Hierarchical factor item response theory models for PIRLS: capturing clustering effects at multiple levels[1]

Frank Rijmen
Educational Testing Service, Princeton, New Jersey, USA

In large-scale assessments, items are often clustered at multiple levels. For example, the Progress in International Reading Literacy Study (PIRLS) assessment consists of item blocks. Each item block contains a reading passage followed by a set of questions. In turn, blocks of items are clustered within a literary versus an informational reading purpose. An alternative item classification scheme that is crossed with item blocks is based on the comprehension process that is involved in each of the items. The conditional dependencies between items of the same cluster can be taken into account by incorporating cluster-specific dimensions in addition to a general dimension representing overall reading ability, resulting in either a higher-order or a hierarchical model. Both types of models are formulated and applied to the PIRLS 2006 assessment. In addition, a hierarchical model is presented that incorporates a multidimensional general structure. Results indicated a moderate effect of item blocks in addition to a predominantly unidimensional general structure.

1 The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational Testing Service. The opinions expressed are those of the author and do not represent the views of Educational Testing Service, the Institute of Education Sciences, or the U.S. Department of Education.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 4

INTRODUCTION

The Progress in International Reading Literacy Study (PIRLS) is an internationally comparative reading assessment that has been carried out every five years since its inception in 2001. In 2006, there were 40 participating countries, totaling more than 200,000 participants. The study is aimed at measuring trends in children's reading literacy achievement (TIMSS & PIRLS International Study Center, Boston College, 2010).

Two overarching purposes of reading are assessed in PIRLS: (a) reading to acquire and use information, and (b) reading for literary experience. In literary reading, "... the reader engages with the text to become involved in imagined events, settings, actions, consequences, characters, atmosphere, feelings, and ideas, and to enjoy language itself" (Mullis, Kennedy, Martin, & Sainsbury, 2006, p. 19). In contrast, when reading for information, "...the reader engages not with imagined worlds, but with aspects of the real universe" (Mullis et al., 2006, p. 9). Each purpose is assessed through a set of questions that are clustered within text materials.

Another aspect of reading literacy that PIRLS focuses on is comprehension. The study distinguishes among four comprehension processes (Mullis et al., 2006): focus on and retrieve explicitly stated information, make straightforward inferences, interpret and integrate ideas and information, and examine and evaluate content, language, and textual elements. Every item assesses a single comprehension process. Each block of items assesses all four comprehension processes.

Taken together, items cluster both within reading purposes and within comprehension processes. Reading purposes are crossed with comprehension processes. Furthermore, item blocks corresponding to test materials are nested within reading purposes, and crossed with comprehension processes. Finally, the cross-classification of items is not completely balanced: the proportions of items measuring the four comprehension processes are not constant across item blocks or reading purposes.

PIRLS reports a scale for overall reading literacy as well as separate scales for purposes of reading and for comprehension processes. As might be assumed, there are two scales for reading purposes. However, there are only two scales for comprehension processes, despite four processes being identified. The first of these two scales combines the first two processes listed above; the second combines the latter two. Item parameters are calibrated separately for each of the five scales using a unidimensional item response theory model. The psychometric analyses used in PIRLS currently do not take into account the clustering of items within item blocks.

In this paper, I will present a set of multidimensional item response theory models that do take into account the effects of item clustering within item blocks and reading purposes, on the one hand, and within comprehension processes, on the other. The use of multidimensional item response theory models has always been hampered by the computational burden to obtain maximum likelihood parameter estimates, a difficulty that is amplified by the sheer size of large-scale assessment datasets.

However, if one is willing to assume specific conditional independence relations between the different dimensions of the model, exact maximum likelihood estimation methods can often be applied, even for a large number of dimensions (Rijmen, 2009, 2010, in press; Rijmen, Vansteelandt, & De Boeck, 2008). I begin my account by describing various multidimensional item response theory models, starting with the simpler and better-known models and moving to the more complex.

