and Carlson (1993)). The probability of a person with trait vector $\boldsymbol{\theta}$ responding correctly to item i is based on a vector of discrimination parameters $\mathbf{a}_i$ and an intercept c_i:

[1] Other, less often used MIRT models are *noncompensatory* (Sympson, 1978). This class of models can be considered the combination of separate unidimensional models. Here the probability of correct response for an item is often formed from the product of the separate probabilities for the latent traits. The models are said to be noncompensatory because the probability of correct response cannot be higher than any of the probabilities in the product.

$$T_i(u_{ij}=1\,|\,\boldsymbol{\theta}_j,\mathbf{a}_i,c_i)=\frac{e^{\mathbf{a}_i\boldsymbol{\theta}+c_i}}{1+e^{\mathbf{a}_i\boldsymbol{\theta}+c_i}}. \quad (2)$$

Unlike the difficulty parameter in the UIRT 2PL model, c_i represents the relative difficulty of an item without respect to a trait dimension. Because more than one dimension affects responses, graphical representations (trace surfaces) are often used to depict the relation between item responses along 2-dimensions. Figure 1 illustrates the trace surface for an item with $a_1 = 3$, $a_2 = 2$, $c = 0$.

Figure 1. Trace surface for an item more discriminating on the primary dimension



That the probability of a correct response increases more rapidly along θ_1 indicates that the first dimension has a greater effect on responses for this item. The compensatory nature of the model is also evident. The exponent in eq. 2 is a linear combination of **θ** and **a**-vectors with an intercept *c*. If the exponent is equal to 0, then eq. 2 simplifies to $T_i = ½$ because $e^0 = 1$. Rearranging the terms in the exponent then provides the line through the **θ**-space where the probability of correct response is 0.5:

$$\theta_2 = -\frac{1}{a_2}(a_1\theta_1 - c)\,. \tag{3}$$

Hence, for the present example, relatively low trait levels on θ_1 (say $\theta_1 = -2$) can be compensated for by high levels of θ_2 (i.e., $\theta_2 = 3$). Note that because this particular item better discriminates on the θ_1 dimension, higher levels of θ_2 are required to compensate for low levels of trait θ_1.

MIRT Scoring

The relative utility of MIRT models may be judged by the scores they produce. In general, MIRT scores may be thought of as a multivariate extension of UIRT scoring. The likelihood of a particular response pattern is computed by the following:

$$L(\mathbf{U} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} T_{u_i}(\theta), \tag{4}$$

where $L(\mathbf{U}|\boldsymbol{\theta})$ is the likelihood of a response pattern to a *n* item test for an individual with response pattern $\mathbf{U} = \{u_1, u_2, \ldots, u_n\}$ (Segall, 1996). For certain extreme response patterns, all correct (or positive) or incorrect (or negative) responses to test items (which are common in short tests), the mean of the likelihood becomes undefined, the mode is infinite, and some heuristic is needed to compute scores. For this and other reasons, it is useful to employ a

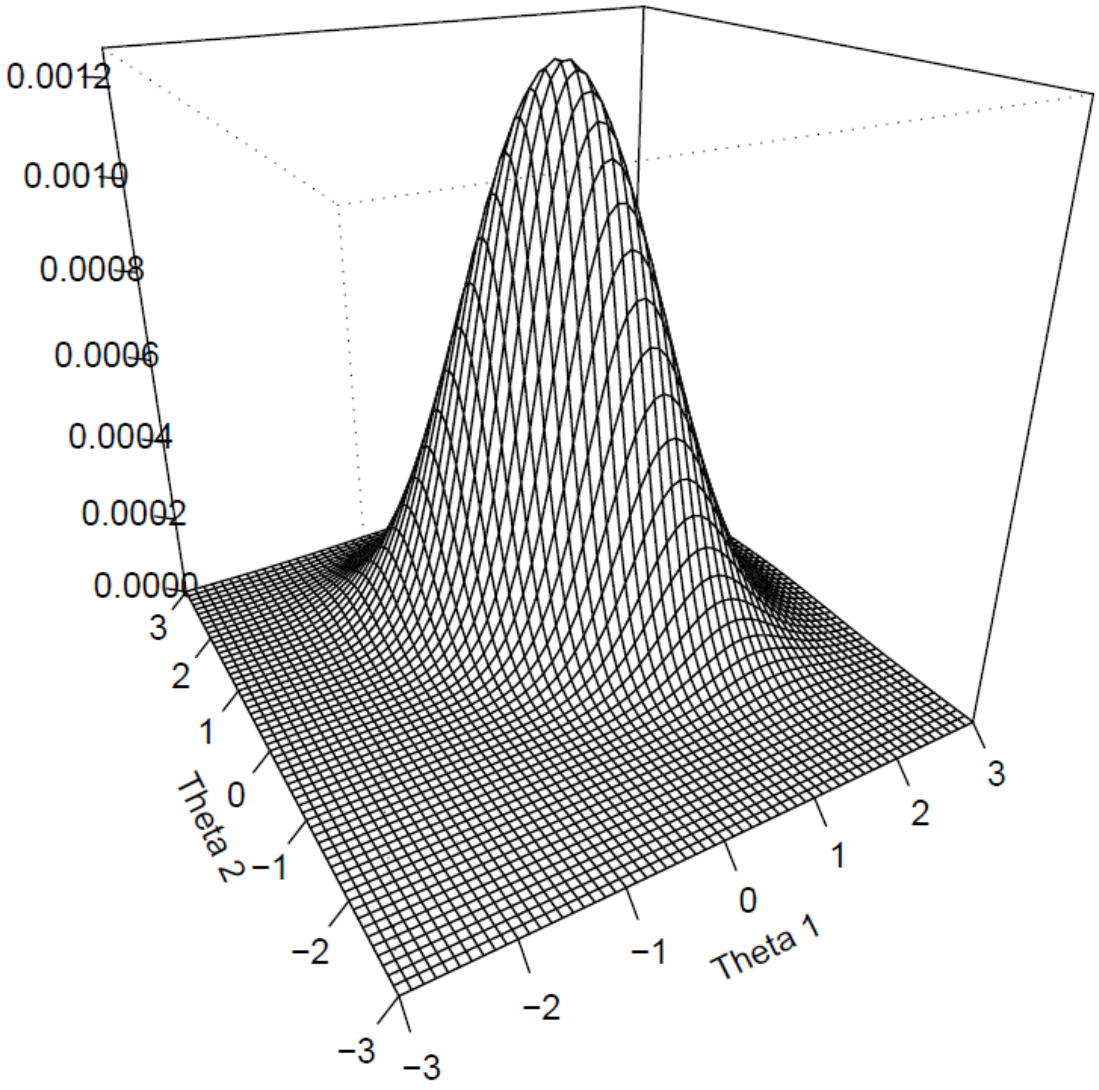
prior distribution and obtain posteriors for response patterns. In the multivariate case, the posterior function takes the form,

$$f(\boldsymbol{\theta} \mid \mathbf{U}) = L(\mathbf{U} \mid \boldsymbol{\theta})\phi(\boldsymbol{\theta}), \tag{5}$$

where the posterior $f(\boldsymbol{\theta}|\mathbf{U})$ is the product of the response pattern likelihood $L(\mathbf{U}|\boldsymbol{\theta})$ and $\varphi f(\boldsymbol{\theta})$, the multivariate normal distribution of $\boldsymbol{\theta}$.[2] Using the parameters from Figure 1, Figure 2 displays the posterior density for a correct item response.

[2] When the model takes on an oblique simple structure form, the multivariate normal distribution has a mean vector of zeros and a variance-covariance matrix $\boldsymbol{\Phi}$ with 1 along the diagonal elements and the population based covariances of the dimensions on the off-diagonal elements. In the case where the dimensions are correlated 1.0, and simple structure is imposed, the posterior equation reduces to the unidimensional case.

Figure 2. Multivariate posterior density for a correct response to an item discriminating on two dimensions



The mode of the posterior density in Figure 2 provides the most likely trait estimate based on a correct response and the normal distribution (Maximum A Posteriori, MAP); the mean of that density is the two-dimensional Expected A Posteriori (EAP).

The only reasonable advantage to using a MIRT-, rather than UIRT-based scoring procedure, is if the scores MIRT models produce have large enough gains in reliability to

warrant the added complexity of the model. Theoretically, so long as the dimensions of a test are correlated, MIRT scores should have greater precision than UIRT scores. This is because of what Segall (2000) refers to as "cross-information"- that scores on one dimension inform scores on another dimension. In other words, if dimensions are correlated, then a high score on one dimension is expected to correspond with a high-score on another dimension. The effect of this additional information is either increased score precision or reduced test length. However, the actual increases in reliability due to MIRT are considered marginal (from about .1 to negligible (Segall, 1996; Luecht, 1996)). It may be that gains are most substantial for domains that begin with relatively low levels of information, but are highly correlated with some other more precisely measured domain (requiring correlations perhaps greater than .6).

When faced with strongly correlated dimensions, researchers are presented with a number of alternatives. If the potential dimensions are weakly correlated, then little is gained from the MIRT model and fitting multiple UIRT models seems a better option. If the dimensions are highly correlated, then some degree of score precision is gained through the MIRT model, but perhaps at the cost of scores with complex interpretations. With highly correlated domains there exists the possibility of a general factor which underlies the items, along with some number of group-specific factors which account for variance particular to only subsets of items.

Bifactor Models

Bifactor models (Holzinger & Swineford, 1937; Tucker, 1958; Gibbons & Hedeker, 1992) are used in situations in which a set of items may be represented by a general (or primary) latent variable in addition to a number of secondary dimensions (or group or content factors) which account for covariance specific to subsets of items. The utility of bifactor

models lies in their broad range of application: Bifactor models are useful when multiple dimensions are expected, or when multidimensionality is caused by unwanted local dependence among item subsets.

Gibbons and Hedeker (1992) describe the bifactor structure in which all items receive one slope parameter on the general dimension and one slope parameter on a secondary dimension (see Table 1). This structure is imposed *a priori* by researchers in situations in which a single dimension is hypothesized to underlie all items on a scale, but additional dimensions are require to account for covariation specific to subsets of items. For example, bifactor models have been used in psychological studies aimed at understanding inter-related but distinct concepts including, but not limited to, depression, anxiety, and anger (e.g., Simms, Gros, Watson, & O'Hara, 2008; Irwin, et al., 2010). In this framework, each concept is represented by a content-specific subfactor, and the primary dimension may be described as general distress/dysphoria. In educational settings, the bifactor model is often used in tests of reading comprehension, where reading passages are followed by a set of related items. The general factor of the bifactor model is reading comprehensions, and additional specific-factors are required for items belonging to each passage.

Table 1. Example of bifactor structure

Item	θ_1	θ_2	θ_3	θ_4
1	a_{11}	a_{12}		
2	a_{21}	a_{22}		
3	a_{31}		a_{33}	
4	a_{41}		a_{43}	
5	a_{51}			a_{44}
6	a_{61}			a_{54}

Alternatively, when bifactor models are employed to account for undesired local dependence, a modification to the bifactor structure is made in which only locally dependent subsets of items receive the additional specific-factor loading (Table 2). These models are

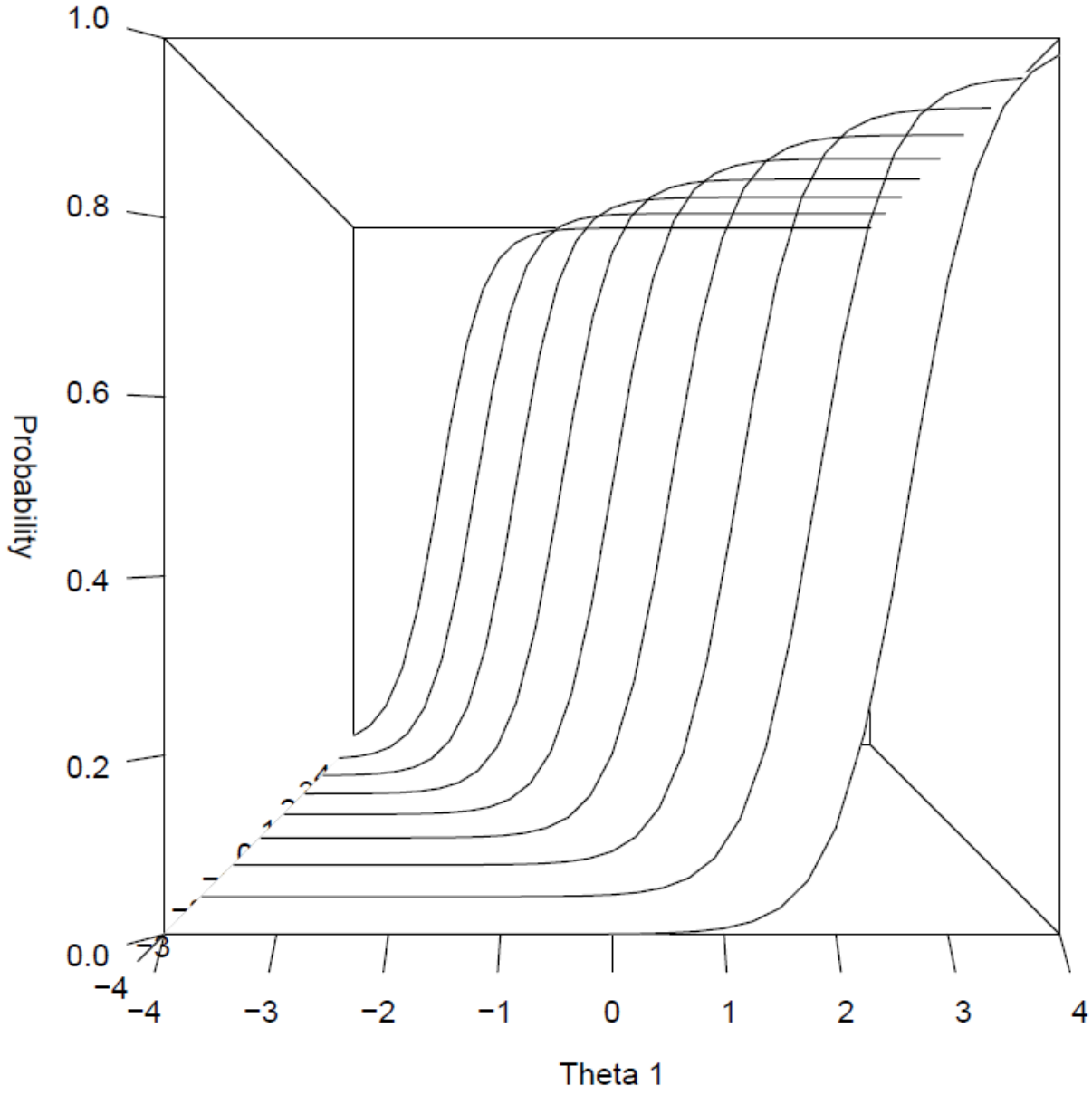
appropriate in situations in which a single latent variable is hypothesized, but subsequent analyses reveal the presence of unaccounted relationships among subsets of test items. In such situations local independence may be achieved by modeling the additional relationships with one or more sub-factors. The following section provides an example of modeling nuisance local dependence in a set of depression items with a modified-bifactor model.

Table 2. Example of modified-bifactor structure

Item	θ_1	θ_2
1	a_{11}	a_{12}
2	a_{21}	a_{22}
3	a_{31}	
4	a_{41}	
5	a_{51}	
6	a_{61}	

Both the bifactor and modified-bifactor models serve as an indication of the dimensionality of a collection of items. In assessing dimensionality, interpretation is focused on the model's slope parameters. Specifically, the magnitude of the secondary dimension slope parameter indicates the influence of this dimension in accounting for relations among responses, but the ratio between general and secondary dimension slopes also indicate the relative strength of each factor. However, interpreting the model in this manner remains limited to assessing the probability of response on one dimension *conditional* on the model's other dimension(s). Figure 3 illustrates this concept. Using the same parameter estimates as Figure 1 ($a_1 = 3$, $a_2 = 2$, $c = 0$), the trace surface is now viewed along the θ_1 dimension and the θ_2 slopes have been removed. What remains are the item's θ_1 trace lines conditional on varying locations of θ_2.

Figure 3. θ_1 trace lines conditional on θ_2



In other words, a slope on θ_1 (or the general factor in the bifactor case) does not indicate the marginal relation between an item response and θ_1, but rather the relation between an item response on θ_1 at various locations on θ_2. Because of this, in bifactor models the primary dimension slope interpretation may be confused or misleading.

Depressive Symptoms Example

To illustrate the difficulties in interpreting bifactor item parameters, considered below are a series of analyses conducted on an eight-item subset of the Patient Reported Outcomes Measurement Information System (PROMIS) pediatric Depressive Symptoms scale (Irwin, et al. 2010). The original 14-item scale was developed from 22 tryout items administered to at least 759 youth aged 12-17 in hospital clinics in North Carolina and Texas. For the purpose of this illustration, eight items were selected to be re-analyzed as a separate scale (the eight items may be found in Table 3). Six of the eight items were from the scale as ultimately assembled and published (Irwin, et al. 2010), and the two additional items were from a locally dependent pair of items identified and set aside during the item tryout period. Specifically, in the analyses reported by Irwin et al. (2010), the items "I cried more than usual" and "I felt like crying" were modeled with a residual correlation in a factor analytic framework.

In this illustration, we fit unidimensional IRT and bifactor MIRT models to the eight-item subset. Table 3 provides the results of fitting two separate models. The first model assumes unidimensionality, while the second bifactor model estimates a general dimension and a secondary dimension for the two locally dependent items with equality constraints on the slope parameters. Comparing the slope parameters on the six unidimensional items between the unidimensional IRT model and bifactor MIRT model indicates that the slope parameters differ little (less than 0.1). However, for the two locally dependent "crying" items, the slope parameters on the primary dimension substantially increase in comparison to the unidimensional estimate (more than 0.4). This effect occurs for any bifactor model in which the slopes on the secondary dimensions are non-zero. The compensatory nature of

model accounts for item responses based on the total number of dimensions present. For items with slopes constrained to zero for the secondary dimension, the interpretation is consistent with the unidimensional model, and any difference in slopes may be due to the additional variance accounted for by the secondary dimensions[3]. However, for items with non-zero slopes on more than one dimension, the interpretation of the primary dimension slope must be made conditional on the secondary dimension slope. In other words, it would be incorrect to interpret the primary dimension slope, for an item with more than one slope, as one would a univariate slope parameter. It may then be desirable to obtain the *marginal* relation between the item response and primary dimension that averages over the secondary dimension(s).

Table 3. Unidimensional and bifactor slope parameters for eight depressive symptoms items.

