PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-18

Pontificia Universidad Católica de Chile
Facultad de Educación

Concurrent Unidimensional and Multidimensional Calibration within Item Response Theory

Calibración concurrente unidimensional y multidimensional dentro de la teoría de respuesta del Ítem

Steffen Brandt
opencampus.sh, Kiel, Germany

Abstract

Today, important educational achievement studies, particularly large-scale assessments, use item response theory (IRT) as the method for their analyses. An important and basic assumption of IRT is on the dimensionality of a test: In order to be interpreted unidimensionally a test has to be unidimensional and hence cannot be multidimensional. Though, this basic assumption is very often neglected. The *Programme for International Student Assessment* (PISA), for example, applies a unidimensional IRT-Model for the analysis of the mathematics achievement and at the same time applies a multidimensional IRT-model for the analysis of the four subscales of mathematics. This contradiction to one of the basic assumptions of IRT is not unique to PISA. This work, at first, discusses the currently used approaches, and presents a new approach: the generalized subdimension model (GSM). It allows the calculation of a weighted mean score within the IRT framework. The model's characteristics are compared to those of other models, particularly hierarchical models. Beyond the comparison of model fit, that is, the reliability of the results, the discussion particularly focuses on the difference in their interpretation, that is, on their validity.

Keywords: hierarchical models, bi-factor model, higher-order models, generalized subdimension model, local item dependence

Post to:
Steffen Brandt
opencampus.sh, Wissenschaftszentrum
Kiel, 24118 Kiel, Germany.
Email: steffen@opencampus.sh

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.4

Resumen

Actualmente, importantes estudios sobre logros educacionales, particularmente a gran escala, usan la Teoría de Respuesta al Ítem (TRI) como método para sus análisis. Uno de los supuestos más importantes y básicos de esta teoría es en la dimensionalidad de los test: Para ser interpretados unidimensionalmente los test deben ser unidimensionales y por lo mismo no pueden ser multidimensional. Aunque este supuesto pareciese muy básico, usualmente es ignorado. El *Programa para la Evaluación Internacional de Estudiantes* (PISA), por ejemplo, aplica un modelo TRI unidimensional para el análisis de logros matemáticos y a su vez aplica un modelo TRI multidimensional para analizar cuatro sub-escalas de matemáticas. Esta contradicción de uno de los supuestos más básicos del modelo TRI no es exclusivo de PISA. Este trabajo discute los acercamientos recientemente usados y presenta un nuevo planteamiento: el modelo de sub-dimensión generalizada. Este modelo permite el cálculo de calificaciones ponderadas dentro del marco TRI. Las características del modelo son comparadas a otros modelos, particularmente los de orden jerárquico. Más allá de las comparaciones sobre fiabilidad de resultados, la discusión se concentra particularmente en la diferencia de interpretaciones, quiere decir, en su validez.

Palabras clave: modelos jerárquicos, modelo bi-factorial, modelos de nivel superior, modelo de sub-dimensión generalizada, dependencia local de ítem

Historically, questionnaires and achievement tests[1] have been analyzed applying methods based on classical test theory (CTT), using, for example, sum scores or mean scores to interpret results, and a large part of today's analyses still are based on these methods. However, CTT has important disadvantages: (a) CTT does not include a theory about item difficulties and thereby limits the investigation of test characteristics as well as the comparison, or linkage, of test results from tests with different sets of items; and (b) CTT includes very strong assumptions on the characteristics of the test that is analyzed (see, e.g., Moosbrugger & Kelava, 2007; Rost, 1996).

As a way to avoid these disadvantages item response theory (IRT) was developed (see, e.g., Moosbrugger & Kelava, 2007; Rost, 1996), and today all important national and international studies are based on IRT analyses. However, IRT still includes many assumptions, which are often difficult to meet. An important and very basic assumption is on the given dimensionality of a test. In IRT (as well as in classical test theory (CTT)) it is true that a test has to be unidimensional in order to be interpreted unidimensionally. This seems to be a redundant notation, however, as a matter of fact in many cases one and the same test is interpreted unidimensionally as well as multidimensionally. In the Programme for International Student Assessment (PISA), for example, a unidimensional mathematics score is reported and at the same time scores in the four subdimensions *Change and Relationships*, *Quantity*, *Space and Shape*, and *Uncertainty and Data*, assuming that mathematics is a multidimensional construct. The same holds for the reading and for the science constructs investigated in PISA (OECD, 2012b) and for similar constructs investigated by other international studies, such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin & Mullis, 2012). Apparently, this is a contradiction to a basic assumption of IRT. Nonetheless, the practical necessity of yielding unidimensional and multidimensional results prevails the necessities of the theory. It is remarkable, however, that neither in PISA, nor in TIMSS, nor in PIRLS is this theoretical contradiction discussed and that its possible effects are ignored.

