Research Report

Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison

Frank Rijmen

December 2009

ETS RR-09-37

Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison

Frank Rijmen
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Three multidimensional item response theory (IRT) models for testlet-based tests are described. In the bifactor model (Gibbons & Hedeker, 1992), each item measures a general dimension in addition to a testlet-specific dimension. The testlet model (Bradlow, Wainer, & Wang, 1999) is a bifactor model in which the loadings on the specific dimensions are restricted to be proportional to the loadings on the general dimension within each testlet (Li, Bolt, & Fu, 2006). In the second-order model, the items only load on the testlet-specific factors. The correlations between the testlet-specific factors are modeled through a second-order factor. The second-order model is formally equivalent to the testlet model. The models were applied to a testlet-based international English assessment test.

Key words: Bifactor models, testlest models, second-order models, IRT

It is not uncommon for a standardized test to consist of item bundles or testlets (Bradlow, Wainer, & Wang, 1999): clusters of items that are based on a common stimulus. For example, in a reading comprehension test, a reading passage is often used as the stem for more than one item.

Responses to items pertaining to the same testlet tend to be conditionally dependent. One way to take testlet effects into account is by incorporating specific dimensions in addition to the general dimension into the item response theory (IRT) model. Three such multidimensional IRT models are described in the following sections: the bifactor model (Gibbons & Hedeker, 1992), the testlet model (Bradlow et al., 1999), and a second-order model. It is also shown how the latter two are formally equivalent and can be formulated as restricted bifactor models.

Notwithstanding the formal relation between IRT and factor analysis for categorical data had been established at least two decades ago (Takane & De Leeuw, 1987), researchers from both research communities seemingly continue to be primarily concerned with "home-grown" models and estimation methods. For example, bifactor and second-order models are most often used by researchers pertaining to the factor analytic tradition, and they are typically estimated using so-called limited-information methods when the data are categorical (Jöreskog, 1994; Muthén, 1984). In the field of educational measurement, on the other hand, the recent literature on modeling the responses stemming from testlet-based tests is dominated by the testlet model of Bradlow et al. (1999), formulated within a Bayesian framework. One underlying motivation of this study is an attempt to further narrow the gap between the two research traditions by describing the formal relations and equivalences between the psychometric models pertaining to these respective research traditions. In addition, it is shown how a common full-information maximum likelihood estimation framework can be used throughout. The latter is illustrated through applying the models to a testlet-based international English assessment test.

The Bifactor Model

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 (e.g., reading ability), whereas the K other dimensions are incorporated to take into account additional dependencies between items belonging to the same cluster. For a test that is composed of testlets, the item clusters correspond to the testlets.

For binary data, the bifactor model can be defined as follows. Let $y_{j(k)}$ denote the binary scored response on the jth item, $j = 1,\ldots, J$, embedded within testlet k, $k = 1,\ldots, K$. There are J_k

items embedded within each testlet *k*, hence $\sum_{k=1}^{K} J_k = J$. The response vector pertaining to testlet *k* is denoted by $\mathbf{y}_k$, and the vector of all responses is denoted by $\mathbf{y}$. Conditional on *K* testlet-specific latent variables θ_k and a general latent variable θ_g that is common to all items, the responses are assumed to be statistically independent,

$$P\left(\mathbf{y}\middle|\boldsymbol{\theta}\right) = \prod_{j=1}^{J} P\left(y_{j(k)}\middle|\theta_g, \theta_k\right), \tag{1}$$

where $\theta = \left(\theta_g, \theta_1, ..., \theta_k, ..., \theta_K\right)$. Furthermore, $\pi_j = P\left(y_{j(k)} = 1\middle|\theta_g, \theta_k\right)$ is related to a linear function of the latent variables through a link function $g\left(\cdot\right)$,

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

where $g\left(\cdot\right)$ 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 being used in the IRT and factor analysis literature for the model presented in Equation 2.

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

Rijmen (2009), generalizing a result of Gibbons and Hedeker (1992), showed under which conditions and how maximum likelihood estimates for the parameters of the bifactor model can be obtained through an expectation-maximization (EM) algorithm in which the E-step is carried out efficiently by exploiting the bifactor structure of the model. The algorithm is a specific instance of a general EM-algorithm in which the conditional independence relations implied by the model are rendered explicit, and are subsequently exploited, through the use of graphical models (see Rijmen, Vansteelandt, & De Boeck, 2008, for a detailed account). The

conditions under which this result holds is the assumption that the specific dimensions are conditionally independent of each other, given the general dimension,

$$p(\boldsymbol{\theta}) = p(\theta_g)\prod_k p(\theta_k|\theta_g). \tag{3}$$

Figure 1 depicts a directed acyclic graph of the bifactor model incorporating the conditional independence of the specific dimensions.