	UIRT	MIRT	
Item	a	$a_{Primary}$	$a_{Subfactor}$
I cried more than usual.	1.78	2.22	1.94
I felt like crying.	1.79	2.33	1.94
I felt everything in my life went wrong.	2.39	2.49	----
I felt like I couldn't do anything right.	2.31	2.45	----
I felt alone	2.20	2.12	----
I felt so bad that I didn't want to do anything.	1.93	1.98	----
Being sad made it hard for me to do things with my friends.	1.92	1.94	----
I wanted to be by myself.	0.73	0.75	----

Note: Items in italics have been previously identified as locally dependent (Irwin et al., 2010).

This dissertation develops a technique for computing the marginal, or average, trace line for the primary dimension after accounting for secondary dimensions in bifactor models and assesses the appropriateness of a logistic approximation (Chapter 2). Next, the

[3] In this example, the difference in slopes for the six unidimensional items suggests that the two locally dependent "crying" items have re-oriented the latent variable in the unidimensional model. When the dependence between the "crying" items is accounted for, the slopes on the primary dimension slightly increase for the six items, indicating primary dimension is less influenced by local dependence.

technique is used to compute IRT-based primary dimension scale score estimates and standard errors, which are then compared to the primary dimension estimates computed from Cai's two tier algorithm (2010) (Chapter 3). Chapter 4 provides a real-data example using the PROMIS Asthma Symptoms scale, places the method of computing marginal trace lines in the context of the two tier algorithm.

CHAPTER 2

COMPUTING AND APPROXIMATING MARGINAL TRACE LINES

In a 2-dimensional MIRT model, to obtain the marginal trace line for θ_1, one must average over the θ_2 dimension of the multivariate trace surface:

$$T_{\text{Marginal}_i}(u_i = 1 \mid \theta_1) = \int T_i(\theta_1, \theta_2)\phi(\theta_2)d\theta_2. \quad (6)$$

In this 2-dimensional example, the product of the θ_1 conditional trace lines from the trace surface, $T_i(\theta_1, \theta_2)$ and the univariate normal distribution, integrated across θ_2, represents the marginal trace line for θ_1, T_{Marginal}. Note that (6) is essentially computing the marginal trace line by weighting the θ_1 conditional trace lines by the normal distribution. Because of this weighting process, marginal trace lines will never be greater in magnitude than the conditional trace lines along θ_1, and depending on the relationship between the conditional slopes (a_1 and a_2) the marginal slope may be much smaller.

Interpretation of the marginal trace line is not unlike the univariate trace line in unidimensional IRT; the marginal trace line is the relationship between the probability of response given θ_1, after accounting for the secondary dimension(s).[4] Using the parameters for the first item in the depressive symptoms example in Chapter 1, *I cried more than usual*, one may illustrate this phenomenon by considering the θ_1 marginal (6) and conditional trace lines (2) at various locations on θ_2 (Figure 4). The varying degrees of line width are meant to suggest that conditional trace lines closer to the mean of the normal distribution receive more

[4] The marginal trace line has also been referred to as an *expected score curve* by Schultz and Lee (2002) and Donoghue (1997); their techniques were used to derive achievement level boundaries.

weight than those near the tails of the distribution. Clearly the slope of the marginal trace line is less than that of the conditional trace lines. Recall that a_1 for the item *I cried more than usual* was 2.22, but after computing the marginal, a_{Marginal} is reduced to 1.46. As researchers interpret such parameter estimates, the (conditional) slopes on the general dimension may be misleading as they suggest a relationship which is in fact inflated due to the secondary dimension. After accounting for the secondary dimension, the marginal trace line gives a more realistic account of the relationship between the item and general factor.

Figure 4. Marginal and conditional trace lines

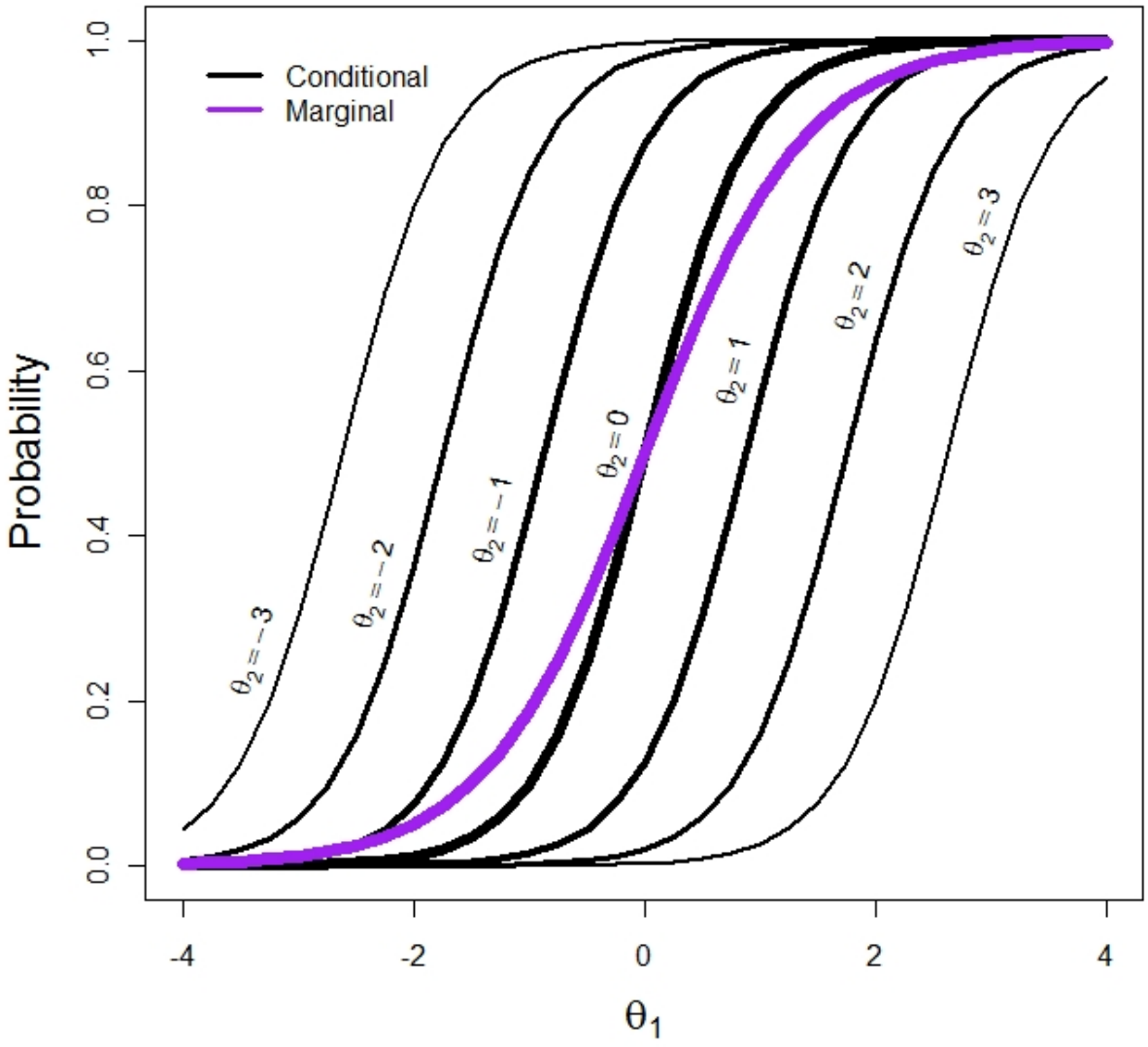


Figure 4 also illustrates the curious fact that when item calibration moves from a unidimensional model to a bifactor model, slope parameters on the factor of interest tend to increase, giving the false impression that items are more representative of the general factor. Rather, in keeping with prior literature (Reckase, 1979), one might expect that unmodeled LD should produce over-estimates of slope parameters on the on the unidimensional factor (i.e., "θ theft"; Thissen and Steinberg, 2010), and that after accounting for LD, slopes on the

general dimension should decrease. This surprising phenomenon is again attributable to slopes in bifactor models being conditional, rather than marginal, representations of item responses; when the marginal trace line is computed (as in Figure 4), the slope parameter takes on a more realistic value.

Logistic Approximations

Note that thus far the marginal trace line has been derived from a MIRT model, but has no item parameters which describe the relationship between the item and marginal θ_1 distribution. The expected score curve from a multidimensional logistic trace surface with a normal population distribution is *not* a logistic function or, indeed, any "simple" function. The marginal trace line may be thought of as an average of the θ_1 conditional trace lines weighted by the normal distribution, and a logistic approximation of this average trace line may suffice.

There is historical precedent for treating the summation of logistic functions as approximately logistic. Winsor (1932) notes that the "sum of a number of logistics does in fact often approximate closely a logistic as has been shown by Reed and Pearl (1927)" (p. 4). Reed and Pearl (1927) use sums of logistics to describe population growth, and later, Merrell (1931) would examine averages of individual growth curves to describe change over time for groups of individuals. From this perspective, one should expect that the marginal trace line, which is itself a weighted average of logistic curves, should be closely approximated by a logistic function.

Given the marginal trace line, to find a logistic approximation, one must estimate item parameters. A potential approximation to be considered involves computing the derivative of the marginal trace line at $T = 0.5$, which is an estimate of the slope of an

approximately logistic function. We may approximate the derivative by taking values of T_{marginal} and θ_1 near $T = 0.5$:

$$\hat{a} = \frac{\log \frac{T_H}{1-T_H} - \log \frac{T_L}{1-T_L}}{\theta_H - \theta_L}, \tag{7}$$

where T_H represents a probability slightly higher than 0.5, T_L a probability slightly lower than 0.5, and θ_H and θ_L the respective θ_1 trait values. The ratio between the difference in the log odds of two probabilities near 0.5 and their respective θ_1 values gives the slope of the function or the marginal trace line $\hat{a}_i$. Next, the threshold or difficulty parameter is the location on θ_1 where T_{Marginal} is 0.5, and in practice is computed directly from the dimension of interest in the MIRT model. For example, in a MIRT model where the marginal is desired for θ_1, the threshold is:

$$\hat{b}_i = \frac{-c_i}{a_{i1}} \tag{8}$$

We may then approximate the marginal trace line using the traditional unidimensional logistic function and $\hat{a}$ and $\hat{b}$ (Birnbaum, 1968). In practice, the $\hat{a}$ parameters computed from the numerical derivative tend to be sensitive to the number of quadrature points provided. Because of this concern, an alternative method is used as originally proposed by Ip (2010a; 2010b).

Ip's method of approximation is equivalent to transforming the MIRT slope parameters into the factor analytic loading metric, and then back-translating to arrive at the marginal slopes. Ip's method is equivalent to computing the marginal factor loading for the dimension of interest, in this case θ_1:

$$\lambda_{\text{Marginal}} = \frac{a_1/D}{\sqrt{1+\sum\left(a/D\right)^2}}, \tag{9}$$

where D is the commonly used scaling constant 1.7. The item variance unexplained by the primary latent dimension is then:

$$\sigma^2_{\text{Marginal}} = 1 - \lambda^2_{\text{Marginal}}\ , \tag{10}$$

and the marginal slope parameter is simply:

$$\hat{a} = \left(\frac{\lambda_{\text{Marginal}}}{\sqrt{\sigma^2_{\text{Maginal}}}}\right)D\,. \tag{11}$$

As with the numerical derivative, $\hat{b}$ is unchanged after computing the marginal. A logistic approximation of the marginal trace line uses $\hat{a}$ and $\hat{b}$ in the traditional fashion of the 2-parameter logistic model:

$$T\left(u_i = 1 \,|\, \boldsymbol{\theta}_1\right) = \frac{1}{1+\exp[-\hat{a}_i(\boldsymbol{\theta}_1 - \hat{b}_i)]}\,. \tag{12}$$

Extensions to the graded response model (GRM; Samejima, 1969) provide no additional complications as the thresholds and slope parameter may be computed as in (8) and (11), respectively. For binary items modeled with the 3-PL to account for guessing, Ip (2010a) notes that the lower asymptote g_i is unaffected by marginalization (i.e., $\hat{g}_i = g_i$).

The Logistic as a Close Approximation of the Marginal Trace Line

In order to justify the use of the logistic, it is important to assess the degree to which it approximates the marginal trace line. Regarding the use of the logistic distribution to appoximate the normal CDF, Haley (1952) notes that the two never differ by probability values greater than 0.01. Here one might anticipate similar results (and Ip (2010a) provides a

graphical illustration of a close approximation), but the degree to which the normal distribution influences the shape of the marginal trace line remains unknown.

The closeness of the logistic approximation to the marginal trace line is here considered both graphically and numerically. For the numerical comparison between marginal trace lines and logistic approximations, a wide range of marginals were computed from various 2-dimensional 2-PL trace surfaces which varied in the magnitude of the a_1 and a_2 slope parameters (all intercept parameters were 0.0). All combinations of trace surfaces were considered from a_1 and a_2 values of 1.0 to 4.5 (a range which liberally incorporates most values seen in practice) in increments of 0.1, resulting in 1,296 unique trace surfaces. Using these trace surfaces, comparisons were made between each marginal trace line and logistic approximation. For each of the 1,296 comparisons, across 81 quadrature nodes between -4 to 4 standard deviations from the mean, the maximum difference in probability between the marginal and logistic approximation of the trace line was no more than ±0.011 (for all positive differences there is a corresponding negative difference of the same magnitude). For each of the 1,296 trace line comparisons, across the range of θ_1, the maximum difference in probabilities between the marginal trace line and logistic approximation ranged from ±0.006 to ±0.011 (mean = 0.010, SD = 0.001). While there was very little difference among the various trace surfaces considered, there was a slight trend that the most precise approximations occurred when the a_2 parameters were low in magnitude, indicating that a weak secondary dimension has little influence on either the marginal trace line or the logistic approximation.

These numeric results may also be illustrated graphically. To demonstrate the appearance of these slight differences between the logistic approximation and the marginal

trace line, four different models were considered (see Table 4). For the first two MIRT models the primary dimension slope is large in magnitude relative to the secondary dimension slope ($a_1 = 3.0$ and $a_2 = 2.0$), and the intercepts are the threshold equivalent to $b =$ 1.5 and -1.5. The second two models have the same intercepts/thresholds, but reverse the magnitude of the primary and secondary dimension slopes.

Table 4. Four 2-PL MIRT models and corresponding marginal parameter estimates

Panels for Figure 5 and 6	a_1	a_2	c	a_{Marginal}	b_{Marginal}
Upper left	3.0	2.0	-4.5	1.94	1.5
Upper right	3.0	2.0	4.5	1.94	-1.5
Lower left	2.0	3.0	-3.0	0.99	1.5
Lower right	2.0	3.0	3.0	0.99	-1.5

Figure 5 illustrates the close approximation between the logistic and marginal trace lines. The marginal is nearly indistinguishable from the logistic approximation, and appears to be unaffected by differences in location parameters, as suggested by Ip (2010a). Figure 6 illustrates the marginal trace lines and logistic approximations after a log odds transformation. Because the logit of a logistic function is linear, the approximations will always be linear. The marginal however, can strictly speaking can never be linear, and any deviation represents missfit of the logistic approximation. The logits in Figure 6 illustrate this fact as deviations in marginals begin to appear in the tails of the distributions. However, at such extreme values in log odds, the probability equivalent is actually quite small (between 0.00002 and 0.003 at the most extreme values of θ_1 for each comparison in Figure 6), providing further evidence of the utility of the logistic as an appropriate appoximation.