1 In the following both "questionnaires" and "achievement tests" will simply be denoted as tests.

A different approach is taken by the National Assessment of Educational Progress (NAEP). In this assessment, reading ability, for example, is composed of Reading for Literary Experience, Reading to Gain Information, and Reading to Perform a Task (Donahue & Schoeps, 2001). The scores for these three subdimensions of reading are calibrated using a (three-)multidimensional IRT model. The comprehensive reading score, however, is calculated as a weighted score based on the estimation results from the calibration of the subdimensions (Allen, Carlson, & Donoghue, 2001). Both the approach of using a scale score from a unidimensional IRT model and the approach of using a composite score based on a multidimensional calibration have their advantages and disadvantages (cf. Table 1), which are discussed in more detail in the next section. Following this discussion, a new approach is presented: the Generalized Subdimension Model (GSM). The GSM represents a combination of the two currently used approaches, a restriction of a multidimensional IRT model that yields a weighted mean score for the comprehensive dimension as an additional parameter of the model.

Table 1.
Pros and Cons of Obtaining a Unidimensional Score for Multidimensional Data Via a Composite Score Versus a Unidimensional Calibration

Unidimensional Calibration	
Pro	Con
• Calculation of maximum likelihood estimates is possible (WLE, MLE, ...)	• Overestimation of reliability due to neglected local item dependence (LID) • Questionable validity of the multidimensional construct if the test was constructed to be unidimensional • Implicit and unclear weighting of the unidimensional score
Composite Score Based on Multidimensional Calibration	
Pro	Con
• Consideration of LID due to the multidimensional structure and therefore more appropriate reliability estimates • Explicit and clear weighting of the unidimensional score	• Calculation of maximum likelihood estimates is *not* possible.[1] • Not appropriate for the Rasch model (standardization of the multidimensional scores leads to increased measurement error)

Advantages and Disadvantages of the Currently Used Approaches in Large-Scale Assessments

NAEP, TIMSS, PIRLS, and PISA are all large-scale assessments that have been conducted on a regular basis over the last two decades. Due to the strong political interest in the results and the considerable attention these studies receive, not only from the public but also from the scientific community, it is a natural obligation for them to conduct the studies always at the current state of the art in measurement. Due to the granted financial resources, the studies are sometimes even capable of contributing significant developments to measurement research. NAEP, for example, was responsible for the introduction of the plausible value approach into measurement, which today is a standard technique to calculate unbiased group-level estimates (Beaton, 1987; Von Davier, Gonzalez, & Mislevy, 2009).

The discussion of the approaches used by these large-scale assessments is therefore considered as a good way to provide an overview of the current state of the art in calculating unidimensional scores for multidimensional data.

Local Item Dependence

A basic underlying assumption of IRT models is local item independence (LII). LII describes the fact that the observed answers for a test are assumed to be conditionally independent given the individuals' scores on the latent variable that is measured. The violation of this assumption is denoted as local item dependence (LID). LID can occur due to various reasons, the most commonly considered is probably LID due to testlets, or item bundles. Testlets refer to items that share a common stimulus. They are popular because they allow a more economic use of the testing time; by answering several items to a single stimulus, persons need less answer time per item in comparison to reading a new stimulus for each item. However, there are drawbacks. If a person correctly answers one item of a given stimulus, the probability that he will answer an item of the same stimulus correctly is often slightly higher than the probability of answering correctly to an item from a different stimulus, in which an item has already been answered incorrectly. That is, the items show LID. The same holds for subdimensions; if a given dimension is assumed to comprise subdimensions, it is assumed that the items pertaining to a common subdimension are stronger related with each other than with items from other subdimensions.

The effects of LID in IRT analyses have been investigated by numerous authors. Unanimously, it is stated that ignoring LID leads to a biased estimation of the difficulty parameters, an overestimation of item discrimination, a bias on the variance estimate, and an overestimation of reliability (see, e.g., Monseur, Baye, Lafontaine, & Quittre, 2011; Tuerlinckx & De Boeck, 2001; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005; Yen, 1984).

The considered large scale assessments take the LII assumption differently into account. In NAEP, TIMSS, and PIRLS the tests for mathematics, for example, simply do not use item bundles but each item is provided with an individual stimulus. It is only in PISA that item bundles are used for the mathematics test. For the reading ability test, on the other hand, all studies use item bundles. In NAEP the potential LID is, therefore, investigated using available LID indices and when necessary items are collapsed to a single item to avoid local item dependence[2] (Allen & Carlson, 1987, pp. 236–237). In the technical reports of TIMSS, PIRLS, and PISA, possible local item dependencies are neither mentioned nor discussed. It is known however that the item bundles used in PISA, for example, result in LID for the respective items (Brandt, 2006; Monseur et al., 2011).