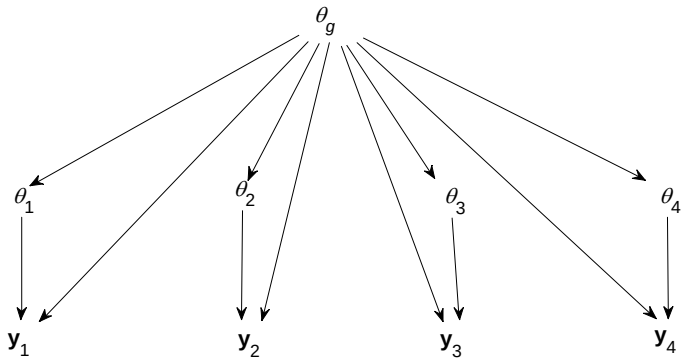

***Figure 1*. Directed acyclic graph of a bifactor model in which the specific dimensions are conditionally independent.**

In order to identify the model, the location and scale of all dimensions have to be fixed. Typically, the mean and variance of each dimension is set to zero and one, respectively. In addition, *K* restrictions are needed stemming from the rotational invariance of the model. This can be achieved by setting the correlations between each of the *K* specific dimensions and the general dimension to zero. Further details can be found in Rijmen (2009).

Commonly, a multivariate normal distribution is assumed for the latent variables. Then, the identification restrictions of the previous paragraph can be imposed by assuming a multivariate standard normal distribution for the latent variables, $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$. Since a correlation of zero implies statistical independence for the multivariate normal distribution, the corresponding directed acyclic graph can be simplified accordingly by removing the directed edges between the general and each of the specific dimensions (see Figure 2). In the application

discussed in this paper, a multivariate normal distribution is assumed for $\boldsymbol{\theta}$ for the bifactor model, as well as for the testlet model and the second-order model.

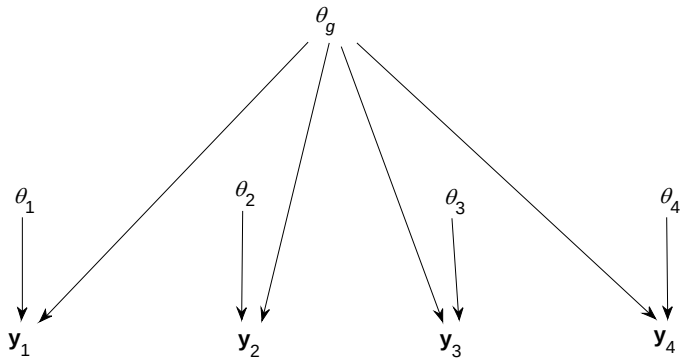

Figure 2. **Directed acyclic graph of a bifactor model with statistically independent dimensions.**

The Testlet Model

The testlet model (Bradlow et al., 1999) is a special case of the bifactor model. It is obtained by constraining the loadings on the specific dimension to be proportional to the loadings on the general dimension within each testlet (Li, Bolt, & Fu, 2006; Rijmen, 2009). Bradlow et al. (1999) formulate the testlet model in a Bayesian framework. Its analogue in a maximum likelihood framework for a model without a guessing parameter can be formulated as

$$g\left(\pi_j\right) = \alpha_{jg}\left(\theta_g + C_k\theta_k\right) + \beta_j, \tag{4}$$

where $\boldsymbol{\theta} \sim N\left(\mathbf{0}, \mathbf{I}\right)$. The testlet-specific proportionality constants C_k , k = 1,…, K, stem from the fact that, as opposed to the bifactor model, the scales of the specific dimensions have not to be fixed for reasons of model identification. The other identification restrictions of the bifactor model do carry over to the testlet model: the locations of all dimensions have to be fixed, the scale of the general dimension has to be fixed, and K restrictions are needed stemming from the rotational invariance of the model. See Li et al. (2006) and Rijmen (2009) for further details.