MULTIDIMENSIONAL ITEM RESPONSE MODELS

Hierarchical Models

The best-known example of a hierarchical model is the bifactor model. The bifactor model first appeared in the factor analysis literature for continuous manifest variables (Holzinger & Swineford, 1937). Gibbons and Hedeker (1992) adapted the model for binary data. In the bifactor model, each item is an indicator of a general dimension and one of K other dimensions. The general dimension stands for the latent variable of central interest (i.e., reading literacy), whereas the K other dimensions are incorporated to take into account additional dependencies between items belonging to the same cluster (i.e., item block, reading purpose, comprehension process).

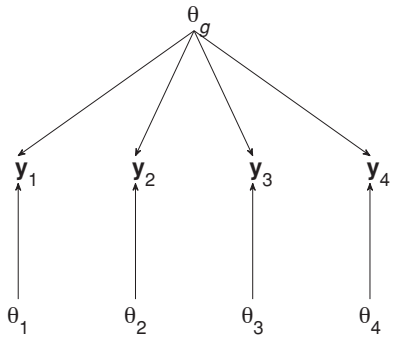
For binary data, the bifactor model can be defined as follows. Let $\mathbf{y}_{j(k)}$ denote the binary scored response on the j^{th} item, $j = 1,\ldots, J$, embedded within testlet k, $k = 1,\ldots, K$. There are J_k items embedded within each item cluster k, hence $\sum_{k=1}^{K} J_k = J$. The response vector pertaining to item cluster k is denoted by $\mathbf{y}_k$, and the vector of all responses is denoted by $\mathbf{y}$. Conditional on K cluster-specific latent variables θ_k and a general latent variable θ_g that is common to all items, statistical independent is assumed between all responses. Thus:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(\mathbf{y}_{j(k)}|\theta_g, \theta_k), \qquad (1)$$

where $\boldsymbol{\theta} = (\theta_g, \theta_{1,\ldots}, \theta_{k,\ldots}, \theta_K)$.

Typically, the latent variables are assumed to be uncorrelated and normally distributed. Figure 1 presents the directed acyclic graph of the bifactor model with uncorrelated latent variables. In the figure, arrows represent conditional dependence relations. The graph represents a model with four item blocks. Items that belong to the same item block are represented by a single node because they depend on the same set of latent variables.

Figure 1: Directed acyclic graph of a bifactor model



Furthermore, $\pi_j = P(y_{j(k)} = 1|\theta_g, \theta_k)$ is related to a linear function of the latent variables through a link function $g(\cdot)$,

$$g(\pi_j) = \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j\,, \qquad (2)$$

where $g(\cdot)$ is typically the probit or logit link function. The parameter β_j is the intercept parameter for item j, and α_{jg} and α_{jk} are the slopes or loadings of item j on the general and specific latent variables. Note that several distinct but formally equivalent parameterizations are in use in the item response theory and factor analysis literature for the model presented in Equation 2.

When the slope parameters α_{jg} and α_{jk} are known constants, a one-parameter bifactor model is obtained. Alternatively, an item-guessing parameter can also be incorporated into the expressions for the π_j's. Furthermore, for polytomous responses, the model can be extended in a straightforward way by choosing a link function $g(\cdot)$ for polytomous data (Fahrmeir & Tutz, 2001).

In order to identify the model, the location and the scale of all dimensions are fixed. Typically, the mean and variance of each dimension is set to zero and one, respectively, so that, under the assumption of normally distributed latent variables, $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$.

In this paper, the logit link function is used for binary items, $g(\pi_j) = \log(\pi_j/(1-\pi_j))$. For polytomous items, the cumulative link functions is incorporated, $g(\pi_j^{c+}) = \log(\pi_j^{c+}/(1-\pi_j^{c+}))$, with c denoting the response category, and $\pi_j^{c+} = P(y_{j(k)c} > c|\theta_g, \theta_k)$ for $c = 0,\ldots, C_j - 1$.