Figure 5. Four 2-PL marginal trace lines and logistic approximations

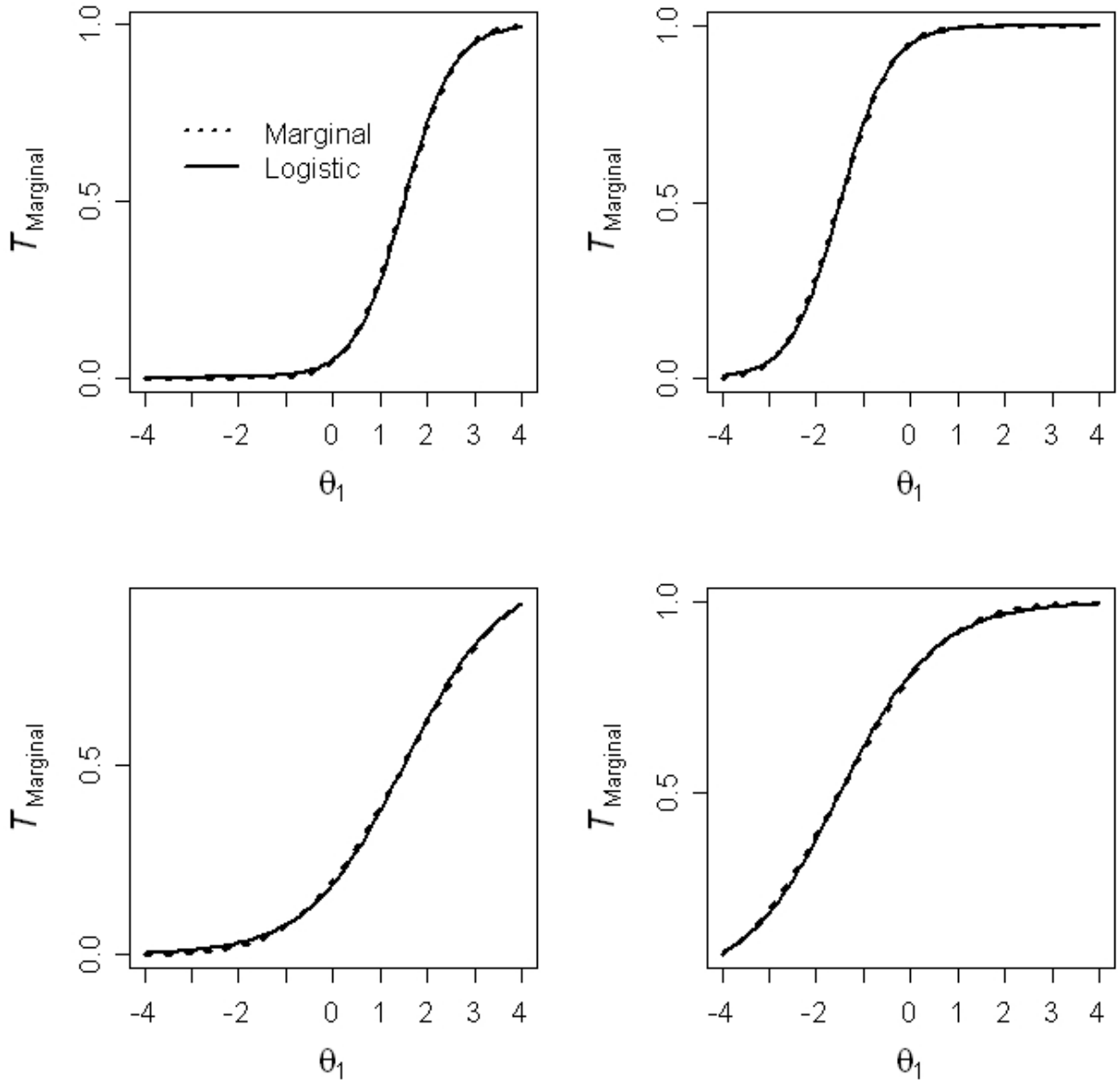
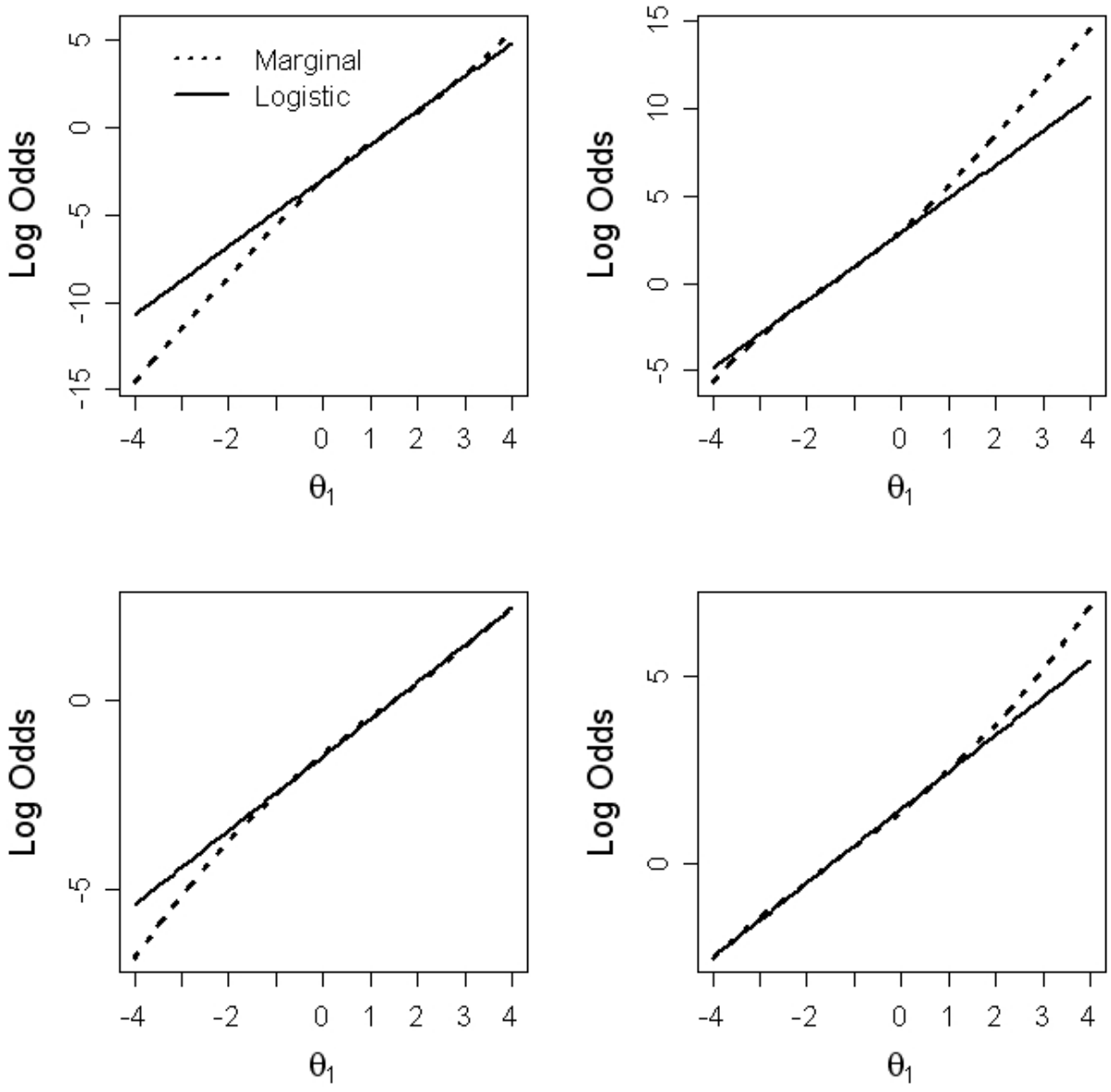


Figure 6. Four 2-PL marginal trace lines and logistic approximations in log odds



The Relation Between Conditional and Marginal Slope Parameters

While Table 4 provides the marginal slope parameter of the logistic approximation for four sets of conditional slope parameters, in practice the computations needed to compute marginal slopes (equation 6) can be carried out for any combination of conditional slopes. Thus, it may be of some interest to provide the relation between conditional slope parameters and the resulting marginal slope parameter. Initially, the magnitude of the marginal slope for some simple bifactor MIRT models which have equal a_1 and a_2 slope parameters is considered.

Figure 7. Marginal slopes for items with equal conditional slopes on two dimensions

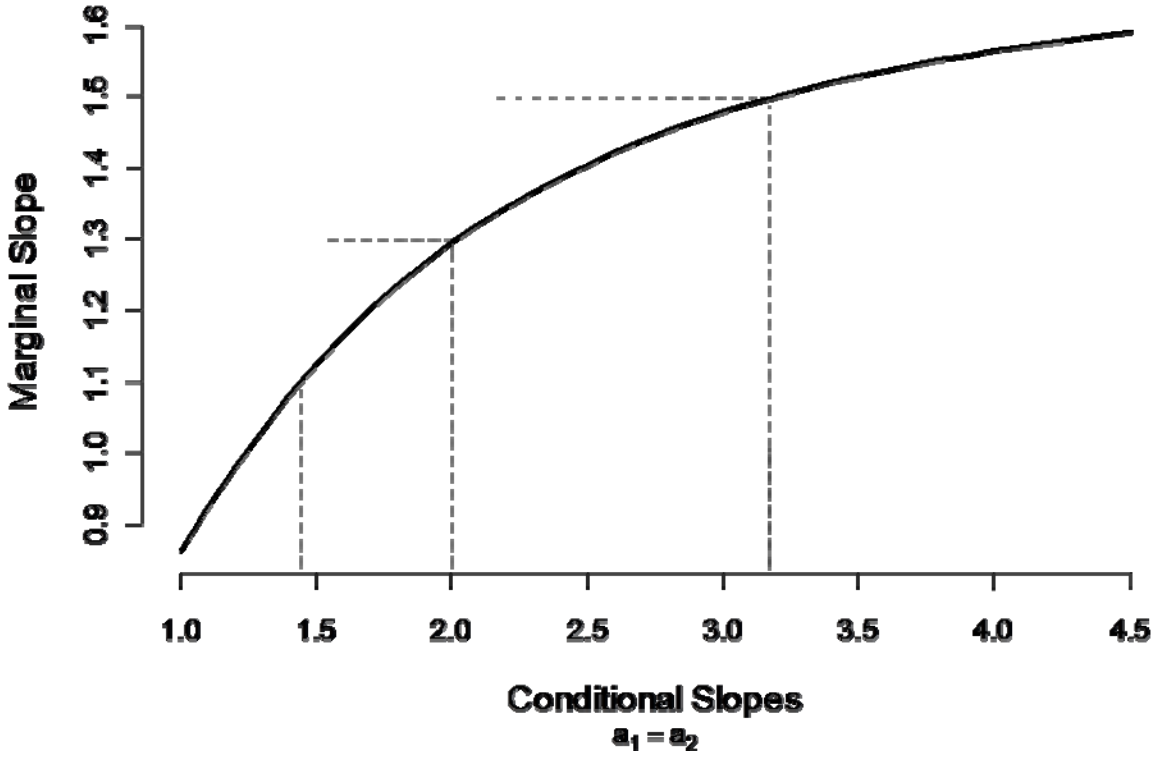


Figure 7 illustrates the relationship between the magnitude of the marginal slope parameter and magnitude of the equal slopes on the primary and secondary dimensions. For

illustrative purposes, horizontal grey lines indicate increases in the marginal slope of 0.2 units, and correspond to marginal slopes of 1.1, 1.3, and 1.5. In general, when conditional slopes are weak (e.g., when $a_{\text{Conditional}}$ is almost 1.5), the marginal slope is also weak and only slightly less than the conditional slopes (e.g., a_{Marginal} =1.1). The marginal slope increases quickly from 1.1 to 1.3 with only slight gains in the conditional slopes. For example, increasing the conditional slopes about 0.5 units from 1.5 to 2.0 results in a marginal slope increasing about 0.2 units from a marginal slope of 1.1 to marginal slope of 1.3. However, gains in the marginal slope quickly diminish as the conditional slopes become large. Continuing with the present example, to achieve an additional gain in the marginal slope of 0.2 (i.e., $a_{\text{Marginal}} = 1.5$) requires an increase in conditional slopes of more than 1.0 units (i.e., $a_{\text{Conditional}} = 3.2$). For most applications, conditional slopes constrained to be equal will not be greater than this, because such slopes correspond to the dimensions accounting for nearly 90% of the item variance.

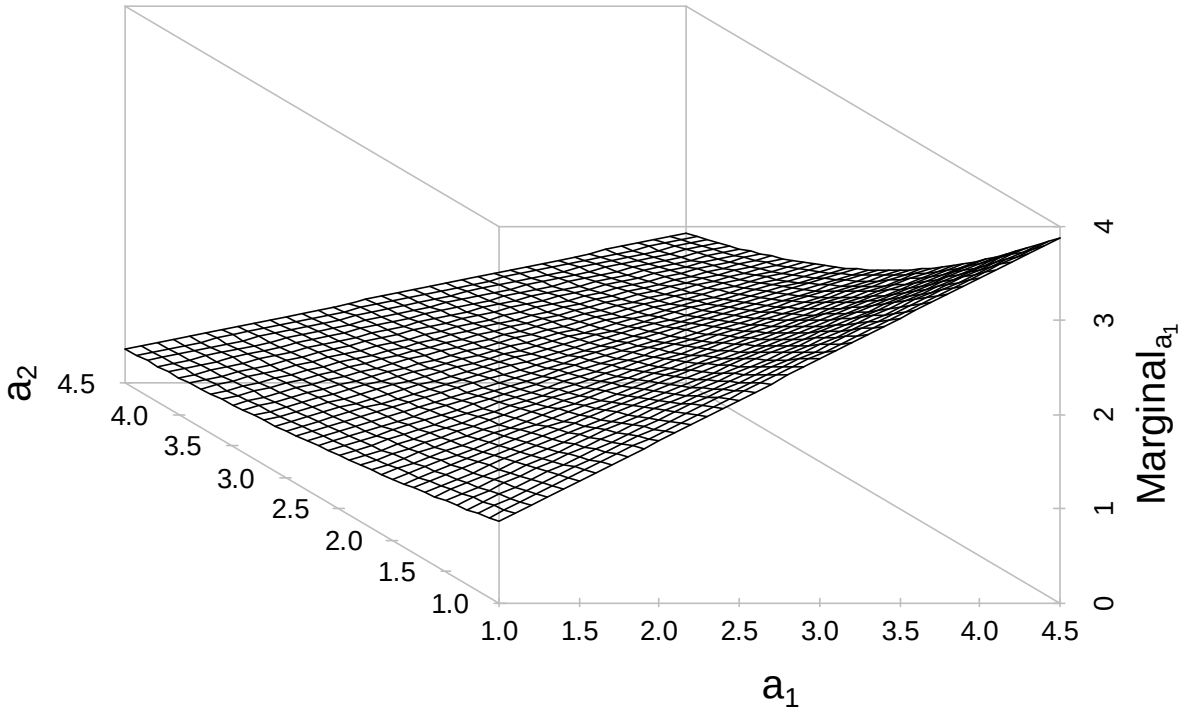
While conditional slopes much greater than this are unlikely and may be evidence of Heywood cases, they do illustrate an interesting fact of the marginal. For bifactor MIRT models with the conditional slopes constrained to be equal on two dimensions, the marginal slope is bounded by the logistic scaling constant. That is, as the conditional slopes increase, the item variance accounted for becomes nearly 50% for each dimension. Using (9) and (10), if the item variance explained is 50%, then the resulting marginal corresponds to the scaling constant (here, 1.7), or a factor loading of about 0.707.

For more general bifactor MIRT models with unequal conditional slopes, Appendix I may serve as a quick reference. The table provides the slope of the logistic approximation of the marginal trace line resulting from conditional slopes which vary from 1.00 to 4.50 in

increaments of 0.25. Because no current software programs compute marginal trace lines, the table should provide interested researchers with the magnitude of the marginal slope parameter for a wide variety of conditional slopes, and interpolation may be used for interpretation purposes.

The relationship between a variety of conditional slopes and marginal slopes may also be considered graphically. Figure 8 shows the functional relationship between conditional and marginal slopes. This non-trivial function cannot be easily approximated and is here presented for illustrative purposes only. The relationship is reasonably linear except for increasingly high values on θ_1 and low values on θ_2, which results in marginal slopes that increase rapidly. This phenomena will continue as a_1 increases and a_2 approaches zero. This relationship can also been seen in Appendix I.

Figure 8. Marginal slopes across a range of conditional slopes on two dimensions



CHAPTER 3

COMPUTING ITEM RESPONSE THEORY SCORES FROM MARGINAL TRACE LINES

This chapter considers the utility of using logistic approximations of marginal trace lines in a variety of test scoring applications. Specifically, logistic approximations of marginal trace lines are used to compute unidimensional IRT-scaled scores for the general, or primary, dimension in bifactor IRT models. It is proposed that these scaled score estimates will provide a close approximation to the primary dimension point estimates used in traditional MIRT scoring (Segall, 1996, 2000). In one recent example of MIRT scoring, Cai (2010) provides a two-tier algorithm in which the dimensionality of the integration for a multidimensional bifactor model is reduced to the number of primary dimensions plus one. Use of this estimation procedure results in a vector of ability estimates for the number of dimensions. From this vector of ability estimates, computed from the multivariate posterior density, the first element $\hat{\boldsymbol{\theta}}_1$ should be the same as the score estimate computed using the marginal posterior distribution (Segall, 2001).

The remainder of this chapter considers the degree to which unidimensional scoring computations using logistic approximations of marginal trace lines provide primary dimension scores and standard error estimates similar to those obtained using the two-tier algorithm for multidimensional models. If the two methods are comparable, then use of the unidimensional logistic approximation technique may provide a simpler, less computationally burdensome method for scoring the primary dimension. What follows is a

comparison of IRT-scores and standard error estimates obtained from the logistic approximation of the marginal trace line with the conventional MIRT scoring technique implemented using the two-tier algorithm.

An Overview of IRT-Scaled Scores for Response Patterns and Summed Scores

Many applications of IRT-based scoring use the individual's complete response pattern in forming the scaled score estimate. Known as response pattern scoring, the point-estimate is the mean or Expected A Posteriori (EAP) from the posterior distribution (Bock & Mislevy, 1982):

$$L(\mathbf{u} \mid \theta) = \prod_{i=1}^{items} T_i(u_i \mid \theta)\phi(\theta), \qquad 12$$

where the posterior distribution is the product of the trace lines for each response *u* to item *i* and the prior density (here normally distributed with a mean of zero and a standard deviation of one). The mean of the posterior density may be computed by approximating the integral over a range of quadrature points *q*:

$$\mathrm{EAP}(\theta) \approx \frac{\sum_{1}^{q}\prod_{i=1}^{items} T_{iq}(u_i)\phi(\theta_q)\theta_q d\theta_q}{\sum_{1}^{q}\prod_{i=1}^{items} T_{iq}(u_i)\phi(\theta_q)d\theta_q}. \qquad 13$$

Likewise, the standard deviation of any given posterior may also be computed by approximation:

$$SD(\theta) \approx \sqrt{\frac{\sum_{1}^{q}\prod_{i=1}^{items} T_i(u_i)\phi(\theta_q)(\theta_q - \mathrm{EAP}[\theta])^2 d\theta_q}{\sum_{1}^{q}\prod_{i=1}^{items} T_i(u_i)\phi(\theta_q)d\theta_q}}. \qquad 14$$

As a function of the item parameters, the posterior standard deviation is allowed to fluctuate across the range of the latent variable.

While response pattern EAPs and SDs incorporate all available information from an individual's responses to a set of items, the number of response patterns (i.e., the number of response categories to the power of the number of items) often makes tables of such response patterns, scores, and standard deviations unwieldy. As an alternative, one may compute the IRT-based expected value of the latent variable given the respondent's summed score rather than response pattern. Scoring tables of summed scores and their associated EAPs and SDs are user-friendly alternatives to response pattern scores and are readily interpretable. It is possible to compute the expected value of the posterior for every summed score x which is itself the sum of the response vector **u**:

$$L_x(\theta) = \sum_{x=\sum \mathbf{u}}^{\text{responsepatterns}} \prod_i T_{u_i}(\theta)\varphi(\theta). \qquad 15$$

A recursive algorithm introduced by Lord and Wingersky (1984), and described in detail by Thissen, Pommerich, Billeaud, and Williams (1995), is used to compute $L_x(\theta)$. Briefly, the recursive algorithm may be viewed as an updating process which is initialized by the trace line for a single item T_1 where the likelihood for a summed score of 1 is $L_{x=1} = T_1$, and the likelihood for summed score of 0 is $L_{x=0} = (1-T_1)$. When a second item is added to the test, the likelihood of a summed score of 0 is $(1- T_1)*(1- T_2)$; the likelihood for summed score of 2 is T_1*T_2; and the likelihood of summed score of 1 is the sum of $T_1(1-T_2)$ and $T_2(1-T_1)$. This updating process continues until the likelihoods for all possible summed scores are evaluated.

Because summed score based EAPs incorporate information available from all response patterns that yield a given summed score, and some of these response patterns may form likelihoods around different locations of the latent variable, any particular summed score likelihood will be slightly wider than the component response pattern likelihoods. For all but

the most extreme response patterns, the loss of information when using IRT-scores from summed scores results in score standard deviations being inflated about 10% (i.e., a 10% loss in precision (Thissen, et al., 1995)), though the correlation between response pattern and summed score-based EAPs is often greater than 0.95.

The decision to use IRT-scores from summed scores also allows intuitive and simple comparison between the two scoring methods. An advantage of using IRT-scores from summed scores is that they are a function of the previously estimated item parameters and all possible response patterns. Because these patterns are known and used in the recursive algorithm, there is no reliance on samples of individuals to provide IRT-scaled score estimates from summed scores.