Maximum likelihood estimates for the testlet model can be obtained by estimating a bifactor model that is restricted accordingly. Likewise, maximum likelihood estimates can be

obtained under the same conditions (i.e., under the assumption of conditional independence of the specific dimensions) through an EM algorithm in which the E-step is carried out efficiently (Rijmen, 2009).

A Second-Order Model Item Response Theory Model

The second-order multidimensional IRT model for testlets incorporates a specific dimension for each testlet, just like the bifactor and the testlet model. It also contains a general dimension, but unlike in the bifactor and testlet models, items do not directly depend on this general dimension. Rather, items only directly depend 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, all associations between the specific dimensions are assumed to be taken into account by the general dimension. The directed acyclic graph for the second-order model is displayed in Figure 3.

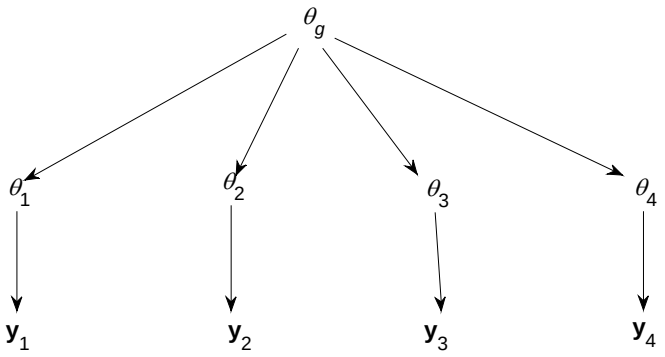

Figure 3. **Directed acyclic graph of a second-order model.**

The model equations look as follows,

$$g\left(\pi_j\right) = \alpha_{jk}\theta_k + \beta_j, \tag{5}$$

$$\theta_k = \alpha_{kg}\theta_g + \xi_k, \tag{6}$$

where α_{kg} indicates to which extent 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 it is assumed that all the

dependencies between the specific dimensions are accounted for through the general dimension, all ξ_k are assumed to be statistically independent from each other and from θ_g. Combining Equations 5 and 6 yields

$$g\left(\pi_j\right) = \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\xi_k + \beta_j, \tag{7}$$

From Equation 7, it is easily verified that model identification requires that the location of all the unique dimensions and of the general dimension have to be fixed. In addition, all scales have to be fixed as well. For the scales of the unique dimensions, this is verified by rewriting Equation 7 as

$$\begin{aligned} g\left(\pi_j\right) &= A_k\alpha_{jk}\alpha_{kg} / A_k\theta_g + A_k\alpha_{jk}\xi_k / A_k + \beta_j \\ &= \alpha^*_{jk}\alpha^*_{kg}\theta_g + \alpha^*_{jk}\xi^*_k + \beta_j, \end{aligned} \tag{8}$$

where $\alpha^*_{jk} = A_k\alpha_{jk}$, $\alpha^*_{kg} = \alpha_{kg} / A_k$, and $\xi^*_k = \xi_k / A_k$. For the general dimension, this is verified analogously by dividing θ_g by a constant and multiplying all α_{kg} with the same constant. Assuming a multivariate normal distribution for the latent variables as before, the second-order model is identified by assuming a standard normal distribution for the latent variables,

$$\left(\theta_g, \xi_1, ..., \xi_K\right)' \sim N\left(\mathbf{0}, \mathbf{I}\right).$$

Comparing Equation 7 to Equation 2, and keeping in mind that a standard normal distribution for the latent variables is assumed for both the bifactor and the second-order model, it follows that the second-order model is a restricted bifactor model, where, within each testlet, the loadings on the specific dimensions are proportional to the loadings on the general dimension. These restrictions are the same as the restrictions on the bifactor model to obtain the testlet model. Indeed, most interestingly, some further algebraic manipulations of Equation 7 clearly show that the second-order model is formally equivalent to the testlet model,

$$\begin{aligned} g\left(\pi_j\right) &= \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\alpha_{kg} / \alpha_{kg}\xi_k + \beta_j \\ &= \alpha_{jk}\alpha_{kg}\left(\theta_g + \xi_k / \alpha_{kg}\right) + \beta_j \\ &= \alpha^+_{jg}\left(\theta_g + C^+_k\xi_k\right) + \beta_j, \end{aligned} \tag{9}$$

where $\alpha_{jg}^{+} = \alpha_{jk}\alpha_{kg}$ and $C_k^{+} = 1/\alpha_{kg}$.