The bifactor model does not suffer from the problem of dimensionality that other multidimensional item response theory models do with respect to maximum likelihood parameter estimation. The reason is that the bifactor structure can be exploited when the expectation step of the expectation maximization (EM) algorithm is carried out. Specifically, the integration over all $K+1$ latent variables can be carried out through a sequence of computations in two-dimensional subspaces, where each subspace consists of the general dimension and one specific dimension. Gibbons and Hedeker

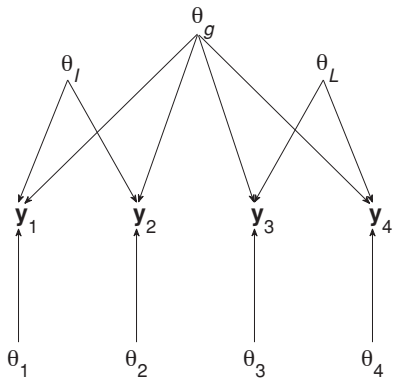
(1992) proved this result under the conditions of normally and independently distributed latent variables, and for the probit link. These limiting conditions were due to the fact that the authors relied on properties of the multivariate normal distribution.

Recently, Rijmen (2009, 2010) showed that the result is a specific example of a general procedure to exploit conditional independence relations during parameter estimation. The procedure is embedded within a graphical model framework for latent variable models (Rijmen, in press; Rijmen et al., 2008). Rijmen (2009) showed that the conditions of independently distributed latent variables can be relaxed to conditional independence of the specific dimensions, given the general dimension. Furthermore, because one does not have to rely on properties of the normal distribution, the result remains valid under any link function other than the probit function, and for latent variables that are not normally distributed. Note, however, that a model defined as such is invariant under rotation of the latent variables and requires *K* additional identification restrictions (see Rijmen, 2009, for a detailed account). I therefore continue to assume independent (and normally distributed) latent variables in the remainder of this paper.

By including a specific dimension for each item block, the bifactor model accounts for the clustering effect of items within item blocks. However, in the PIRLS assessment, there is, as noted earlier, an additional level of nesting: items are nested within item blocks, which in turn are nested within purposes of reading. This additional level can be incorporated by adding an additional layer to the bifactor model. A model defined as such is again a hierarchical model, and could be called a trifactor model: every item depends on the overall reading literacy factor, a factor specific to the reading purpose that the item is assessing, and a specific factor for the item block to which the item belongs.

Figure 2 presents the directed acyclic graph for the trifactor model. The factors at the intermediate level of the hierarchy are denoted by θ_L (reading for literary experience) and θ_I (reading to acquire and use information). Using the graphical-modeling framework, one can show that maximum likelihood estimation of a trifactor model involves a sequence of computations in three-dimensional subspaces. Each subspace contains one latent variable for each of the three levels: an item block factor, a factor for the purpose of reading in which the item block is nested, and the overall factor.

Figure 2: Directed acyclic graph of a trifactor model



Higher-Order Models

Higher-order models offer an alternative approach for modeling a nested item structure. Like hierarchical models, higher-order models originate from factor analysis for continuous variables. However, analogous models can be formulated for discrete outcome variables.

The second-order multidimensional item response theory model that takes into account the effects of item blocks incorporates a specific dimension for each of them, just like the bifactor model. It also contains a general dimension. However, items do not directly depend on this general dimension as is the case in the bifactor model. Rather, items depend directly only on their respective specific dimensions, which, in turn, depend on the general dimension. It is assumed that the specific dimensions are conditionally independent; that is, the general dimension is assumed to take into account all associations among the specific dimensions. Figure 3 displays the directed acyclic graph for the second-order model.

The model equations for the second-order model defined for binary data are

$$g(\pi_j) = \alpha_{jk}\theta_k + \beta_j, \quad (3)$$

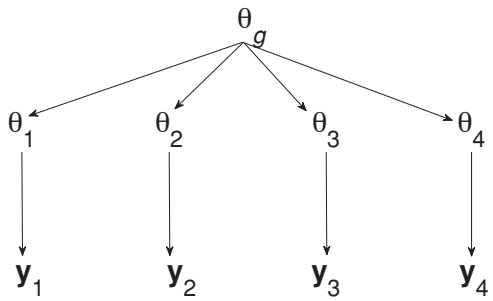
$$\theta_k = \alpha_{kg}\theta_g + \xi_k, \quad (4)$$

where α_{kg} indicates the extent to which the specific dimension θ_k is explained by the general dimension θ_g, and ξ_k is the part of θ_k that is unique. Because of the assumption that the general dimension accounts for all the dependencies between the specific dimensions, all ξ_k are assumed to be statistically independent from one another and from θ_g. Combining Equations 3 and 4 yields

$$g(\pi_j) = \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\xi_k + \beta_j . \quad (5)$$

The second-order model is identified by assuming a standard normal distribution for the latent variables $(\theta_g, \xi_1 \ldots, \xi_k)' \sim N(\mathbf{0}, \mathbf{I})$. A comparison of Equation 5 with Equation 2 shows that the second-order model is a restricted bifactor model, where, within each item block, the loadings on the specific dimensions are proportional to the

Figure 3: Directed acyclic graph of a second-order model



loadings on the general dimension. In general, a higher-order model can always be reformulated as a hierarchical model with proportionality constraints on the loadings (Yung, Thissen, & McLeod, 1999).