Rather than comparing individual's scaled scores using samples of response patterns, comparing summed score-based EAPs and SDs from the logistic approximation of the marginal trace line and the MIRT two-tier algorithm is quick and easy and may be computed directly from the MIRT item parameters. For instance, consider a multidimensional six-item binary test with seven possible summed scores on the primary dimension (0, ... , 6). Any difference in the seven EAPs between the two methods is interpreted as score bias when using the logistic approximations. The ratio between the score standard deviations represents potential bias in score precision between the two methods. For instance, if a particular score had standard deviation of 0.60 for the logistic approximation method and 0.80 for the two-tier approach, the ratio between the two-tier score standard deviations and logistic approximation of the marginal trace lines (0.80/0.60) would indicate that the logistic approximation scores appeared to be 1.33 times more precise. Such a finding would indicate

a bias of the logistic approximation method. Using such simple comparisons, many scores and standard errors can be evaluated from a variety of models.

The Method of Evaluating Primary Dimension Scores across MIRT Models

The methods used in this dissertation involve comparing scaled scores and score standard deviations between the logistic approximation and the two-tier algorithm for a variety of MIRT models (or *tests*) for binary items. All MIRT model parameter estimates are considered known.[5] The steps involved for the comparisons are as follows: (1) For all multidimensional tests, the primary dimension scaled scores from summed scores are initially computed using the two-tier algorithm, (2) next the marginal trace lines and logistic approximations of them are computed for all items using the methods presented in Chapter 2, (3) finally, the recursive algorithm is used to compute the comparable primary dimension scores and score standard deviations from the logistic approximations. These steps are repeated for all tests.

To evaluate the utility of scoring tests using the logistic approximation of the marginal trace line, the two scoring approaches are compared across a variety of MIRT models. Comparisons between the two methods will take into account three model-related conditions, *factor loadings*, *test length*, and *dimensionality*. The following model conditions are meant to reflect a wide range of bifactor models used in research settings.

First, to compare models which vary in influence of the secondary dimension, the *factor loading*[6] conditions will consider different ratios between the magnitude of the primary and secondary dimension loadings. A range of factor loadings will be divided into three groups

[5] For simplicity, the thresholds of all MIRT parameter estimates (modeled as intercepts) were fixed at the mean of the latent variable ($\theta = 0$). Findings in Chapter 2 indicate that fit of the logistic approximation to the marginal trace line is independent of the location of item's location parameter.
[6] To provide a more readily interpretable metric, factor loadings are reported. All computations were performed with slope parameters converted from factor loadings.

(low (.3 to .5), medium (.5 to .7), or high (.7 to .9)), following guidelines used by McDonald (1999) and Reise, Cook, and Moore (under review). Items with multidimensional structure may have primary and secondary dimension slopes which are combinations of low, medium, and high (e.g., "high" primary and "low" secondary, "low" primary and "high" secondary", etc.). Note that "high" factor loadings on both the primary and secondary dimensions may result in negative residual variances or so-called *Heywood* cases. Thus, this condition was eliminated resulting in eight different primary and secondary factor loading conditions. Varying the factor loadings across dimensions in this manner provides a means of detecting potential biases in scores based on the strength of a particular dimension. These biases, however minor, may be compounded depending on the strength of loadings and test length.

In addition to differences in factor loadings across dimensions, it is also of interest to consider multiple test lengths. For instance, for a long test with one pair of LD items, there may be little utility in computing the marginal trace line given simpler traditional methods (e.g., setting items aside to eliminate LD), whereas for a short test, which provides less score information, it may be more desirable to consider marginal trace lines as a means of gaining all possible information from the data. Thus, to uncover how the *test length* condition affects scoring the logistic approximations, a few practical test lengths are considered. Based on common lengths of scales in both health outcomes and psychological research, short (6 items), medium (12 items), and long (24 items) tests are considered.

Finally, the design of the *dimensionality* condition will take into account two model fitting situations in which bifactor models are commonly used. The first situation is one in which all items load on both the primary dimension and, because of hypothesized dependence among clusters of items, one secondary dimension (i.e., complete bifactor

structure). The second situation represents a modified-bifactor model in which the items are generally unidimensional, but because of some unplanned nuisance dimensionality, secondary factors in the form of item doublets are needed to achieve conditional independence. For both situations, the number of item clusters and doublets modeled is dependent on the length of the test. For instance, while a short test may be limited to one or two secondary dimensions (modeled as doublets, or two locally dependent items), the medium and long test length conditions includes high-dimensional models which have only recently become practical following advances in MIRT parameter estimation via the two-tier method (Cai, 2010). Because the number of dimensions possible depends on test length, or is nested within test length, the dimensionality condition allows the short test condition to have 1 or 2 secondary dimensions (i.e., 1 or 2 doublet pairs); the medium length test has 3 secondary dimensions with 4-item clusters or 3 doublet pairs of items; and the long test has 6 secondary dimensions with 4-item clusters or 6 doublet pairs of items.

Given these conditions, the study design crosses the strength of the factor loadings on the primary and secondary dimensions with test length (and also the number of dimensions that are nested within test length). This scoring design results in three factor loading conditions (which when crossed yields 8 conditions), three test length conditions, and two dimensionality conditions within each test length condition. The total number of conditions which compare scores computed from the logistic approximation to the two-tier algorithm is then $8_{\text{factor loadings}} \times 3_{\text{test length}} \times 2_{\text{dimensionality}} = 48$. This study design covers the majority of test conditions seen in research settings. These conditions provide insights into the use of logistic approximations of marginal trace lines in providing a better understanding of the relation

between item responses and the primary dimension, and if so, whether or not these techniques are useful in providing an IRT-score for the primary dimension.

An IRT-based Scoring Example

To further illustrate these MIRT scoring conditions, this section follows one of the forty-eight conditions through the entire scoring process. This condition uses the long test (24 items), with complete bifactor structure (six secondary clusters with four items each), and has medium factor loadings on the primary dimension and low factor loadings on the secondary dimensions. The multidimensional factor structure is provided in Table 5. Note from the table that the primary dimension loadings are balanced between the lower (.5) and higher (.7) loadings for the "medium" factor loading condition, and the secondary dimension loadings are balanced between the lower (.3) and higher (.5) loadings for the "low" factor loading condition.

Table 5. Example of bifactor structure for scoring.

Item	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
1	0.5	0.3					
2	0.6	0.3					
3	0.6	0.3					
4	0.7	0.3					
5	0.5		0.4				
6	0.6		0.4				
7	0.6		0.4				
8	0.7		0.4				
9	0.5			0.5			
10	0.6			0.5			
11	0.6			0.5			
12	0.7			0.5			
13	0.5				0.3		
14	0.6				0.3		
15	0.6				0.3		
16	0.7				0.3		
17	0.5					0.4	
18	0.6					0.4	
19	0.6					0.4	
20	0.7					0.4	
21	0.5						0.5
22	0.6						0.5
23	0.6						0.5
24	0.7						0.5

After converting the factor loadings in Table 5 into slopes, the two-tier algorithm, as implemented in the software program IRTPRO (Cai, du Toit, & Thissen, forthcoming) is used to compute primary dimension IRT-scores and standard deviations along with their associated summed scores. After tabulating these values, the marginal trace lines for all 24 items are computed from the MIRT item parameters using the R language for statistical computing and 81 quadrature points equally spaced between -4 and +4. Logistic approximations using the 2PL model are then made from these marginal trace lines. Once the 2PL item parameters are obtained, the recursive algorithm is used to compute the IRT-scaled scores and standard deviations from summed scores. This process results in two sets

of IRT-scaled scores and score standard deviations for the primary dimension. A summed score to scale score translation table is then used to compare the values from the two approaches. Below, Table 6 provides the results for this first of forty-eight scoring comparisons.

Table 6. Example of a score translation table using the logistic approximation of marginal trace lines and the two-tier algorithm.

Summed Score	Two-Tier EAP	Logistic EAP	Two-Tier SD	Logistic SD
0	-2.07	-2.17	0.59	0.56
1	-1.73	-1.83	0.53	0.50
2	-1.47	-1.56	0.49	0.45
3	-1.25	-1.33	0.46	0.42
4	-1.07	-1.13	0.44	0.39
5	-0.90	-0.95	0.42	0.36
6	-0.75	-0.79	0.41	0.35
7	-0.61	-0.64	0.40	0.33
8	-0.48	-0.50	0.39	0.32
9	-0.36	-0.37	0.39	0.32
10	-0.24	-0.25	0.38	0.31
11	-0.12	-0.12	0.38	0.31
12	0.00	0.00	0.38	0.31
13	0.12	0.12	0.38	0.31
14	0.24	0.25	0.38	0.31
15	0.36	0.37	0.39	0.32
16	0.48	0.50	0.39	0.32
17	0.61	0.64	0.40	0.33
18	0.75	0.79	0.41	0.35
19	0.90	0.95	0.42	0.36
20	1.07	1.13	0.44	0.39
21	1.25	1.33	0.46	0.42
22	1.47	1.56	0.49	0.45
23	1.73	1.83	0.53	0.50
24	2.07	2.17	0.59	0.56

Note: Because the difficulty parameters are fixed at zero for all items, the EAP and standard deviations are symmetrical around the mean 0.0.

This table provides the first appearance of bias in scoring when using the logistic approximations of marginal trace lines. For each summed score, across the entire range of

the primary dimension, the scaled scores for the logistic approximation are more extreme than those computed directly from the MIRT model. Though potentially minor, the difference in the logistic-based point estimates is at most ±0.10 standard deviations from the two-tier-based estimate. Additionally, the standard deviations indicate overly narrow posteriors for the logistic approximation (a spurious increase in score precision of about 5% to 18% depending on the location of the latent variable). In other words, the logistic approximation gives the impression that the items provide more information about the latent variable than should be present.

Prior to scoring all 48 conditions, this phenomena, where use of the logistic approximation of the marginal trace line results in a spurious increase in score precision (i.e., overly precise scores), is investigated in a few selected conditions. Figure 9 provides a graphical illustration of the values in Table 6. Parallel results for a 12-item scale with medium slopes on the primary dimensions and low secondary slopes with three doublet factors are shown in Figure 10, and results from an example with one doublet, 6-item scale with high slopes on the secondary dimension are in Figure 11.

Figure 9.

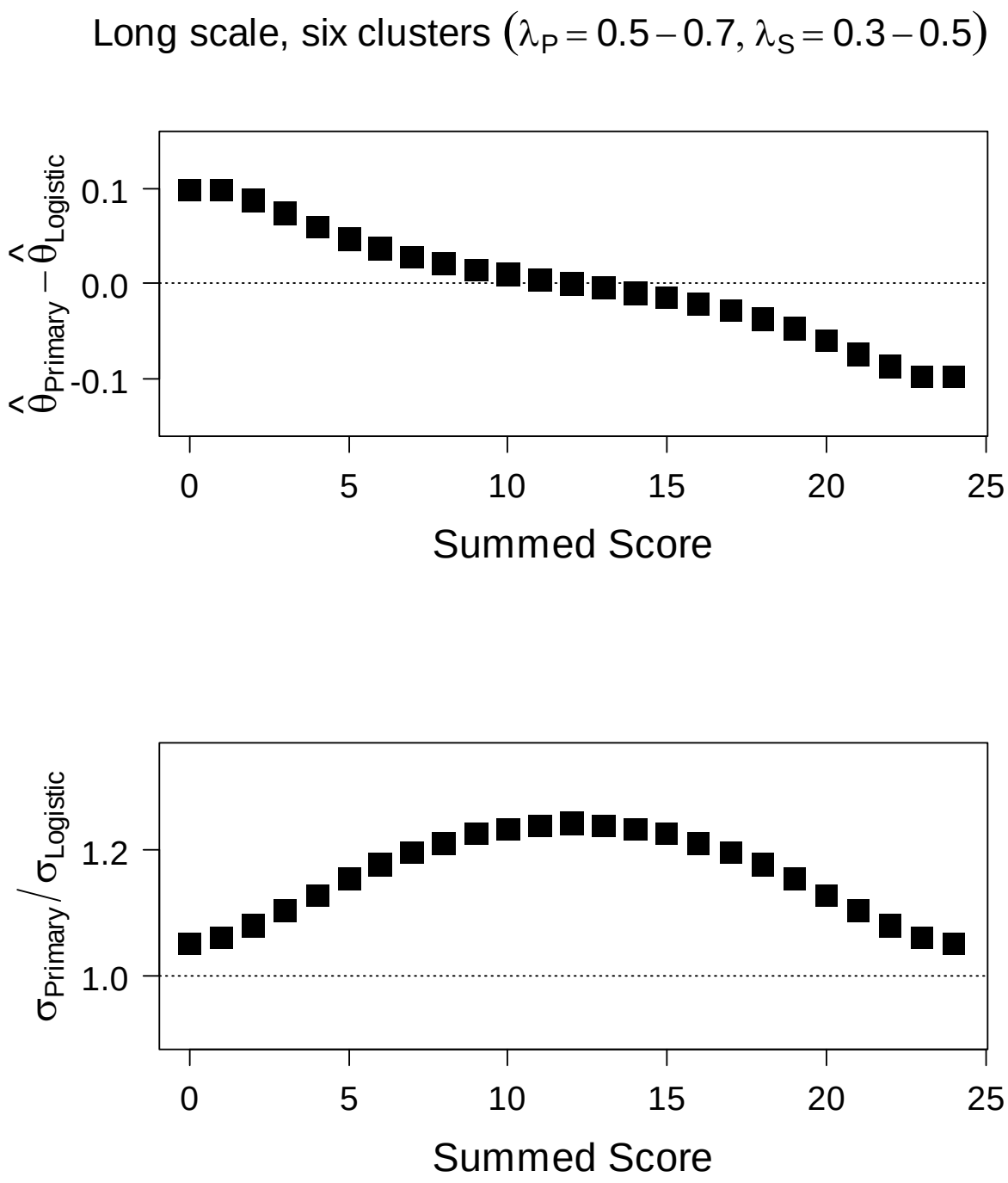

Long scale, six clusters ($\lambda_P = 0.5 - 0.7, \lambda_S = 0.3 - 0.5$)
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
0.1
0.0
-0.1
0
5
10
15
20
25
Summed Score
$\sigma_{Primary} / \sigma_{Logistic}$
1.2
1.0
0
5
10
15
20
25
Summed Score

Figure10.

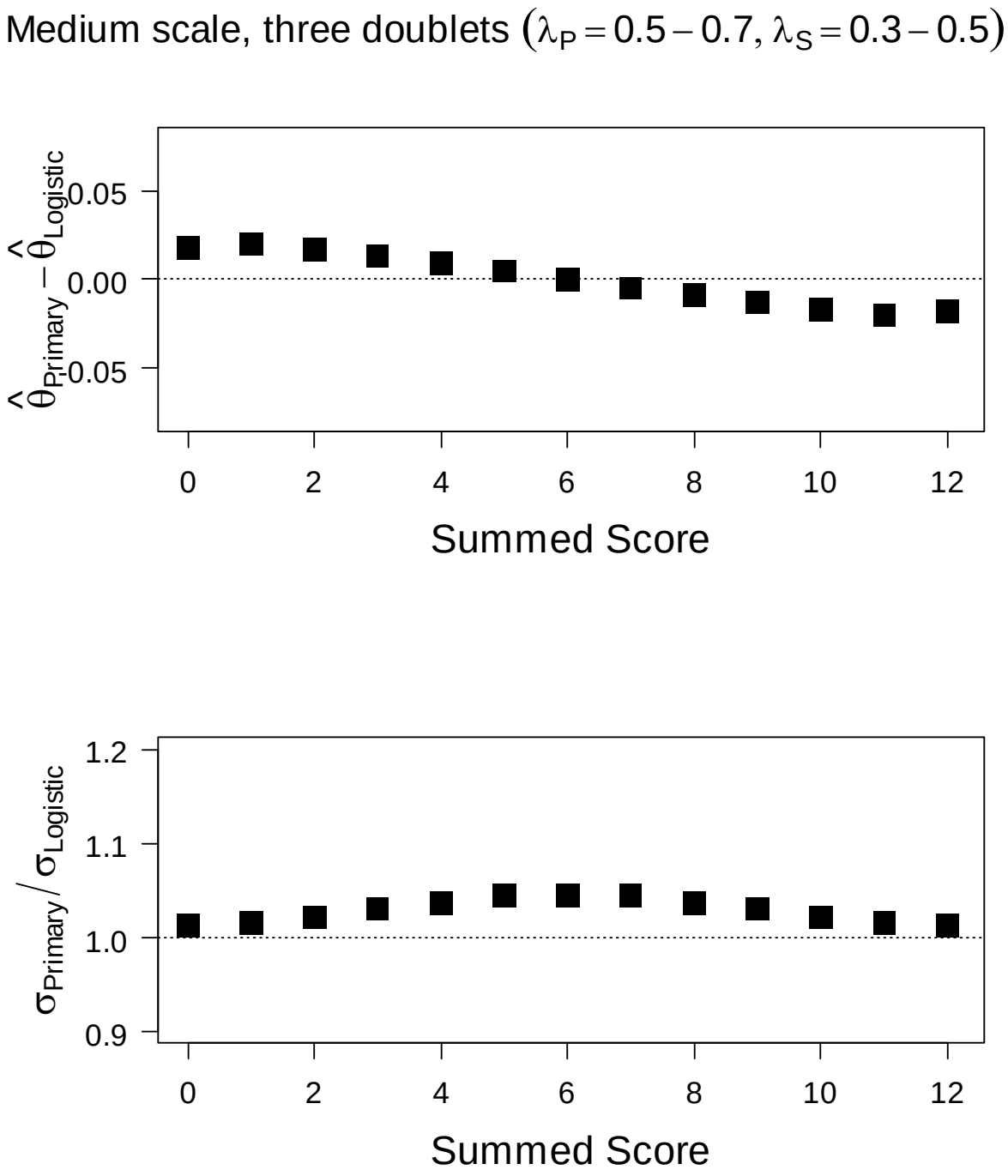
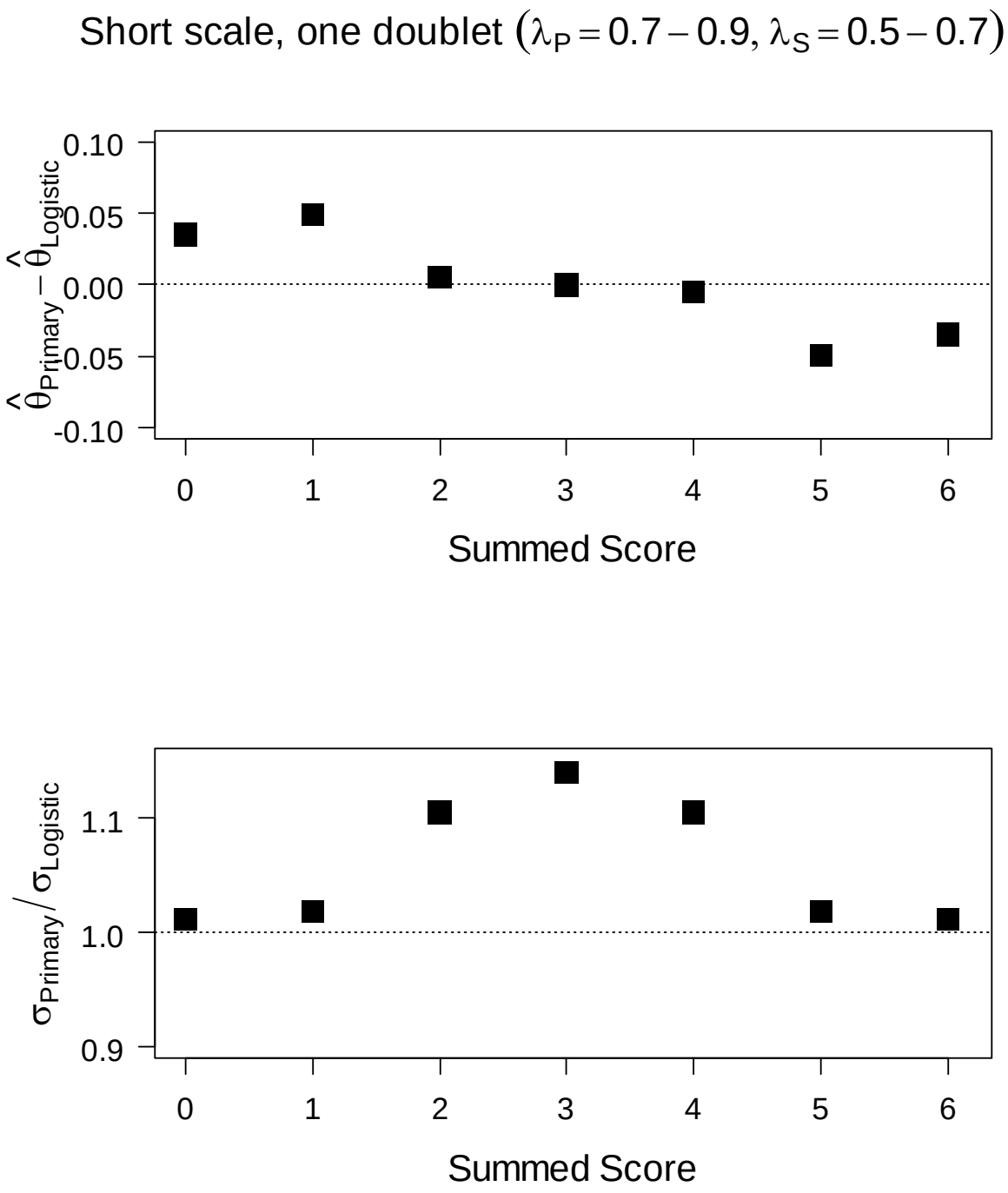