Upon further reflection, the finding of Li et al. (2006) that the testlet model is a restricted bifactor model, and the mathematical equivalence between the testlet model and the second-order model reported here, should not entirely come as a surprise for the researcher versed in latent variable modeling. In the field of factor analysis, these relations are merely specific instances of a more general result that has been established for continuous observed variables by Yung, Thissen, and McLeod (1999). These authors presented the general result that a higher-order model is formally equivalent to a so-called Schmid-Leiman hierarchical factor model (Schmid & Leiman, 1957). For a second-order model, the equivalent Schmid-Leiman hierarchical factor model is a bifactor model that incorporates the constraints that, within each testlet, the loadings on the specific factor are restricted to be proportional to the loadings on the general factor. Hence, the testlet model is a second-order model that has been subjected to a Schmid-Leiman transformation.

Application: An International English Assessment Test

The data from the application stem from an international English assessment test. A subset of 20 reading comprehension items organized into 4 testlets of 5 items each was selected. The reported analyses were carried out on a sample of 13,508 persons who took the test for the first time and reached the end of the test.

Three models were fitted: a unidimensional two-parameter logistic model, a second-order model incorporating a first order dimension for each testlet, and a bifactor model with a specific dimension for each testlet. For all three models, the link function $g(\cdot)$, linking the conditional response probabilities π_j to the predictor containing the latent variables (cf. Equation 2), was the logit link, $g(\pi_j) = \ln\left[\pi_j / (1-\pi_j)\right]$.

The parameters were estimated with an EM-algorithm. Because the marginal likelihood of a response pattern contains an integral over the continuous latent variable(s) of the model for which there is no closed-form solution, any maximum likelihood method requires some form of numerical approximation. For this application, integrals over the latent variable(s) were approximated over a discrete grid using the method of Gauss-Hermite quadrature (Bock & Aitkin, 1981).

For the multidimensional models, the posterior probabilities were calculated efficiently during the E-step by exploiting the conditional independence relations implied by the bifactor structure. The description of a generic EM-algorithm in which the E-step is carried out efficiently by exploiting the conditional independence relations implied by the model can be found in Rijmen et al. (2008). Rijmen (2009) describes in detail how such an efficient EM-algorithm can be constructed for the bifactor model.

Obtaining updated parameter estimates in the M-step is relatively straightforward for the two-parameter logistic and the bifactor model, since for both models the expected complete data score function corresponds to the score function of a generalized linear model in which the weight factors correspond to the posterior probabilities, and hence the standard technique of Fisher scoring can be used.

The second-order model can be estimated as a bifactor model in which the loadings on the specific dimensions are restricted to be proportional to the loadings on the general dimensions. These restrictions imply that the expected complete data score function no longer corresponds to the score function of a generalized linear model in that the predictor now contains bilinear terms, that is, the products of the loadings on the general dimension and the testlet-specific proportionality constants. The method of Fisher scoring is easily adapted to such an extension of generalized linear models (Fahrmeir & Tutz, 2001). Alternatively, a conditional maximization scheme can be implemented: given the proportionality constants, a generalized linear model is obtained with a predictor that is linear in the loading and intercept parameters. Given the latter two sets of parameters, again a generalized linear model is obtained, now with a predictor that is linear in the proportionality constants.

The standard errors were computed from the empirical observed information matrix, which is a commonly used approximation to the observed information matrix (Meilijson, 1989).

The models were specified in BNL (Bayesian networks with logistic regression nodes; Rijmen, 2006).

The number of model parameters, deviance (minus twice the log-likelihood evaluated at the maximum likelihood estimates), Akaike information criterion (AIC; the deviance plus twice the number of parameters), and Bayesian information criterion (BIC; the deviance plus the logarithm of the sample size times the number of parameters) values are presented in Table 1 for the three estimated models. In general, the model with the lowest values on the information

criteria is the preferred one, since that models offers the best balance between model misfit, as measured by the deviance, and model complexity, as measured by the term involving the number of parameters. According to both the BIC and AIC, the bifactor model is the model to be selected.