The second-order model for discrete observed variables was introduced in the literature on item response theory under the name of the testlet model (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Wang, 2007). The fact that these authors used a slightly different notation may have contributed to the formal equivalence between the testlet model and a second-order model having generally been ignored.

Analogous to hierarchical models, higher-order models can be formulated for assessments in which items are nested at more than one level. In the context of the PIRLS assessment, a third-order model could be formulated. In this model, item blocks would constitute the first-order factors, purposes of reading the second-order factors, and reading literacy the single third-order factor. However, this third-order model is not identified without further constraints in the specific context of the PIRLS assessment because there are only two indicators for the overarching reading literacy factor (the two reading purposes). One way to identify the model is to impose an equality constraint on the loadings of the two reading-purpose factors on the third-order factor. Alternatively, a second-order model can be formulated with two correlated factors at the second level. I discuss this model in the next section.

Bifactor or Second-Order Model with a General Factor for Each Reading Purpose

Rather than specifying a third-order model for the PIRLS assessment, one can specify a second-order model with two correlated reading purpose factors at the second level (see Figure 4). Because of the undirected edge between θ_L and θ_I, the graph is no longer a directed graph but a chain graph. Analogously, a bifactor structure can be specified that incorporates two correlated dimensions instead of a single general dimension. Cai (2010) has proposed a similar model. Figure 5 presents a bifactor model with two correlated general dimensions.

Figure 4: Chain graph of a second-order model with a second-order factor for each reading purpose

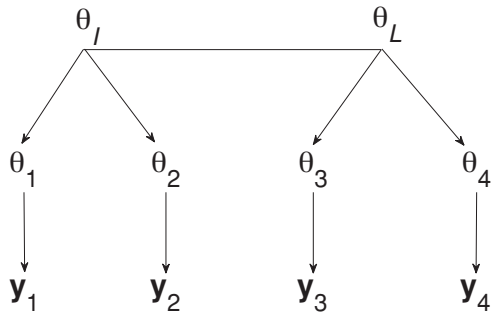
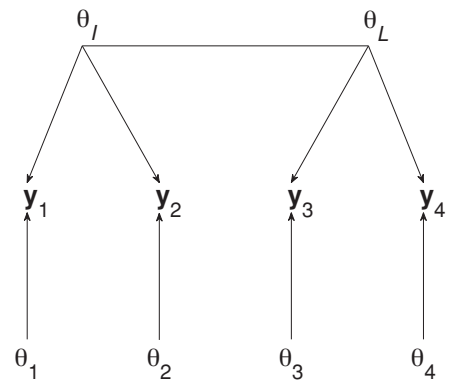


Figure 5: Chain graph of a bifactor model with a general factor for each reading purpose