Medium scale, three doublets $(\lambda_P = 0.5 - 0.7, \lambda_S = 0.3 - 0.5)$
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
0.05
0.00
−0.05
0
2
4
6
8
10
12
Summed Score
$\sigma_{Primary} / \sigma_{Logistic}$
1.2
1.1
1.0
0.9
0
2
4
6
8
10
12
Summed Score

Figure 11.


Short scale, one doublet ($\lambda_P = 0.7 - 0.9, \lambda_S = 0.5 - 0.7$)
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
0.10
0.05
0.00
-0.05
-0.10
0
1
2
3
4
5
6
Summed Score
$\sigma_{Primary} / \sigma_{Logistic}$
1.1
1.0
0.9
0
1
2
3
4
5
6
Summed Score

The figures illustrate the difference between primary dimension EAPs and SDs between the two-tier algorithm and the logistic approximation of the marginal trace line. The upper panel of each figure indicates the differences in EAPs between the two scoring methods. Because the thresholds are fixed at zero, the mean IRT score is always 0.0 and is located in the middle of the summed score scale. For instance, in Figure 9 the mean summed score is 12 (out of 24) and is associated with an EAP of 0.0. EAPs for summed scores greater than 12 are positive values, and EAPs less than 12 are negative values. In all figures, when the difference between two-tier and logistic approximation EAPs is positive for summed scores below the mean, the difference indicates that the logistic approximation is providing an EAP which is lower than the EAP for the two-tier algorithm.

The lower panel of each figure indicates the ratio in score standard deviations of the two-tier algorithm to the logistic approximation. Because all ratios are greater than 1.0, each of the three examples indicates that use of the logistic approximation provides overly precise score estimates (i.e., overly narrow posteriors).

The pattern of results is also similar across the three examples. That is, the logistic approximations of marginal trace lines produce more extreme scores and scores with more apparent information. While there appear to be differences in the magnitude of these deviations, the patterns remain consistent. The logistic approximation provides overly precise scores near the mean of the latent variable, and performs somewhat more expectedly near the tails of the distribution. The EAPs for scaled scores at the mean are exactly the same as the two-tier algorithm suggests, indicating that the mean summed score is the middle value and corresponds with an EAP of zero. However, differences from the mean are symmetrical such that the logistic approximation-based EAPs suggest scores that are farther

from the mean of the latent variable than the two-tier algorithm provides (i.e., more extreme scores).

Controlling Local Dependence

Differences between scaled scores and standard deviations in Figures 9-11 indicate the same problem: More weight or information is being provided by the logistic approximations of marginal trace lines. These findings may be best explained by a failure to account for local dependence. In other words, while the marginal slope does provide the relation between the item response and the primary dimension after controlling for the secondary dimension(s) for each given item, it does not take into account the other items loading on the same dimension. Each item's marginal slope is the correct relation between the item response and the primary dimension for the individual item, but pooling multiple items from the same secondary dimension is still in essence providing more items of the type that are described by the secondary dimension, which is a violation of local independence when scored with a unidimensional model.

This phenomenon may be explained through an example. Suppose responses to a set of items were best characterized by a bifactor model in which *negative affect* is the primary dimension, but with the secondary factors *anger*, *anxiety*, and *depression*. If a researcher using only the anger subdomain items was interested in a score for the general factor, negative affect, then (as was proposed) computing the marginal trace lines for the negative affect dimension (integrating over the anger secondary factor) should provide the parameters to be used in IRT-based scoring. However, this suggestion essentially ignores the items' multidimensionality and treats them as if they were a unidimensional set measuring negative

affect. Put more simply, one should not administer a set of anger items under the assumption that the scores measure only negative affect.

However, for each item separately, the logistic approximation of the marginal trace line does provide the correct relation between the item response and the dimension of interest. So, if each item is considered separately, the corresponding scale scores from the two-tier algorithm or logistic approximation approaches should be nearly identical. As a brief illustration, consider a single binary, multidimensional item with a primary dimension slope of 2.0, a secondary slope of 2.0, and a corresponding threshold of 0.0. From Table A1, the marginal slope is 1.30 for the primary dimension (or secondary dimension). If this single item were administered and scored separately using both techniques, the EAPs for correct responses using the two-tier, multidimensional model and the logistic approximation of the marginal trace line are 0.488 and 0.480 (a difference of 0.008 standard deviations), respectively. The score standard deviations between the two methods are also noticeably close (0.877 and 0.873, respectively). This example illustrates that individual items selected from a multidimensional cluster of items can be appropriately scored using the logistic approximation of the marginal trace line, which is equivalent to achieving local independence among a set of locally dependent items. If the previous example is extended to a scale of items with bifactor structure, selecting a single item, regardless of which item, from each LD cluster and proceeding with the logistic approximation method of scoring should provide scaled scores which reflect those from the two-tier algorithm.

This method is actually similar to traditional approaches of controlling for local dependence. It is common in practice to model the relations among a set of tryout items with a bifactor model (Hill, et al., 2007). After identifying violations of unidimensionality (in the

form of doublets or clusters of locally dependent items loading on a secondary dimensions), a single item is retained for the final scale and all other locally dependent items are set aside. Removing an item's locally dependent partners eliminates the shared relation above and beyond the primary dimension. This process results in a unidimensional scale with a subset of the initially multidimensional items.

It is now proposed that the logistic approximation of the marginal trace line will produce scores similar to those from the two-tier algorithm if a single item is selected from each cluster of locally dependent items. To test this revised hypothesis, the scoring design previously proposed is amended to include scoring runs which control for local dependence. The additional forty-eight scoring runs which select only a single item from each locally dependent cluster of items are referred to as using only *locally independent* items. The original forty-eight conditions are referred to as using *all items*. All forty-eight conditions are repeated to allow for comparisons between the *all items* set of scores, which violate local independence when using the logistic approximation, and the *locally independent* condition.

To select a single item from each locally dependent subset of items some decisions are required. In practice, it is best to select the item which most reflects the dimension of interest, along with other substantive concerns. Because the conditions currently assessed are meant to be an evaluation of the method, the item selected from each locally dependent cluster of items tended to be an item in the middle of cluster (i.e., an item that is neither extreme on the primary dimension, nor extreme on the secondary dimension).

Additionally, selecting a single item from each cluster reduces the number of items being scored. The reduction in the number of scores depends on the test length and dimensionality conditions, but can range from just a single item (e.g., a 6-item scale with one locally

dependent doublet, which is reduced to 5 items which are scored) to the majority of the original scale (e.g., the 24-item, 6-cluster condition is reduced to 6 scored items).

Scoring Results

All 48 conditions were scored using *all item* and *locally independent* items only. Prior to comparing results for all conditions, illustrations from a few selected examples are provided. Figures 12-17 provide differences in EAPs and SDs between the two-tier method and the logistic approximation. The top panels of each figure plot bias in the EAPs by taking the difference between the two-tier algorithm and the logistic approximation. Deviations from zero indicate scaled score differences between the two-tier algorithm and the logistic approximation. The bottom panels plot the ratio between the score standard deviations for the two-tier algorithm and logistic approximation. Positive SD ratios indicate overly precise scores for the logistic approximation. The left panels indicate the bias in scores and SDs when all items are used and locally dependence is ignored. The right panels indicate the correction in scores and SDs when a single item is selected from each locally dependent cluster of items and local independence is achieved.

Recall that the original Figures 9-11 detected bias in the logistic approximation when ignoring local dependence. Figures 12-14 plot the corresponding correction in scores when only locally dependent items are scored (i.e., the left panels of Figures 12-14 are equivalent to Figures 9-11, respectively, and the right panels of Figures 12-14 illustrate the correction in scores and SDs when local dependence is controlled). Figures 15-17 are notable results from three of the remaining scoring conditions. Figure 15 is illustrative of medium length scales (12 items), which are primarily unidimensional but with three pairs of locally dependent items. Figures 16 and 17 represent long scales (24 items) which have complete bifactor

structure. Figure 16 has a commonly seen factor structure with large loadings on the primary dimension and medium loadings on the secondary conditions. Figure 17 is the most extreme example from the forty-eight conditions and represents a scale with low loadings on the primary dimension and high loadings on the secondary dimensions.

Figure 12.

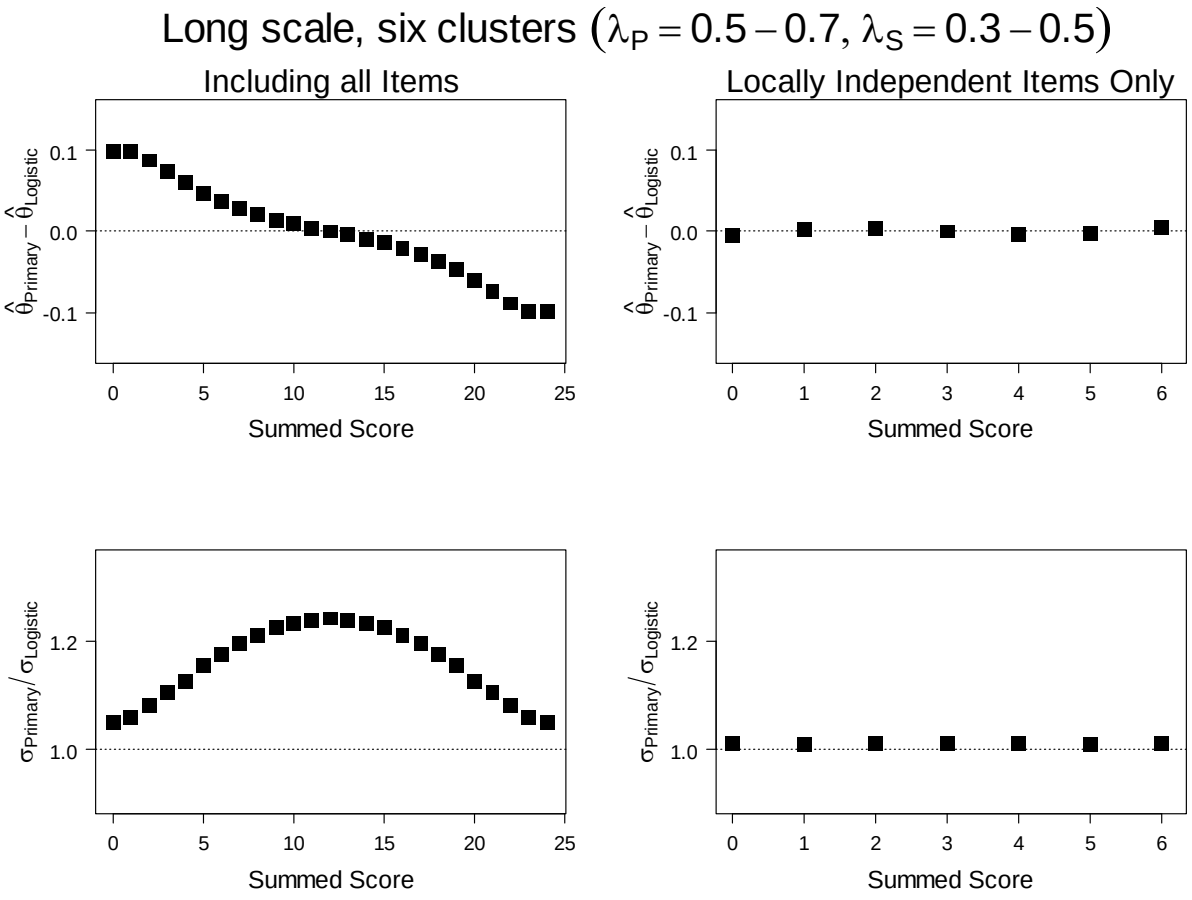

Long scale, six clusters ($\lambda_P = 0.5 - 0.7, \lambda_S = 0.3 - 0.5$)
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary}/\sigma_{Logistic}$
Summed Score

Figure 13.

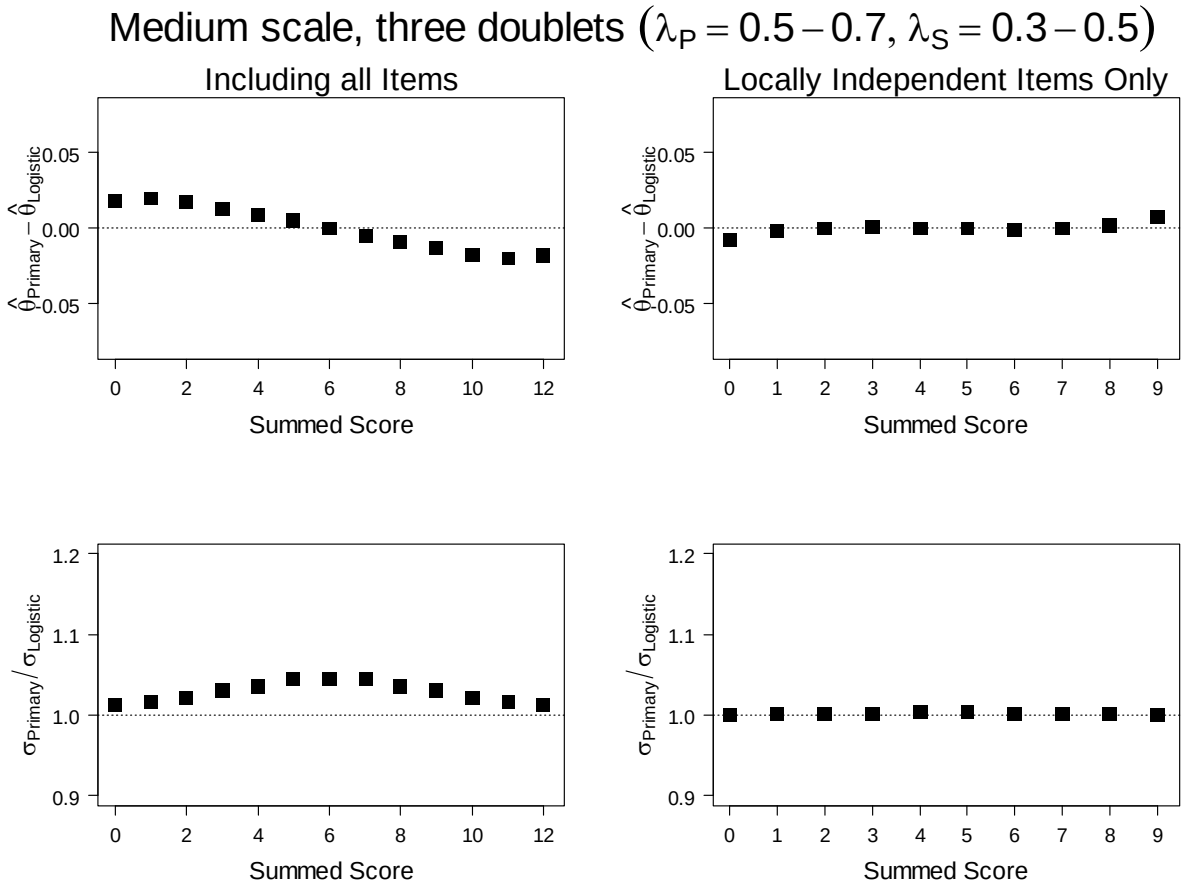

Medium scale, three doublets ($\lambda_P = 0.5 - 0.7$, $\lambda_S = 0.3 - 0.5$)
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary} / \sigma_{Logistic}$
Summed Score

Figure 14.