Table 1

Deviance, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) for the Two-Parameter Logistic (2PL) Model, the Second-Order Model, and the Bifactor Model

	Number of parameters	Deviance	AIC	BIC
2PL	40	322780	322860	323161
Second-order	44	321694	321782	322112
Bifactor	60	321293	321413	321864

Because it is not uncommon in an operational context to ignore dependencies due to testlet effects, it is worthwhile to explore the consequences of ignoring testlet effects. In Figure 4, the estimates of the item intercepts and the 95% confidence intervals are displayed for both the two-parameter logistic model and the bifactor model, which was the preferred model for the dataset under consideration. Figure 5 displays the same information for the loadings.

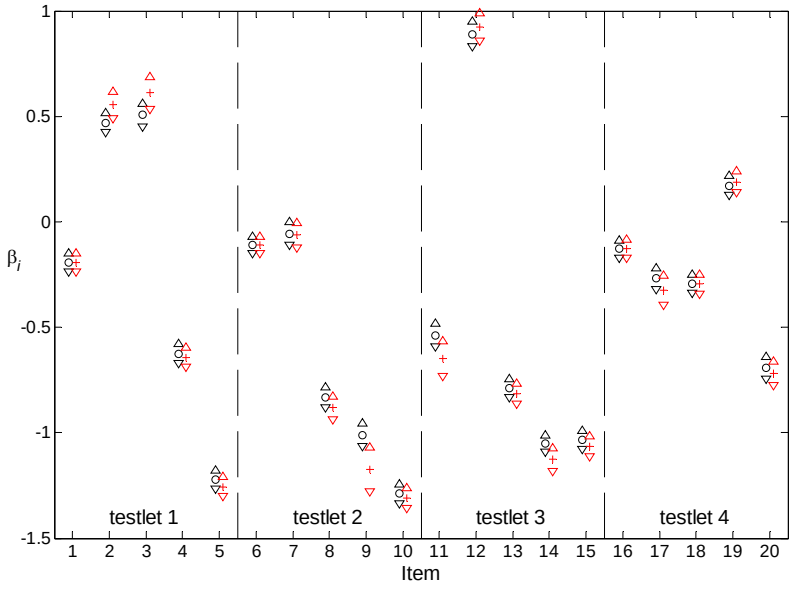

***Figure 4*. Estimates of the item intercepts and the 95% confidence intervals (triangles) for the two-parameter logistic model (o) and the bifactor model (+).**

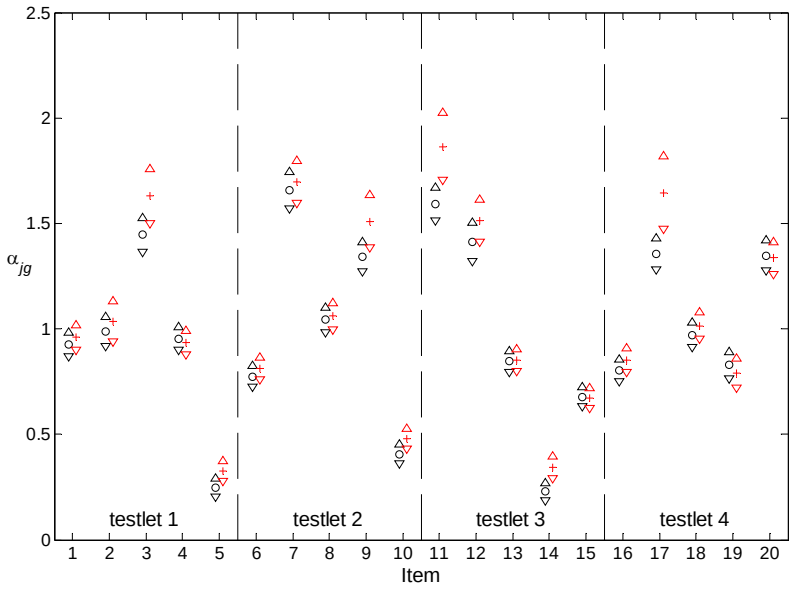

***Figure 5*. Estimates of the item loadings and the 95% confidence intervals (triangles) for the two-parameter logistic model (o) and the bifactor model (loadings on the general dimension only; +).**

The estimates of the item intercepts are quite close together, with the estimates for the two-parameter logistic model somewhat shrunk towards zero, compared to the estimates for the bifactor model for the most part. For the loadings on the general dimension, the estimates for the two-parameter logistic model are again somewhat shrunk towards zero, compared to the estimates for the bifactor model for the most part. The discrepancies in estimated values is larger than was observed for the intercept parameters.