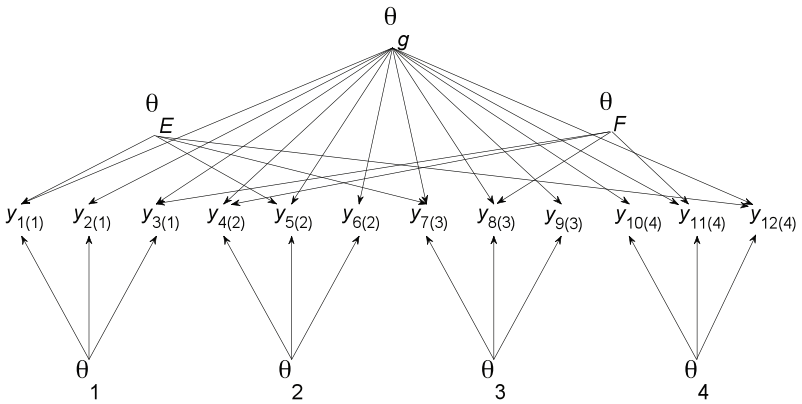
In principle, both models can be estimated with an efficient EM algorithm that involves computations in two-dimensional latent spaces. However, technically, when the latent variables are assumed to be normally distributed, the model is estimated through a Cholesky decomposition of the covariance matrix of the latent variables. As a consequence, one of the correlated reading purpose factors, say θ_L, is reformulated as a weighted sum of the other purpose factor, say θ_I, and an independent residual factor, say $\theta_{I,\,res}$. For the bifactor model, the items loading on θ_L are now loading on three dimensions: the item block factor θ_I and the residual purpose factor $\theta_{L,\,res}$. For the second-order factor, the item-block factors for the literacy item blocks now load on the uncorrelated factors θ_L and $\theta_{L,\,res}$. Therefore, computations in three-dimensional latent spaces are required in the expectation step of the efficient EM algorithm.

Double-Structure Bifactor Model

Although the models discussed up to now can be used to incorporate either the effects of item clustering within item blocks (and within purposes of reading) or the effects of items clustered within comprehension processes, they cannot take into account the crossed classification structure of item blocks with comprehension processes. Figure

6 presents a model that does take into account a crossed classification structure. The model contains a general factor and two sets of specific factors: one set for item blocks and another set for comprehension processes. To keep the visual representation clear, the figure represents only two comprehension processes: focus on and retrieve explicitly stated information (θ_F), and examine and evaluate content, language, and textual elements (θ_E). Note that because of the crossed classification structure, it is no longer possible to have a single node represent all of the items of a given item block. More specifically, this situation occurs because items within an item block relate to different comprehension processes.

Figure 6: Directed acyclic graph of a double-structure bifactor model



Similar model structures could be defined within a higher-order model structure. Furthermore, the additional nesting of item blocks within reading purposes could be incorporated by adding a layer for reading purposes on the item-block side of the model. In the context of factor analysis, the multitrait-multimethod model is also an example of a model with a crossed classification latent structure (Campbell & Fiske, 1959), which is most often defined for continuous outcome variables.

A feature that all double-structure models have in common is the fact that maximum likelihood estimation of such models scales with the smallest set of latent variables in them. Thus, for the model presented in Figure 6, maximum likelihood estimation would involve numerical integration over the sets of latent variables consisting of the general factor, one item block factor, and all of the four comprehension processes. This can be understood intuitively as follows: in the expectation step of the EM algorithm, the posterior probabilities of the latent variables, given the observed data, are computed. However, given that a response is observed, the latent variables for which the item is an indicator become conditionally dependent. For example, a person who gives a correct answer to a particular item is more likely to have a high level of reading literacy, *or* to be good at solving that particular item block, *or* to having mastered the involved comprehension process very well. Observing a high value on one of these three latent variables would make a high value on any of the other two latent variables less likely.

APPLICATION TO THE PIRLS 2006 ASSESSMENT

Description of the Dataset

The 2006 PIRLS assessment was administered in 40 countries, with a total student sample size of 215,137. The assessment contained five text passages for each reading purpose, with a total of 126 questions. The number of items within an item block varied from 11 to 14 items (Martin, Mullis, & Kennedy, 2007). A balanced incomplete booklet design was used, where each booklet consisted of two item blocks. Both constructed-response and multiple-choice items were included. All multiple-choice items were binary items, whereas some of the constructed-response items were partial credit items. In all analyses reported below, the logit link function was used for binary items, and the cumulative logit link function was used for polytomous items.

As is the case in other large-scale assessments, participants in PIRLS are sampled according to a complex two-stage clustered sampling design. The sampling design calls for the use of sampling weights during model estimation, as recently discussed by Rutkowski, Gonzalez, Joncas, and von Davier (2010). Within a country, sampling weights are computed as the inverse of the selection probability, and their sum approximates the size of the population (Foy & Kennedy, 2008). In situations in which data from several countries are combined, using these "total" weights would lead to results that are heavily influenced by the data from the large countries. Therefore, in the following analyses, "senate weights" were used. Senate weights are a renormalization of the total weights within each country so that they add up to the same constant for each country and thereby give equal weight to each country in the analyses.