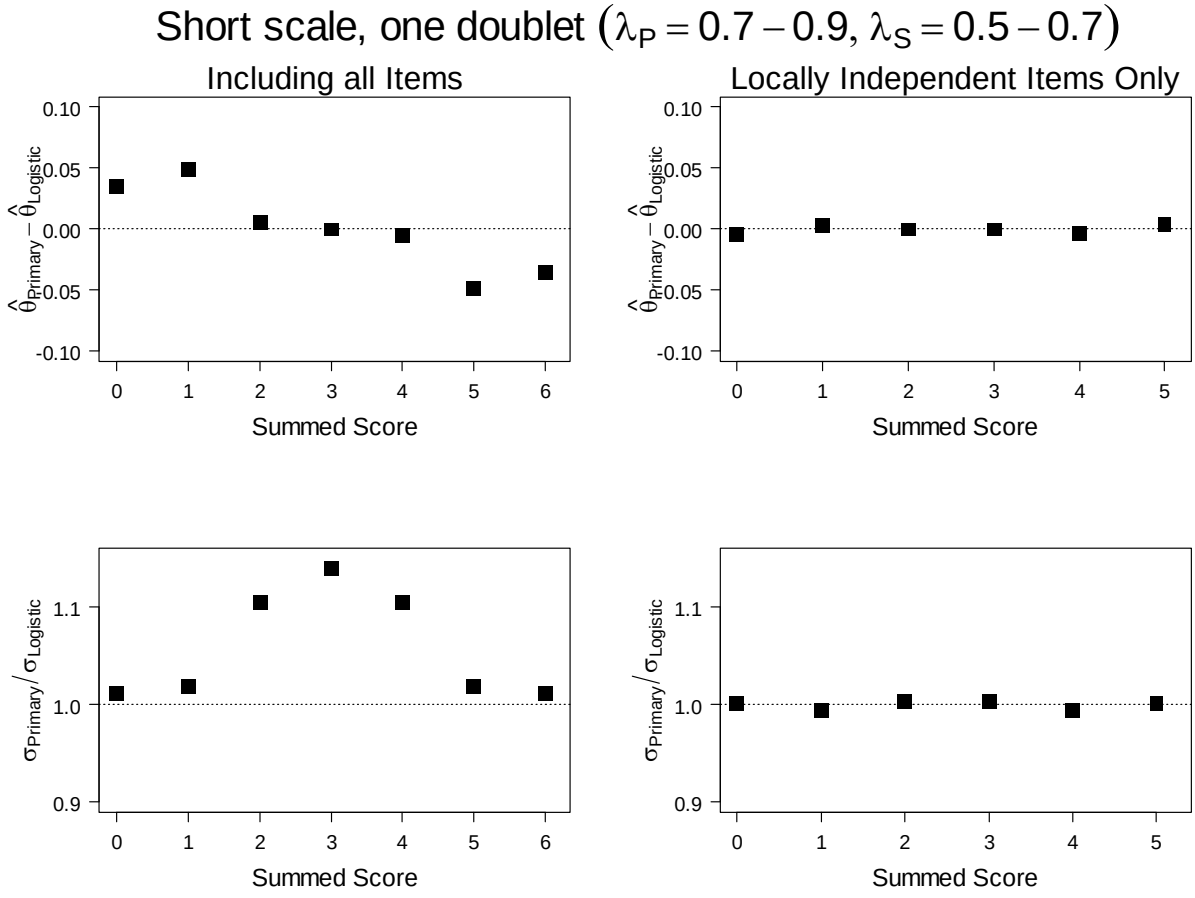

Short scale, one doublet ($\lambda_P = 0.7 - 0.9, \lambda_S = 0.5 - 0.7$)
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary} / \sigma_{Logistic}$
Summed Score
0.10
0.05
0.00
-0.05
-0.10
1.1
1.0
0.9
0
1
2
3
4
5
6

Figure 15.

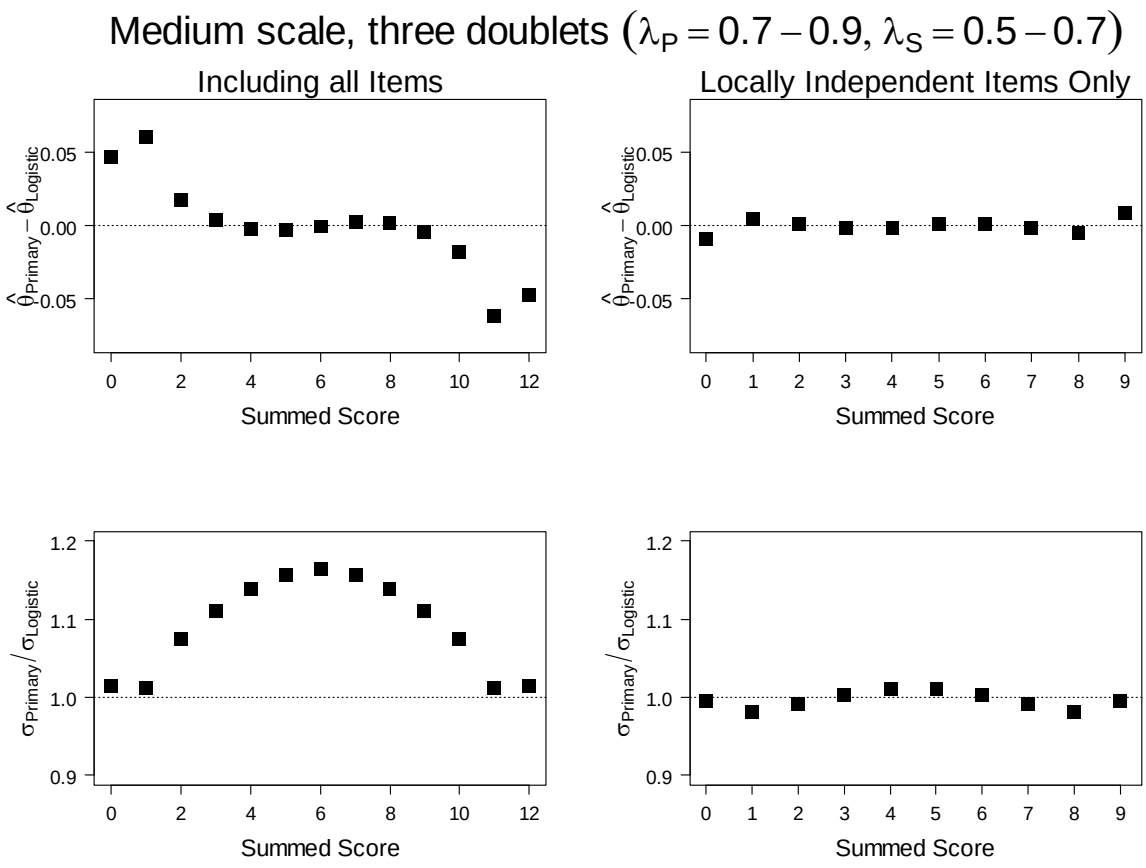

Medium scale, three doublets ($\lambda_P = 0.7 - 0.9, \lambda_S = 0.5 - 0.7$)
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary} / \sigma_{Logistic}$
Summed Score

Figure 16.

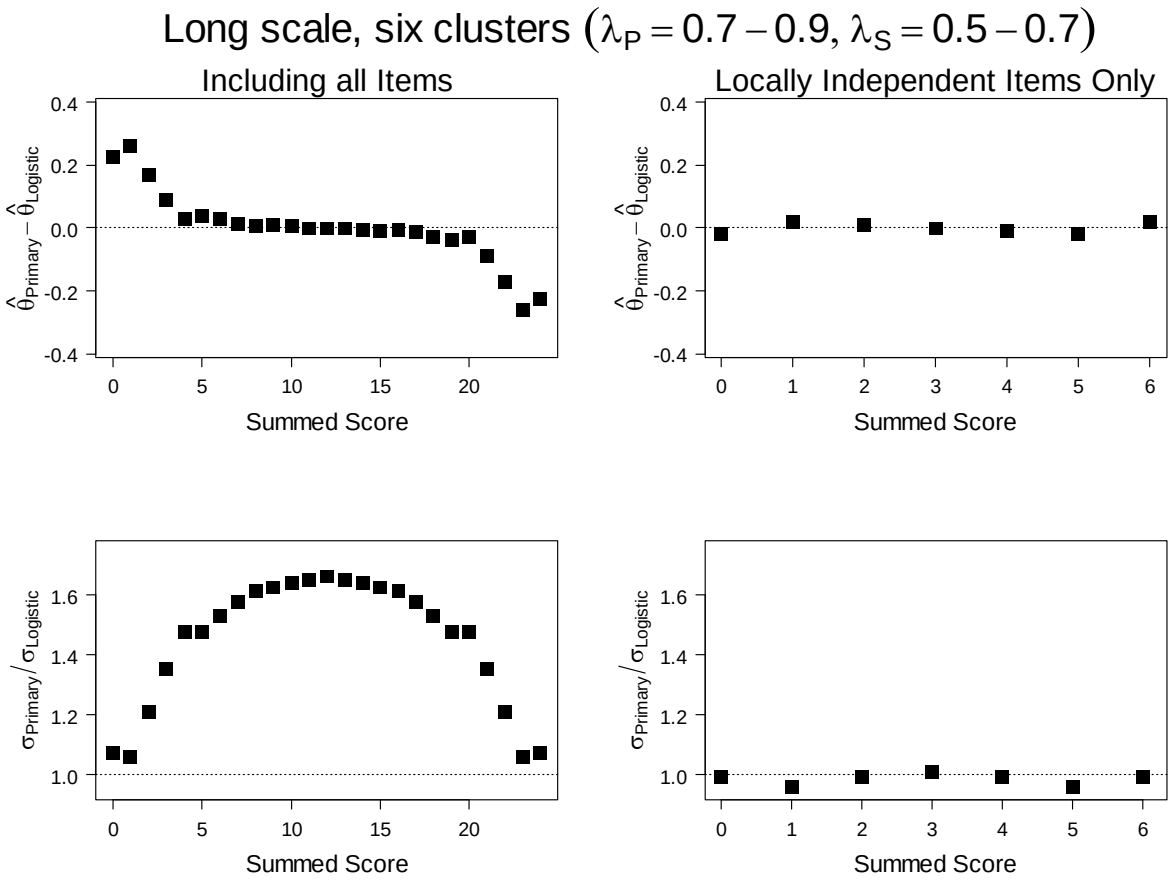
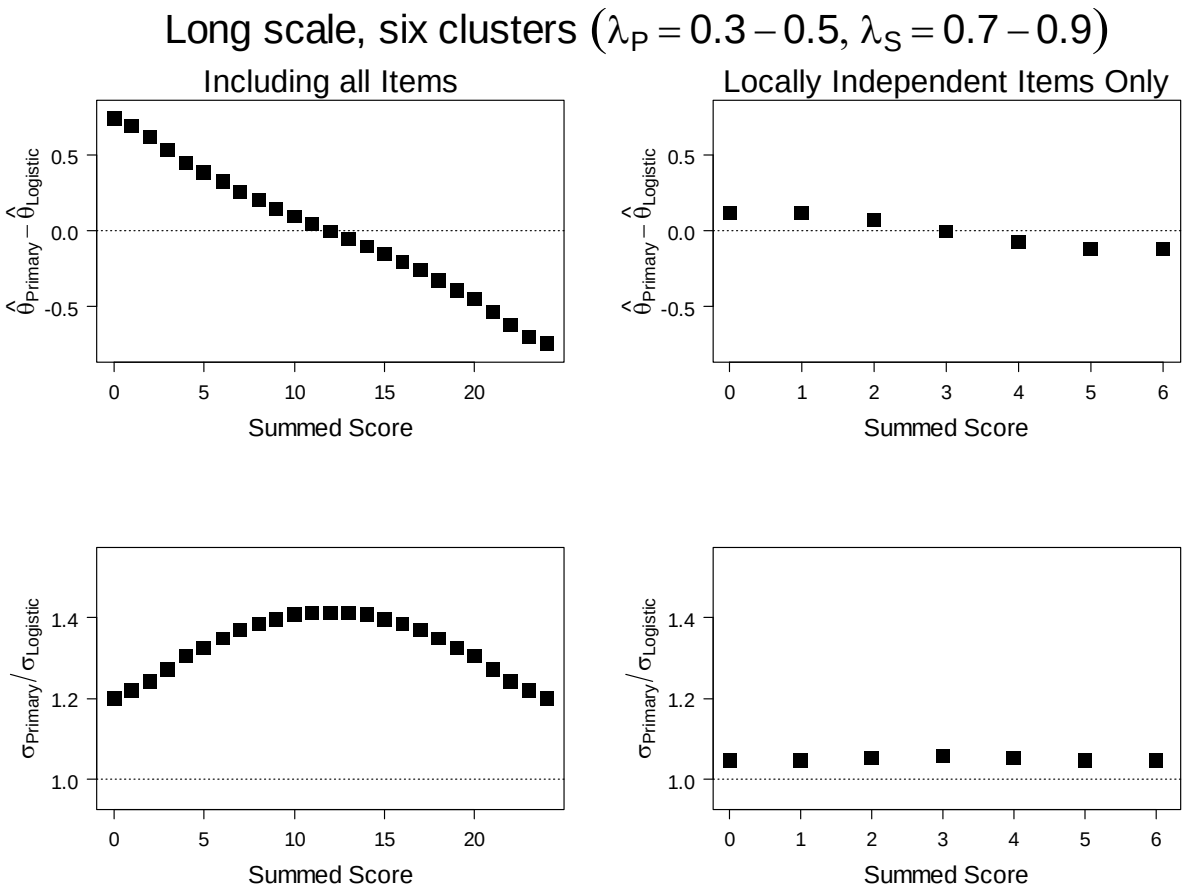

Long scale, six clusters ($\lambda_P = 0.7 - 0.9, \lambda_S = 0.5 - 0.7$)
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary} / \sigma_{Logistic}$
Summed Score

Figure 17.


Long scale, six clusters $(\lambda_P = 0.3 - 0.5, \lambda_S = 0.7 - 0.9)$
Including all Items
Locally Independent Items Only
$\hat{\theta}_{Primary} - \hat{\theta}_{Logistic}$
$\sigma_{Primary} / \sigma_{Logistic}$
Summed Score

Figures 12-14 illustrate the bias in scoring when using all items, ignoring local dependence, and the correction made by selecting locally independent subsets of items. As expected, ignoring local dependence can create large biases in scores and standard deviations when using the logistic approximation of the marginal trace line. In Figure 12 the maximum difference in scaled scores when using all items is ±0.099, which is reduced to ±0.005 when using locally independent subsets of items. In Figure 13, because the secondary factor is weak, and only three pairs of the 12 items have secondary factors, little is lost by ignoring LD or gained by controlling for it. However, the correction is greater in Figure 14 which contains only one doublet from a 6-item scale, but results in a maximum spurious increase in score precision of 12.2% for the logistic approximation method when ignoring local dependence (which is corrected to -0.6%, when using locally independent items).

Figures 15-17 illustrate some interesting examples that reflect common multidimensional models. Figure 15 illustrates a largely unidimesional model with three doublets. Ignoring unidimensionality does not affect scaled scores to a large degree; however, score precision can be over estimated by up to 14%. Figure 16 represents a traditional bifactor model with high loadings on the primary dimension and medium loadings on the secondary dimensions. In this case, the effects of ignoring local dependence are widespread as scaled scores may be overestimated by 0.26 standard deviation units with a spurious increase in score precision of nearly 40%. Finally, Figure 17 is an extreme example of a long scale with low loadings on the primary dimension and high factor loadings on the secondary dimensions. Ignoring local dependence in this situation can result in the scale score differences of up to 0.74 standard deviations and score precision overestimation by nearly 30%.

Summary of Findings: Ignoring Local Dependence

EAP and SD estimates were computed using the two-tier algorithm and the logistic approximation of the marginal trace line for all forty-eight scoring conditions. For each condition, across the range of summed scores, the maximum deviation between EAPs and SD estimates is presented in Tables 7 and 8, respectively. This approach is akin to selecting the most egregious case from each scoring run. This presentation of findings indicates the bias in scores and score precision when ignoring local dependence, and after selecting locally independent subsets of items the proximity between primary dimension scores computed from either the two-tier algorithm or the logistic approximation.

Tables 7 and 8 suggest that bias in logistic approximation-based scores is not the product of a single factor. The current results indicate (1) the deleterious effect of many instances of multidimensionality (i.e., local dependence), (2) the relative strength of the local dependence in relation to the strength of the primary dimension, and (3) the total number of items on the scale, which may either protect against, or amplify the effects of ignored local dependence.

As an overview of Tables 7 and 8, it appears that ignored local dependence results in the most severe biases for scales with a weak primary dimension but strong secondary dimensions (the second column from the right in Tables 7 and 8) and many items on each secondary dimension (the last few rows in Tables 7 and 8). Indeed, scales with such covariance structure may indicate the appropriateness of multiple unidimensional scales rather than a bifactor model. However, the number of items present on the scale may serve as a protective factor against local dependence overwhelming the orientation of the latent variable. Consider the second and third columns from the left (which indicate strong primary dimensions relative to the secondary dimensions) and the first three rows (where there are

fewer instances of local dependence); these conditions illustrate Harrison's (1986) observation that effects of local dependence may be dampened by a strong primary dimension with many items relative to fewer subsets of LD items.

Summary of Findings: Controlling for Local Dependence

Obtaining unidimensionality by setting aside items from clusters of locally dependent items greatly reduces the bias in the logistic approximation-based scores. The correction in scores and score precision by selecting locally independent items is presented in three parts: short, medium, and long test lengths. First, considering the short test length with models including either one or two doublets, ignoring local dependence resulted in a maximum EAP deviation of 0.200 standard deviations and a 17.0% over-estimation of score precision (for the "low" primary and "high" secondary loading conditions, and the "high" primary and "medium" secondary conditions, respectively). In both cases, selecting locally dependent subsets of items resulted in a maximum bias in EAPs of ±0.007 standard deviations and a maximum over-estimation of score precision of 1.5%, which for most practical purposes reflects nearly identical scores and standard deviations between the two-tier and logistic approximation of marginal trace lines.

Results for the medium test length condition with three pairs of doublet items are consistent with those from the short test length conditions; however, ignoring local dependence in the bifactor structure condition resulted in maximum score bias of 0.724 standard deviations, and an inflation in score precision of up to 30.7 percent (for the "low/high" and "medium/high" primary and secondary factor loading conditions, respectively). For these cases, setting aside locally dependent items resulted in scores and

score standard deviations much closer to the two-tier findings (i.e., a ±0.019 standard deviation difference in EAPs and only a 1.9% spurious increase in precision).