In Figure 6, the quotient of the estimates of the loadings on the specific dimensions and the loadings on the general dimension are plotted. Were the testlet model to hold, these quotients should be (approximately) constant within each testlet, since the testlet model is a bifactor model in which the loadings on the specific dimensions are, within each testlet, proportional to the loadings on the general dimension. As Figure 6 shows, these quotients are far from constant within a testlet.

Concluding Remarks

In this paper, three multidimensional models for testlet-based tests were described. All three models have in common that the conditional dependencies between items pertaining to the

same testlet are taken into account through the incorporation of testlet-specific dimensions. The bifactor and especially the testlet model have received a substantial amount of attention in the IRT literature, whereas the second-order model is more often encountered in the factor analysis literature.

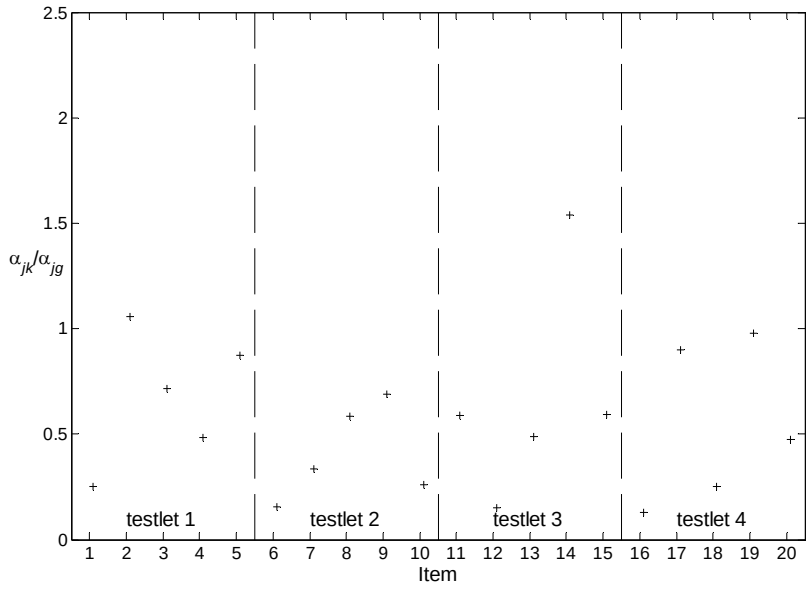

***Figure 6*. The estimates of the loadings on the specific dimensions of the bifactor model divided by the estimates of the loadings on the general dimension.**

Interestingly, it was shown in this paper that the testlet model is formally equivalent to the second-order model with a dimension for each testlet. As was shown by Li et al. (2006), the testlet model in turn is a bifactor model in which, for each testlet, the loadings on the specific dimensions are proportional to the loadings on the general factor.

Though the formal equivalence between the testlet model and the second-order model has not been explicitly described before, to the knowledge of the author, results on equivalences between similar types of models have been obtained about a decade ago. In the context of factor analysis for continuous data, Yung et al. (1999) presented the general result that a higher-order model is formally equivalent to a so-called Schmid-Leiman hierarchical factor model (Schmid & Leiman, 1957). For a second-order model, the corresponding Schmid-Leiman hierarchical factor model is a bifactor model that incorporates the constraints that within each testlet, the loadings

on the specific factor are restricted to be proportional to the loadings on the general factor. Hence, the testlet model is a second-order model after a Schmid-Leiman transformation.

The bifactor and second-order/testlet model were applied to an international English assessment test. The proportionality restrictions that the second-order model imposes on the bifactor model turned out to be implausible in the application. From a pragmatic point of view, ignoring the testlet effects and fitting a unidimensional two-parameter logistic model resulted in a mild shrinkage of the parameter estimates towards zero.

References

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

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

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

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

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381-389.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, B, 51*, 127-138.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Rijmen, F. (2006). BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes [Software manual]. Available from the Matlab Central Web site: http://www.mathworks.com/matlabcentral/fileexchange/13136

Rijmen, F. (2009). *An efficient EM algorithm for multidimensional IRT models: Full information maximum likelihood estimation in limited time* (ETS Research Rep. No. RR-09-03). Princeton, NJ: ETS.

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

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

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

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.