Analysis Sequence

The modeling framework outlined in the previous sections is quite flexible. Consequently, the type and number of psychometric models that can be estimated from a given dataset become quite large. This calls for a strategy that allows one to determine which models to estimate. A sequential approach was followed in the present study. First, a unidimensional two-parameter logistic model was estimated. The unidimensional model served as the background model by which to evaluate the more complex models. Next, the effects of item blocks were taken into account through both a second-order and a bifactor model, with one general dimension and 10 specific dimensions (one for each testlet). Similarly, in order to take into account the effects of comprehension processes, both a second-order and a bifactor model were estimated, with one general dimension and a specific dimension for each of the four comprehension processes. Further models were specified contingent upon the results of these analyses, and therefore these models are not discussed until after presentation of the results for the bifactor and second-order models.

Results

The two-parameter logistic, the bifactor, and the second-order models

The top six rows of Table 1 present the number of parameters, the number of dimensions, the deviance, and the Akaike information criterion (AIC) (Akaike, 1973) for the two-parameter logistic, the bifactor, and the second-order models. The two-parameter model was estimated with both 10 and 20 quadrature points, whereas 10 quadrature points were used for all multidimensional models. The individual contributions of the tested students to the log-likelihood were weighted by their sampling weights.

Table 1: The number of parameters, number of dimensions, deviance, and Akaike information criterion (AIC) for the estimated models

	#Par	#Dim	Deviance	AIC
2PL_10	290	1	706956	707386
2PL_20	290	1	706431	707011
BF_IB	415	11	704925	**705755**
2O_IB	300	11	705603	706203
BF_CP	415	5	705965	706795
2O_CP	294	5	706743	707331
2D2PL	291	2	706161	706743
2DBF_IB	416	12	704757	**705569**

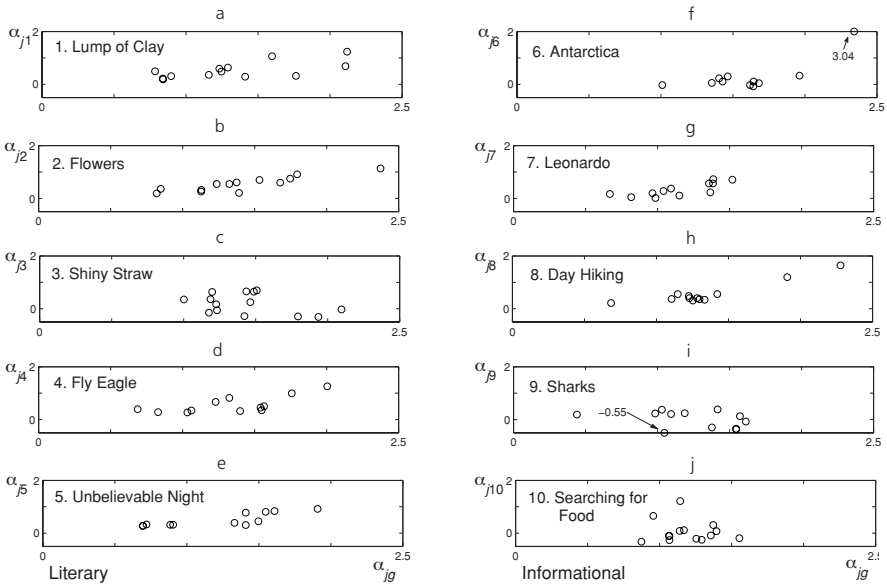
Note: #Par = number of parameters; #Dim = number of dimensions; AIC = Akaike information criterion; 2PL_10 = two-parameter logistic model with 10 quadrature points; 2PL_20 = two-parameter logistic model with 20 quadrature points; BF_IB = bifactor model with item blocks as specific dimensions; 2O_IB = second-order model with item blocks as specific dimensions; BF_CP = bifactor model with comprehension processes as specific dimensions; 2O_CP = second-order model with comprehension processes as specific dimensions; 2D2PL = between-item two-dimensional two-parameter logistic model with reading purposes as dimensions; 2DBF_IB = bifactor model with item blocks as specific dimensions and two, general dimensions representing reading purposes.