Finally, the long test length condition with complete bifactor structure provided the most challenging set of results for the logistic approximation. Ignoring local dependence resulted in a maximum EAP difference of ±0.741 standard deviations and a 39.8% maximum overestimation of score precision (for the "low/high" and "medium/high" primary and secondary factor loading conditions, respectively). Obtaining a locally independent subset of items reduced the maximum EAP difference to ±0.120 and the score precision to -4.0% (an underestimation of score precision). While these results are the most extreme example of the logistic approximation deviating from the two-tier algorithm, it is unlikely in practice that a bifactor model would be useful in the presence of a weak primary dimension and six strong and distinct secondary dimensions.

Table 7. Maximum difference in EAPs between tests scored with the two-tier algorithm and the logistic approximation of the marginal trace line.[†]

Test Length (Dimensionality) ↓	Primary Loadings → Secondary Loadings →	Low Low	Medium Low	High Low	Low Medium
Short (6 items) (1 doublet pair)	All items	0.028	0.018	0.018	0.056
	Locally Independent	0.001	0.002	0.003	0.002
Short (6 items) (2 doublet pairs)	All items	0.055	0.038	0.038	0.067
	Locally Independent	0.004	0.003	0.005	0.007
Medium (12 items) (3 doublet pairs)	All items	0.035	0.020	0.021	0.082
	Locally Independent	0.003	0.007	0.005	0.004
Medium (12 items) (3 clusters with 4 items each)	All items	0.224	0.138	0.118	0.455
	Locally Independent	0.009	0.005	0.011	0.016
Long (24 items) (6 doublet pairs)	All items	0.025	0.012	0.020	0.066
	Locally Independent	0.008	0.007	0.004	0.014
Long (24 items) (6 clusters with 4 items each)	All items	0.187	0.099	0.105	0.422
	Locally Independent	0.005	0.005	0.013	0.009

(Continued)

Test Length (Dimensionality) ↓	Primary Loadings → Secondary Loadings →	Medium Medium	High Medium	Low High	Medium High
Short (6 items) (1 doublet pair)	All items	0.043	0.049	0.093	0.066
	Locally Independent	0.003	0.004	0.003	0.003
Short (6 items) (2 doublet pairs)	All items	0.091	0.105	0.200	0.174
	Locally Independent	0.004	0.007	0.007	0.003
Medium (12 items) (3 doublet pairs)	All items	0.055	0.061	0.150	0.087
	Locally Independent	0.011	0.009	0.002	0.012
Medium (12 items) (3 clusters with 4 items each)	All items	0.329	0.111	0.724	0.459
	Locally Independent	0.008	0.011	0.019	0.009
Long (24 items) (6 doublet pairs)	All items	0.042	0.038	0.131	0.072
	Locally Independent	0.015	0.005	0.018	0.019
Long (24 items) (6 clusters with 4 items each)	All items	0.271	0.262	0.741	0.482
	Locally Independent	0.018	0.020	0.120	0.017

Note: Because deviations in scores are symmetrical around the mean, each value in the table has a corresponding negative value.

† The values displayed indicate the maximum difference in scores across the range of scale scores from summed scores.

Table 8. Maximum percentage difference in score precision between tests scored with the two-tier algorithm and the logistic approximation of the marginal trace line. †

Test Length (Dimensionality) ↓	Primary Loadings → / Secondary Loadings →	Low / Low	Medium / Low	High / Low	Low / Medium
Short (6 items) (1 doublet pair)	All items	1.2	2.3	5.4	2.6
	Locally Independent	0.1	-0.3	0.4	0.4
Short (6 items) (2 doublet pairs)	All items	2.2	4.0	7.4	4.1
	Locally Independent	0.5	0.4	0.4	0.8
Medium (12 items) (3 doublet pairs)	All items	2.5	4.3	8.4	5.2
	Locally Independent	0.4	0.4	0.9	0.6
Medium (12 items) (3 clusters with 4 items each)	All items	10.0	16.6	27.8	16.8
	Locally Independent	0.8	1.0	-1.6	1.4
Long (24 items) (6 doublet pairs)	All items	3.4	5.3	9.6	6.5
	Locally Independent	0.5	0.8	1.8	0.9
Long (24 items) (6 clusters with 4 items each)	All items	13.1	19.5	31.0	22.1
	Locally Independent	1.0	1.1	2.0	1.8

(Continued)

Test Length (Dimensionality) ↓	Primary Loadings → Secondary Loadings →	Medium Medium	High Medium	Low High	Medium High
Short (6 items) (1 doublet pair)	All items	5.3	12.2	4.7	7.6
	Locally Independent	0.3	-0.6	0.5	0.3
Short (6 items) (2 doublet pairs)	All items	8.1	17.0	7.1	10.8
	Locally Independent	0.7	1.5	1.0	1.1
Medium (12 items) (3 doublet pairs)	All items	9.0	14.1	7.9	11.2
	Locally Independent	1.2	-1.8	0.8	1.0
Medium (12 items) (3 clusters with 4 items each)	All items	26.9	26.2	22.2	30.7
	Locally Independent	1.7	-1.6	1.6	1.8
Long (24 items) (6 doublet pairs)	All items	9.8	13.7	10.3	12.5
	Locally Independent	1.4	2.8	1.3	1.6
Long (24 items) (6 clusters with 4 items each)	All items	31.2	39.8	29.2	36.3
	Locally Independent	1.4	-4.0	5.5	2.0

Note: Positive values indicate the logistic approximation's tendency to over-estimate of score precision.

† The values displayed indicate the maximum difference in scores across the range of scale scores from summed scores. For each score comparison, the percentage difference is $\left[(\sigma_{\text{Two-Tier}} - \sigma_{\text{Logistic}})/\sigma_{\text{Two-Tier}}\right] * 100\%$.

CHAPTER 4

AN APPLICATION OF MARGINAL TRACE LINES FOR BIFACTOR ITEM RESPONSE THEORY MODELS

Thus far, logistic approximations of marginal trace lines have only been considered for 2PL models with known item parameters. To further examine the utility of the logistic approximation, a real-data application is now considered which makes use of the 5-category graded response model (GRM; Samejima, 1969) and data from one of the pediatric Patient Reported Outcomes Measurement Information System (PROMIS) scales. The PROMIS pediatric network involves researchers whose goal is to develop item banks across several general health domains (e.g., physical function, pain, fatigue, emotional distress, social function, and one disease-specific scale for asthma) for youth ages 8–17 years (e.g., Irwin, et al., 2010; Yeatts, et al., 2010; Varni, et al., in press). In developing these health outcome measures, PROMIS scale construction and item assembly methodology used bifactor models to account for nuisance dimensionality and local dependence, and IRT models to calibrate unidimensional subsets of items.

The present example uses data from the 33 tryout items of the PROMIS Asthma Symptoms health outcomes domain previously analyzed by Yeatts et al. (2010) [7]. Participants were recruited from hospital clinics and public schools in Texas and North Carolina and included 622 children ages 8-17 (55% Male, 46% White). The researchers' original item factor analytic model identified one primary dimension and seven nuisance

[7] The initial set of tryout items for the asthma symptoms domain contained an item which cross-loaded on three dimensions, and is excluded from the present set of analyses.

dimensions. The high-degree of multi-dimensionality was due to the item development process. The items were generated primarily from existing asthma symptoms scales, which had a moderate degree of content overlap; subfactors and doublets were needed to account for redundancy in content. After substantive review, one item was selected from each doublet or subfactor to remain on the scale resulting in the calibration of an 18-item unidimensional model (an additional item was set aside following calibration yielding the 17-item Pediatric Asthma Impact Scale (PAIS; Yeatts et al., 2010)).

We now re-visit and re-estimate the models described by Yeatts et al. (2010) to illustrate the utility of logistic approximations of marginal trace lines. Specifically, the techniques discussed here allow the original 33-item bifactor model to have marginal trace lines for all items with slope parameters on more than one dimension. Following guidelines used by Yeatts et al (2010), items are set aside for local dependence and the remaining unidimensional subset of 18-items is calibrated. This process allows for the comparison of item parameters, EAPs, and posterior standard deviations between the logistic approximations of marginal trace lines and the univariate IRT model with the same final set of items.

Re-evaluating an Asthma Symptoms Scale

The final factor analytic model reported by Yeatts, et al (2010) was used to guide the current analyses. Specifically, Yeatts et al. report 34 items loading on one primary dimension with five doublets (e.g., asthma attacks, trouble sleeping, hospital/emergency room visits, etc.), and two secondary factors best described as being "scared or worried" by having asthma and difficulties with "sports or exercise" due to asthma. The present analyses fit this same model to the original asthma symptoms items using the software program

IRTPRO and the two-tier algorithm for item parameter estimation (an item reported by Yeatts et al. that cross-loaded on three dimensions was set aside).

Table 9 provides the slope parameters for this 8-dimensional model. A few examples of conditional slopes may illustrate the challenge in using bifactor models in scale development scenarios. Consider the items "I had trouble breathing because of my asthma" and "I felt out of breath because of my asthma." Both items had relatively strong primary-dimension slopes (3.18 and 2.75, respectively) and weak, though significant secondary dimension slopes (1.13, each). This information alone may make for a challenging interpretation of the relation between item responses and the primary dimension; however, after computing the marginal trace lines and logistic approximations (see Table 9 for conditional and marginal slopes) it is clear that this weak secondary factor has little influence on the marginal slopes for the primary dimensions (marginal slopes = 2.65 and 2.29, respectively). Use of the marginal slope parameters in this case suggests that either item would be useful in a final scale for uncovering information regarding the latent variable.

While the previous item pair was indicative of relatively weak local dependence, the item pair "I went to the hospital for my asthma" and "I went to the emergency room for my asthma" had evidence of strong local dependence (secondary slopes = 3.31, each), though each is only moderately related to the primary dimension (primary slopes = 2.00 and 1.96). However, integrating over the secondary dimension illustrates how misleading conditional slopes may be as the marginal trace lines for both items are weak and indicate a relatively weak relationship with the primary dimension (marginal slopes = 0.91 and 0.90, respectively). With knowledge of the marginal slope parameter, there may be little utility in including either item in a scale measuring asthma symptoms.

Table 9. A comparison of conditional and marginal slope parameters for 33 asthma symptoms items.

Item	$a_{marginal}$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
I was scared at night because of my asthma.	1.85	2.20	1.09	---	---	---	---	---	---
I felt scared that I might have trouble breathing because of my asthma.	1.90	2.09	0.78	---	---	---	---	---	---
I worried I would have an asthma attack.	1.52	1.95	1.36	---	---	---	---	---	---
I was scared that I might have to go to the emergency room or hospital because of my asthma.	1.53	1.86	1.17	---	---	---	---	---	---
I worried that other people would not know what to do if I had an asthma attack.	1.11	1.48	1.51	---	---	---	---	---	---
It was hard for me to play sports or exercise because of my asthma.	1.91	2.71	---	1.72	---	---	---	---	---
I had trouble playing with other kids because of my asthma.	2.06	2.27	---	0.79	---	---	---	---	---
It was hard for me to play outside because of my asthma.	1.87	2.19	---	1.03	---	---	---	---	---
I limited my activities because of asthma.	1.70	2.15	---	1.32	---	---	---	---	---
I was unable to take part in active sports, like running because of my asthma.	1.41	1.80	---	1.34	---	---	---	---	---
I felt short of breath when I did active sports because of my asthma.	1.55	1.70	---	0.77	---	---	---	---	---
Asthma attacks bothered me.	1.87	2.55	---	---	1.57	---	---	---	---
I had asthma attacks.	1.62	2.20	---	---	1.57	---	---	---	---
I had trouble breathing because of my asthma.	2.65	3.18	---	---	---	1.13	---	---	---
I felt out of breath because of my asthma.	2.29	2.75	---	---	---	1.13	---	---	---
I had trouble walking because of my asthma.	2.01	2.64	---	---	---	---	1.44	---	---
I had trouble talking because of my asthma.	1.61	2.11	---	---	---	---	1.44	---	---
I went to the hospital for my asthma.	0.91	2.00	---	---	---	---	---	3.31	---
I went to the emergency room for my asthma.	0.90	1.96	---	---	---	---	---	3.31	---
I had trouble sleeping at night because of my asthma.	2.01	3.01	---	---	---	---	---	---	1.90
I woke up because of my asthma.	1.77	2.66	---	---	---	---	---	---	1.90

(Continued)

Item	$a_{marginal}$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
My asthma bothered me.	---	2.46	---	---	---	---	---	---	---
I felt wheezy because of my asthma.	---	2.13	---	---	---	---	---	---	---
It was hard to take a deep breath because of my asthma.	---	2.10	---	---	---	---	---	---	---
My chest felt tight because of my asthma.	---	1.87	---	---	---	---	---	---	---
My asthma was really bad.	---	1.72	---	---	---	---	---	---	---
I was bothered by the amount of time I spent wheezing.	---	1.86	---	---	---	---	---	---	---
My body felt bad when I was out of breath.	---	1.83	---	---	---	---	---	---	---
My asthma bothered me when I was with my friends.	---	1.55	---	---	---	---	---	---	---
I missed school because of asthma.	---	1.54	---	---	---	---	---	---	---
I coughed because of my asthma.	---	1.47	---	---	---	---	---	---	---
I got tired easily because of my asthma.	---	1.39	---	---	---	---	---	---	---
It was hard for me to play with pets because of my asthma.	---	1.15	---	---	---	---	---	---	---

Note: All conditional slope estimates are more than twice their standard errors

Next, following the scale construction techniques described by Yeatts et al. (2010), a unidimensional subset of items was selected using the bifactor model for guidance. Specifically, Yeatts et al. (2010) selected a single item from each secondary factor based on the strength of the primary dimension slope or based on substantive content review. Both the "hospital" and "emergency room" items were set aside. This process resulted in the unidimensional calibration of 18 remaining items.

Table 10 provides a comparison between the conditional MIRT or marginal slope parameters and the univariate IRT slope parameters. The first six entries list the marginal slope parameters for the six items selected from secondary factors. The remaining twelve items (with no secondary dimensions) have their conditional slopes listed. These marginal and conditional slope parameters may be compared to the slope parameters resulting from the unidimensional calibration of the same 18 items. If the bifactor model accounts for local dependence among subsets of items, then the marginal trace lines should closely correspond to the univariate IRT slope parameter estimates. Likewise, slopes for items modeled in the bifactor model that were only represented by a single dimension should be close to the estimates from the unidimensional IRT model.

Not surprisingly, differences between marginal or conditional slope parameters (and thresholds, see Table 11) and unidimensional IRT slopes and small. The average slope parameter value difference between marginal/conditional and univariate slopes is -0.05, suggesting that the process of computing marginal slopes and logistic approximations may re-create the slope and threshold parameters from a unidimensional calibration. However, the three items with the largest parameter estimates for the unidimensional IRT model and for the conditional/marginal model were also the three items with the largest slope

differences (-0.29 to -0.16), which may be evidence for unmodeled local dependence, though the increase in slope parameters is minor in magnitude.

Table 10. A comparison of conditional, marginal, and univariate slope parameters for 33 asthma symptoms items.

Item	$a_{marginal}$	a_1	$a_{univariate}$	Slope Difference
I had trouble breathing because of my asthma.	2.65	---	2.94	-0.29
I had trouble walking because of my asthma.	2.01	---	1.96	0.05
I had trouble sleeping at night because of my asthma.	2.01	---	2.00	0.01
It was hard for me to play sports or exercise because of my asthma.	1.91	---	2.00	-0.09
I felt scared that I might have trouble breathing because of my asthma.	1.90	---	1.93	-0.03
I had asthma attacks.	1.62	---	1.59	0.03
My asthma bothered me.	---	2.46	2.64	-0.18
I felt wheezy because of my asthma.	---	2.13	2.29	-0.16
It was hard to take a deep breath because of my asthma.	---	2.10	2.15	-0.05
My chest felt tight because of my asthma.	---	1.87	1.97	-0.10
I was bothered by the amount of time I spent wheezing.	---	1.86	1.89	-0.03
My body felt bad when I was out of breath.	---	1.83	1.86	-0.03
My asthma was really bad[†]	---	1.72	1.75	-0.03
My asthma bothered me when I was with my friends.	---	1.55	1.56	-0.01
I missed school because of asthma.	---	1.54	1.50	0.04
I coughed because of my asthma.	---	1.47	1.54	-0.07
I got tired easily because of my asthma.	---	1.39	1.42	-0.03
It was hard for me to play with pets because of my asthma.	---	1.15	1.13	0.02

[†] indicates an item set aside from the PAIS due to gender DIF.

Table 11. A comparison of marginal and univariate thresholds for the reduced 18-item scale.

	Marginal/Conditional Thresholds			
Item	b_1	b_2	b_3	b_4
I coughed because of my asthma.	-1.36	-0.53	0.60	1.44
My chest felt tight because of my asthma.	-1.07	-0.46	0.67	1.39
I had trouble breathing because of my asthma.	-0.92	-0.36	0.82	1.51
My body felt bad when I was out of breath.	-0.97	-0.33	0.83	1.52
My asthma bothered me.	-0.96	-0.33	0.84	1.52
I felt scared that I might have trouble breathing because of my asthma.	-0.82	-0.13	1.05	1.85
I got tired easily because of my asthma.	-1.04	-0.24	1.09	2.12
It was hard to take a deep breath because of my asthma.	-0.66	-0.01	0.97	1.63
I felt wheezy because of my asthma.	-0.64	0.00	1.09	1.89
It was hard for me to play sports or exercise because of my asthma.	-0.50	0.06	1.15	1.78
My asthma was really bad.	-0.94	-0.14	1.43	2.31
I was bothered by the amount of time I spent wheezing.	-0.52	0.11	1.19	1.91
I had trouble sleeping at night because of my asthma.	-0.09	0.48	1.31	1.87
My asthma bothered me when I was with my friends.	-0.18	0.49	1.83	2.85
I had asthma attacks.	0.34	0.93	1.87	2.60
I missed school because of asthma.	0.34	0.83	1.93	2.56
I had trouble walking because of my asthma.	0.59	1.14	1.86	2.63
It was hard for me to play with pets because of my asthma.	0.43	1.19	2.18	3.00

(Continued)

Item	Univariate Thresholds				
	b_1	b_2	b_3	b_4	$b_{severity}$
I coughed because of my asthma.	-1.33	-0.54	0.55	1.35	0.01
My chest felt tight because of my asthma.	-1.06	-0.47	0.61	1.31	0.09
I had trouble breathing because of my asthma.	-0.91	-0.38	0.75	1.41	0.20
My body felt bad when I was out of breath.	-0.97	-0.36	0.77	1.46	0.22
My asthma bothered me.	-0.94	-0.34	0.78	1.43	0.23
I felt scared that I might have trouble breathing because of my asthma.	-0.83	-0.17	0.97	1.78	0.42
I got tired easily because of my asthma.	-1.03	-0.26	1.04	2.04	0.43
It was hard to take a deep breath because of my asthma.	-0.67	-0.04	0.91	1.56	0.44
I felt wheezy because of my asthma.	-0.65	-0.04	1.02	1.79	0.51
It was hard for me to play sports or exercise because of my asthma.	-0.53	0.00	1.07	1.69	0.55
My asthma was really bad.	-0.94	-0.17	1.37	2.24	0.61
I was bothered by the amount of time I spent wheezing.	-0.54	0.08	1.13	1.85	0.62
I had trouble sleeping at night because of my asthma.	-0.13	0.42	1.26	1.86	0.85
My asthma bothered me when I was with my friends.	-0.21	0.45	1.78	2.78	1.17
I had asthma attacks.	0.28	0.86	1.82	2.64	1.38
I missed school because of asthma.	0.30	0.80	1.91	2.56	1.38
I had trouble walking because of my asthma.	0.53	1.08	1.82	2.64	1.49
It was hard for me to play with pets because of my asthma.	0.41	1.17	2.16	2.99	1.68

Note: Items were arrainged by order of severity according to the univariate thresholds (i.e., $b_{severity}$). Using the expected score curve for each graded response item, the "psuedo-threshold" was computed for each item and then sorted according to items most likely to be endorssed to least likely. The single severity parameter is the location on the latent variable where a score in the middle response category is most likely. For most response functions $b_{severity}$ is similar to the average of the univariate thresholds.

Finally, a score translation table is provided (Table 12). Scaled scores for summed scores may be compared between the 18-item unidimensional IRT model and the 18-item MIRT model with logistic approximations of six items from secondary factors. Given findings in Chapter 3, one would expect that the scores and posteriors would be close between these two methods, however, rather than assuming known item parameters and comparisons between the two-tier method for bifactor models and the unidimensional use of marginal trace lines, in this case item parameters are being estimated for the unidimensional model and are being compared to items with marginal trace lines.

Given the key differences between the Chapter 3 scoring conditions and the current application, score differences appear to be as minor as previously reported. Specifically, scores and posterior standard deviations were similar using the unidimensional IRT model or the marginal trace line approach. Across the range of summed scores, the average difference in EAP estimates was 0.04 and none differred by more than 0.06 (e.g., an individual with a summed score of 64 would received a scaled score 0.06 standard deviations higher using the unidimensional slope estimates rather than the marginal estimates). Similarily, both scoring methods resulted in nearly the same precision estimates. Across the range of the latent variable, the average differenece in score precision was 1.6% (with the marginal method resulting in less precise scores), the maximum difference was 3.5%, and overall score reliability was quite similar (marginal reliability = 0.927 and 0.930, for the marginal slope and unidimensional IRT models, respectively).

Table 12. Marginal/Conditional and Univariate EAPs and SDs for 18 Asthma Symptoms Items.

	Marginal/Conditional		Univariate			
Summed Score	EAP	SD	EAP	SD	EAP Difference	SD Difference
0	-2.27	0.50	-2.26	0.50	-0.01	0.00
1	-1.93	0.42	-1.91	0.41	-0.01	0.01
2	-1.76	0.40	-1.75	0.39	-0.01	0.01
3	-1.59	0.37	-1.58	0.36	-0.01	0.01
4	-1.45	0.35	-1.45	0.34	-0.01	0.01
5	-1.33	0.33	-1.33	0.32	0.00	0.01
6	-1.22	0.31	-1.22	0.30	0.00	0.01
7	-1.12	0.30	-1.12	0.29	0.00	0.01
8	-1.03	0.29	-1.03	0.28	0.01	0.01
9	-0.94	0.28	-0.95	0.27	0.01	0.01
10	-0.86	0.27	-0.87	0.27	0.01	0.01
11	-0.79	0.27	-0.80	0.26	0.01	0.01
12	-0.71	0.26	-0.73	0.26	0.01	0.01
13	-0.64	0.26	-0.66	0.25	0.02	0.01
14	-0.57	0.26	-0.59	0.25	0.02	0.01
15	-0.51	0.26	-0.53	0.25	0.02	0.01
16	-0.44	0.25	-0.46	0.25	0.02	0.01
17	-0.38	0.25	-0.40	0.24	0.02	0.01
18	-0.32	0.25	-0.34	0.24	0.03	0.01
19	-0.26	0.25	-0.28	0.24	0.03	0.01
20	-0.20	0.25	-0.22	0.24	0.03	0.01
21	-0.14	0.25	-0.17	0.24	0.03	0.01
23	-0.02	0.25	-0.11	0.24	0.03	0.01
24	0.04	0.25	-0.05	0.24	0.03	0.01
25	0.09	0.25	0.00	0.24	0.03	0.01
26	0.15	0.24	0.06	0.24	0.04	0.01
27	0.20	0.24	0.11	0.24	0.04	0.01
28	0.26	0.24	0.17	0.24	0.04	0.01
29	0.31	0.24	0.22	0.24	0.04	0.01
30	0.37	0.24	0.27	0.24	0.04	0.01
31	0.42	0.24	0.33	0.24	0.04	0.01
32	0.48	0.24	0.38	0.24	0.04	0.01
33	0.53	0.24	0.43	0.24	0.04	0.01
34	0.58	0.24	0.48	0.24	0.05	0.01
35	0.64	0.24	0.54	0.24	0.05	0.01
36	0.69	0.24	0.59	0.24	0.05	0.00

(Continued)

	Marginal/Conditional		Univariate			
Summed Score	EAP	SD	EAP	SD	EAP Difference	SD Difference
37	0.74	0.24	0.69	0.24	0.05	0.01
38	0.79	0.24	0.74	0.23	0.05	0.00
39	0.85	0.24	0.80	0.23	0.05	0.00
40	0.90	0.24	0.85	0.23	0.05	0.00
41	0.95	0.24	0.90	0.23	0.05	0.00
42	1.00	0.24	0.95	0.23	0.05	0.00
43	1.05	0.24	1.00	0.23	0.05	0.00
44	1.11	0.24	1.05	0.23	0.05	0.00
45	1.16	0.24	1.11	0.23	0.05	0.00
46	1.21	0.24	1.16	0.24	0.05	0.00
47	1.27	0.24	1.21	0.24	0.05	0.00
48	1.32	0.24	1.26	0.24	0.06	0.00
49	1.37	0.24	1.32	0.24	0.06	0.00
50	1.43	0.24	1.37	0.24	0.06	0.00
51	1.48	0.24	1.43	0.24	0.06	0.00
52	1.54	0.25	1.48	0.24	0.06	0.00
53	1.60	0.25	1.54	0.24	0.06	0.00
54	1.66	0.25	1.60	0.25	0.06	0.00
55	1.71	0.25	1.66	0.25	0.06	0.00
56	1.78	0.26	1.72	0.25	0.06	0.00
57	1.84	0.26	1.78	0.26	0.06	0.00
58	1.90	0.26	1.84	0.26	0.06	0.00
59	1.97	0.27	1.91	0.27	0.06	0.00
60	2.04	0.27	1.98	0.27	0.06	0.00
61	2.11	0.28	2.05	0.28	0.06	0.00
62	2.18	0.28	2.13	0.28	0.06	0.00
63	2.26	0.29	2.21	0.29	0.06	0.00
64	2.35	0.30	2.29	0.30	0.06	0.00
65	2.43	0.31	2.38	0.31	0.06	0.00
66	2.53	0.32	2.48	0.32	0.05	0.00
67	2.63	0.33	2.58	0.33	0.05	0.00
68	2.74	0.34	2.69	0.35	0.05	0.00
69	2.86	0.36	2.81	0.36	0.05	0.00
70	3.01	0.38	2.96	0.38	0.05	0.00
71	3.17	0.40	3.13	0.40	0.04	0.00
72	3.43	0.45	3.39	0.45	0.04	0.00

Note: EAP differences were computed using more decimals than shown in the rounded values in the table, accounting for apparent discrepancies.

Comparing the Logistic Approximation and Two-Tier Algorithm

This application and results from Chapter 3 indicate that IRT-scores using logistic approximations of marginal trace lines are equivalent to scores computed using the two-tier algorithm when only a single item is used from each secondary dimension. This section provides more detail on the key differences between the two methods. Consider a bifactor model for binary responses to items that are influenced by a primary dimension θ_i and K secondary dimensions θ_j. Cai (2010, p. 607-608) notes that the contribution to the marginal likelihood may be approximated by quadrature points Q:

$$L_i = \sum_{q_i=1}^{Q}\left[\prod_{k=1}^{K}\sum_{q_j=1}^{Q} L_k(\theta_i,\theta_j)\varphi(\theta_j)\right]\varphi(\theta_i), \tag{16}$$

where L_k is the product of the θ_i and θ_j trace surfaces for responses to items on the kth secondary dimension. It is convenient to consider the expected score on the primary dimension from a set of responses to items on the secondary dimension k as

$$E_{ik} = \sum_{j=1}^{Q} L_k(\theta_i,\theta_j)\phi(\theta_j), \tag{17}$$

and the product of all E_{ik} over K secondary dimensions forms the marginal likelihood for the primary dimension as shown in (16). In other words, Cai integrates over the secondary dimension in the likelihood formed as the product of the IRT trace surfaces belonging to each secondary dimension.

By comparison, the method of using logistic approximations of marginal trace lines initially proposed in Chapter 3 formed the contribution to the marginal likelihood using the marginal trace line of each item from all secondary dimension θ_j:

$$L_i = \sum_{q_i=1}^{Q}\left[\prod_{1}^{nitems}\sum_{q_j=1}^{Q} T(\theta_i,\theta_j)\varphi(\theta_j)\right]\varphi(\theta_i). \tag{18}$$

where

$$T_{\text{marginal}} = \sum_{q_j=1}^{Q} T(\theta_i, \theta_j)\varphi(\theta_j). \quad (19)$$

The difference in the two techniques is that the two-tier algorithm simultaneously considers all items belonging to a secondary dimension and then integrates over the secondary dimension to form the contribution to the marginal likelihood for each secondary dimension, while the marginal trace line technique approximates the integral for each given trace surface regardless of whether or not an item belongs to a particular secondary dimension. Thus, the marginal trace line technique of using all items from all secondary dimensions ignores local dependence and improperly weights the marginal likelihood if more than one item is used from any secondary cluster.

The slight difference between the way the marginal likelihoods are formed using the two-tier algorithm and the marginal trace lines is compounded when there are both many items within each secondary dimension and many secondary dimensions. In both techniques, products of E_{ik} and T_{marginal} are used to form the marginal likelihood, but because the marginal from a likelihood computed from more than one item is not the same as the product of the marginal trace lines for more than one item, E_{ik} will not equal T_{marginal}. However, if a single item is considered from each secondary dimension, then E_{ik} will be equivalent to T_{marginal} because the integration occurs for the same trace surface. In other words, the expected score curve on the primary dimension for a single item from a secondary dimension E_{ik} is the marginal trace line T_{marginal}. Thus, when scoring tests, selecting a single item from each secondary dimension produces equivalent marginal likelihoods, and any difference in IRT-scores or SDs will be due entirely to the logistic approximation.

CHAPTER 5

CONCLUSIONS

As the use of bifactor models gains popularity, test analysts will be increasingly faced with the challenge of interpreting slope parameter estimates that must be made conditional on other dimensions. The present work provides a useful method to ease interpretability of the relation between an item response and the primary dimension by computing marginal trace lines for items represented by more than one dimension. In addition, findings suggest that a logistic function, common in many applications of item response theory, closely approximates the marginal trace line. In particular, the fit the logistic approximation was compared to marginal trace lines computed across a wide range of varying bifactor parameter estimates, and under each condition the marginal trace line was closely approximated by a logistic approximation.

Additionally, a method of IRT-based scoring is proposed that uses logistic approximations of marginal trace lines to compute unidimensional scaled scores and posterior standard deviations for the primary dimension. Using a variety of bifactor models which varied in the degree of dimensionality, test length, and factor loadings, IRT-scores and standard deviation estimates were compared between the logistic approximation of marginal trace lines and the two-tier algorithm. Contrary to initial hypotheses, it was shown that use of the logistic approximations to conduct item response theory-based scoring should be restricted to selecting a single item from each secondary factor in order to control for local dependence. Given the restriction, the contribution to the marginal likelihood for the primar

dimension is the same using either marginal trace lines or the two-tier algorithm, a MIRT-based estimation procedure which reduces integration to two-dimensions for bifactor models. Subsequently, the two methods result in nearly equivalent scaled scores and posterior standard deviation estimates. A real-data application using a bifactor model is provided which illustrates the convenience of scoring a single dimension using the logistic approximation of marginal trace lines and the utility of considering marginal slope parameters in item selection and scale development scenarios.

Regarding scoring, it was hypothesized that some computational gains may accrue from the use of marginal trace lines in a unidimensional fashion rather than multidimensional models estimated with the two-tier algorithm. Following the findings in Chapter 4, it is clear that given the restriction that a single item must be selected from each secondary factor to control for local dependence, there are rare opportunities for improvement in computational efficiency. However, in computer adaptive testing scenarios in which the restriction is imposed that only one item from any locally dependent cluster may be used for a particular respodent, scores on the primary dimension may be obtained more simply using unidimensional marginal trace lines rather than full implementation of a multidimensional adaptive test.

Appendix I:

Marginal slope parameters from combinations of primary and secondary dimension slopes

Secondary → / Primary ↓	4.50	4.25	4.00	3.75	3.50	3.25	3.00	2.75
4.50	1.59	1.67	1.76	1.86	1.97	2.09	2.22	2.37
4.25	1.50	1.58	1.66	1.75	1.86	1.97	2.10	2.23
4.00	1.41	1.49	1.56	1.65	1.75	1.85	1.97	2.10
3.75	1.33	1.39	1.47	1.55	1.64	1.74	1.85	1.97
3.50	1.24	1.30	1.37	1.45	1.53	1.62	1.73	1.84
3.25	1.15	1.21	1.27	1.34	1.42	1.51	1.60	1.71
3.00	1.06	1.11	1.17	1.24	1.31	1.39	1.48	1.58
2.75	0.97	1.02	1.08	1.14	1.20	1.27	1.36	1.45
2.50	0.88	0.93	0.98	1.03	1.09	1.16	1.23	1.31
2.25	0.80	0.84	0.88	0.93	0.98	1.04	1.11	1.18
2.00	0.71	0.74	0.78	0.83	0.87	0.93	0.99	1.05
1.75	0.62	0.65	0.68	0.72	0.76	0.81	0.86	0.92
1.50	0.53	0.56	0.59	0.62	0.66	0.70	0.74	0.79
1.25	0.44	0.46	0.49	0.52	0.55	0.58	0.62	0.66
1.00	0.35	0.37	0.39	0.41	0.44	0.46	0.49	0.53

(Continued)

Secondary → / Primary ↓	2.50	2.25	2.00	1.75	1.50	1.25	1.00
4.50	2.53	2.71	2.91	3.14	3.37	3.63	3.88
4.25	2.39	2.56	2.75	2.96	3.19	3.42	3.66
4.00	2.25	2.41	2.59	2.79	3.00	3.22	3.45
3.75	2.11	2.26	2.43	2.61	2.81	3.02	3.23
3.50	1.97	2.11	2.27	2.44	2.62	2.82	3.02
3.25	1.83	1.96	2.10	2.26	2.44	2.62	2.80
3.00	1.69	1.81	1.94	2.09	2.25	2.42	2.59
2.75	1.55	1.66	1.78	1.92	2.06	2.22	2.37
2.50	1.41	1.51	1.62	1.74	1.87	2.01	2.15
2.25	1.27	1.36	1.46	1.57	1.69	1.81	1.94
2.00	1.12	1.21	1.30	1.39	1.50	1.61	1.72
1.75	0.98	1.05	1.13	1.22	1.31	1.41	1.51
1.50	0.84	0.90	0.97	1.05	1.12	1.21	1.29
1.25	0.70	0.75	0.81	0.87	0.94	1.01	1.08
1.00	0.56	0.60	0.65	0.70	0.75	0.81	0.86

REFERENCES

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

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *21*, 179-185.

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

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

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*, 145-168.

Donoghue, J. R. (1997, March). *Item mapping to a weighted composite scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373–389.

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

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

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

Hill, C.D., Edwards, M.C., Thissen, D., Langer, M.M., Wirth, R.J., Burwinkle, T.M., & Varni, J.W. (2007). Practical issues in the application of item response theory: A demonstration using items from the Pediatric Quality of Life Inventory™ (PedsQL™) 4.0 Generic Core Scales. *Medical Care*, 45, 39-47.

Holzinger, K. J., & Swineford, R. (1937). The bifactor method. *Psychometrika*, *2*, 41-54.

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

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

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

Yeatts, K., & DeWalt, D. D. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*, 595-607.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychology Measurement*, *8*, 453-461.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389-404.

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

McKinley, R. L., & Reckase, M. N. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research report ONR 83-2). Iowa City, IA: American College Testing.

Merrell, M. (1931). The relationship of individual growth to average growth. *Human Biology, 3*, 37-70.

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

R Development Core Team (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.

Reckase, M. D. (2009). Estimation of item and person parameters. In M. D. Reckase (Eds.), *Multidimensional Item Response Theory*. New York, New York, 2009.

Reed, L. J. & Pearl, R. (1927). On the summation of logistic curves. *Journal of the Royal Statistical Society*, *90*, 729-746.

Reise, S. P., Cook, K. F., & Moore, T. M. (under review) Evaluating the impact of multidimensionality on unidimensional item response theory model item parameters.

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

Schulz, E. M., & Lee, W. (2002, April). *Describing NAEP mathematics achievement using domain scores*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA. ERIC Identifier: ED464917.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-74). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66*, 79-97.

Simms, J. L., Gross, D. F., Watson, D., O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, *25*, 34-46.

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

Thissen, D., & Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality (pp. 123-144). In Embertson, S. E. (Ed.) *Measuring Psychological Constructs: Advances in Model-Based Approaches*. Washington, D. C.: American Psychological Association Books.

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

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

Varni, J., Stucky, B. D., Thissen, D., DeWitt, E. M., Irwin, D. E., Lai, J. S., Yeatts, K., & DeWalt, D. D. (2010). PROMIS Pediatric Pain Interference Scale: An item response theory analysis of the pediatric pain item bank. *Journal of Pain*, *11*, 1109-1119.

Winsor, C. P. (1932). The Gompertz curve as a growth curve. *Proceedings of National Academy of Sciences*, *18*, 1-8.

Yeatts, K., Stucky, B. D., Thissen, D., Irwin, D. E., Varni, J., DeWitt, E. M., Lai, J. S., & DeWalt, D. D. (2010). Construction of the Pediatric Asthma Impact Scale (PAIS) for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Asthma, 47*, 295-302.