According to the AIC, the bifactor model with specific factors corresponding to item blocks emerged as the preferred model. Closer inspection of the item-parameter estimates reveals whether this model provided a better fit to the data. In Figures 7a–j, the loadings of the items on the specific dimensions are plotted against the item loadings on the general dimension, separately for each item block. Although the loadings on the specific dimensions are smaller than the loadings on the general dimension, many of them are still substantially different from zero. This pattern explains why the bifactor model provides a better fit than the two-parameter logistic model. The results vary somewhat across item blocks. For the item block labeled "Antarctica" (Figure 7f), all loadings on the specific dimensions are close to zero, except for one item that has an outlying estimated value of 3.04. For four item blocks (see Figures 7c, 7f, 7i, and 7j), the loadings on the specific dimensions are negative for some items and positive for other items, indicating both negative and positive conditional dependencies, given the general dimension. For the six other item blocks, all loadings on the specific dimension are larger than zero, indicating that all conditional dependencies are positive for the items in those item blocks.

It is furthermore clear from Figures 7a–j that, for most item blocks, the loadings on the specific dimension are not proportional to the loadings on the general dimension. If this were the case, the dots within each figure would form approximately a straight line. The lack of proportionality within each item block is a violation of the assumption of the second-order model, since the second-order model is a bifactor model in which the loadings on the specific dimension are constrained to be proportional to the loadings on the general dimension within each item block. The lack of proportionality of the loadings explains why the bifactor model provides a better fit to the data then the second-order model.

For the bifactor model in which the comprehension processes constituted the specific dimensions, the loadings on the specific dimensions were closer to zero than was the case for the bifactor model with item blocks as specific dimensions. These means were heavily influenced by some outlying estimated values for the bifactor model with comprehension processes as specific dimensions. The median value of the loadings on the specific dimensions amounted to 0.15 for this model, which is less than half of the median for the bifactor model with item blocks as specific dimensions (0.32). These results may explain why the bifactor model with the item blocks as specific dimensions provided a better fit than the bifactor model with the comprehension processes as specific dimensions.

Figure 7 a–j: Scatter plots of the loadings on the general dimension versus the loadings on the specific dimensions for the bifactor model with item blocks as specific dimensions



Note: For the "Antarctica" item block, one loading on the specific dimensions was truncated at 2.0 in Panel 6. For the "Sharks" item block, one loading on the specific dimensions was truncated at -0.5 in Panel 9.

In terms of model fit, it pays to incorporate item-cluster effects that stem from the item blocks rather than to incorporate item-cluster effects related to comprehension processes. Another indication that taking into account the comprehension processes does not have a substantial impact is that the deviance for the second-order model with comprehension processes as first-order dimensions has a higher deviance than the unidimensional two-parameter logistic model with 20 quadrature points. It seems that increasing the quadrature points from 10 to 20 for the two-parameter logistic model leads to more model-fit improvement than does modeling the comprehension processes with a second-order model.

From the first set of estimated models, it is apparent that the item blocks do constitute a separate source of individual differences. In contrast, the four comprehension processes do not seem to constitute separate dimensions, but rather are blended together into one overall dimension.

Models incorporating dimensions for purposes of reading

Because item blocks are nested within two purposes of reading, a valid question is whether the effects of item blocks that were found in the previous section are effects that can be attributed to the item blocks per se, or whether these effects merely reflect individual differences linked to the two different purposes of reading. In order to investigate this matter further, two more models were estimated: a between-item two-dimensional model, with one dimension corresponding to each reading purpose, and a bifactor model with specific dimensions corresponding to item blocks and with two general dimensions, one for each reading purpose (see Figure 5). The corresponding higher-order models were not considered because the results presented in the previous section indicated that the second-order model structure was less suited to the PIRLS 2006 dataset. The last two lines of Table 1 present the number of parameters, number of dimensions, deviance, and AIC for these two models.

The two-dimensional two-parameter logistic model provided a better fit than the unidimensional two-parameter logistic model, but it was not as good as the bifactor model with item blocks as specific dimensions. The estimated correlation between the two purposes of reading was 0.91. A visual inspection of the scatterplot revealed that the estimated loadings were very similar for the two-dimensional and unidimensional two-parameter logistic models.

After the item blocks had been taken into account through the incorporation of specific dimensions corresponding to item blocks, it was evident that the model with a separate dimension for each of the two reading purposes provided a better fit than the corresponding model with only a single general dimension, according to the AIC. The correlation between the two reading purposes was 0.93. A visual inspection of the scatterplots revealed that the estimated loadings were very similar to the bifactor model with one general dimension. This was the case for both the general dimensions and the specific dimensions corresponding to item blocks. The median of the loadings on the specific dimensions was 0.23, which is about one third lower than the median of the loadings on the specific dimensions for the bifactor model with a single general dimension.

Overall, the results indicate that the effect of item blocks can be attributed in part but not entirely to the fact that item blocks are clustered within reading purposes. However, the high correlations between the two reading purposes in the two models presented in this section, and the fact that the item loadings were very similar between the models with one general dimension and the models with a dimension for each reading purpose, indicate that the two purposes of reading do not constitute substantially different sources of individual differences; rather, they blend together into one overall dimension for reading.

CONCLUDING REMARKS

The items of large-scale assessments are often clustered at multiple levels. For example, the items used in PIRLS are clustered within item blocks related to text materials, which are further clustered within two purposes of reading. At the same time, items can be clustered according to comprehension processes. The clustering according to comprehension processes is crossed with the clustering within item blocks and reading purposes.

Several multidimensional item response theory models were presented. The models differ with respect to *how* clustering effects were taken into account; that is, they incorporated a hierarchical versus a higher-order structure. The presented models also differ with respect to the *degree* to which clustering effects were taken into account.

Some of the models presented have been previously proposed as multidimensional item response theory models. In particular, Gibbons and Hedeker (1992) presented the bifactor model for categorical data, while Bradlow and colleagues presented a second-order model (see Bradlow et al., 1999; Wainer et al., 2007). Although the more complex models presented in this paper have not been examined in the context of item response theory models, related models have been proposed in the context of factor analysis. But what is of more importance than the degree to which the proposed models are new is the explicit recognition that all of these models can be embedded within a graphical model framework for latent variable models (Rijmen, 2008, 2010, in press).

Through the application of algorithms operating on the graphical representation of the models, the conditional independence relations implied by the structure of the model can be exploited, leading to efficient maximum likelihood estimation methods. What determines the computational burden of a model is not so much its dimensionality as whether or not the latent variables of the model can be partitioned into conditionally independent subsets, after observation of the data. It is crucial to realize that prior independence of latent variables does not imply that they are independent after observation of the responses. For example, all latent variables of the model that incorporates latent variables for both item blocks and comprehension processes (the double structure model depicted in Figure 6) are assumed to be independent a priori. However, given the observed item responses, a complex dependency structure arises, as I explained when discussing the double structure model.

A sequence of models was fitted to the complete PIRLS 2006 dataset. The results indicated substantial item clustering effects related to the organization of items within item blocks. In contrast, the effects of comprehension processes and purposes of reading were minor. The loadings of the items on the specific dimensions corresponding to comprehension processes were quite low, and the correlation between the two dimensions corresponding to the two purposes of reading was very high. The loadings on the specific dimensions corresponding to item blocks remained substantial after taking into account the clustering of items within the two purposes of reading. Taken together, the findings suggest that design factors such as item blocks are more substantial sources of residual dependencies between items than content factors such as comprehension processes and purposes of reading.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kaidó.

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

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York, NY: Springer.

Foy, P., & Kennedy, A. M. (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

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

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Rijmen, F. (2009). *An efficient EM algorithm for multidimensional IRT models: Full information maximum likelihood estimation in limited time* (ETS Research Report RR-09-03). Princeton, NJ: ETS.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rijmen, F. (in press). The use of graphs in latent variable modeling: Beyond visualization. In G. R. Hancock (Ed.), Advances in latent class analysis: *A festschrift in honor of C. Mitchell Dayton*. Charlotte, NC: Information Age Publishing.

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

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142–151.

TIMSS & PIRLS International Study Center, Boston College. (2010). *About PIRLS*. Retrieved from http://timss.bc.edu/pirls2006/about.html

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

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.