Considering LID due to subdimensions NAEP also follows a different approach. In NAEP a multidimensional IRT model is used to calibrate plausible values for each person and each subdimension, and the comprehensive scores across the subdimensions are calculated as weighted means of the plausible values (Allen et al., 2001, p. 155). This way possible negative effects on the IRT calibration due to LID by the subdimensions is avoided. In TIMSS, PIRLS, and PISA the

2 Collapsing items into a single item is one possible strategy to avoid LID; the drawback, however, is a loss in information since only the sum score of the items is considered in the IRT model and not any more the individual scores of the respective items (Yen, 1993).

subdimensions are calibrated jointly, using a unidimensional IRT model; possible effects due to LID are not considered.

Subdimension Weighting

In NAEP the subdimensions are calibrated as separate dimensions and subject experts provide a weight for each subdimension within the overarching comprehensive dimension. This way, the score for Reading in Grade 12, for example, is composed by considering *Reading for Literary Experience* with a weight of 35%, *Reading to Gain Information* with a weight of 45%, and *Reading to Perform a Task* with a weight of 20% (Donahue & Schoeps, 2001).

Table 2
Mathematics Score Distributions by Subdimension for the PISA 2003 Paper-Based Main Survey and the PISA 2012 Paper-Based and Computer-Based Main Survey

Subdimension	PISA 2003 Paper-Based	PISA 2012 Paper-Based	PISA 2012 Computer-Based
Change and Relationships	30.4% (28 points)	26.1% (24 points)	29.2 (14 points)
Quantity	23.9% (22 points)	23.9% (22 points)	20.8% (10 points)
Space and Shape	22.8% (21 points)	25.0% (23 points)	31.3% (15 points)
Uncertainty and Data	22.8% (21 points)	25.0% (23 points)	18.8% (9 points)
Total	100% (92 points)	100% (92 points)	100% (48 points)

Note. The scores were calculated based on the item classifications given in Appendix 12 of the PISA 2003 Technical Report and Annex A of the PISA 2012 Technical Report (OECD, 2005, 2012b).

In TIMSS, PIRLS, and PISA the weights of the subdimensions are less clear, and in fact vary from test to test. To demonstrate this, the weights for the PISA mathematics test are exemplarily depicted in Table 2. The IRT model used in PISA to calibrate the achievement scales is based on the Rasch model (Rasch, 1980). In the Rasch model, the weight of an item corresponds to the maximum score achievable for this item. Dividing the maximum score for each subdimension by the total score achievable therefore provides the weights of the subdimensions. In PISA 2003 as well as in PISA 2012 the assessment framework for the mathematics tests (in these two PISA cycles mathematics was the focus domain and included the estimation of the subdimensions) specified an equal weighting of 25% for each of the subdimensions *Change and Relationships*, *Quantity*, *Space and Shape*, and *Uncertainty and Data* (OECD, 2012a, 2004). The actual weightings, however, varied between 22,8% and 30,4% in PISA 2003, between 23,9% and 26,1% for the paper-based test in PISA 2012, and for the computer-based assessment in PISA 2012 between 18,8% and 31,3%. An important reason for the variations in the weightings is the fact that it is very difficult to predict the final total score of a subdimension at the moment the test is administered. The final scores are fixed only after knowing the response data and the resulting item characteristics. For PISA 2012 for example—even

though all items went through an extensive field trial—, a mathematics item included in the main trial was deleted because of concerns regarding the consistency with which the intended coding rule was applied across countries, and for six countries an item was deleted on the national level because the item (in each case a different one) showed a difficulty that was inconsistent with the difficulty observed across the remaining countries (OECD, 2012b, pp. 231–232). As a consequence, the deleted items' subdimensions will have a smaller weight within the total score. In fact, considering the weightings of the subdimensions for the six countries with national deletions, these will even slightly differ from the weighting for the remaining countries. Another reason for the final maximum score to change might be that an item that was administered to differentiate the persons in three scoring categories (0 points, 1 point, and 2 points) does not show sufficient variability in the answers, so the scoring categories are collapsed to just two (0 points and 1 point).

Considering the final weightings in TIMSS and PIRLS a prior definition of the actual weights is even more complicated since these two studies use a 2-PL IRT model (Birnbaum, 1968). While the Rasch model only estimates a difficulty parameter for each item, the 2-PL model additionally estimates a discrimination parameter for each item, which allows modeling the data more accurate and increasing the reliability. The drawback, however, is that the items obtain different weights for the calibration of the final score, in which items with higher discriminations get a higher weight and items with a lower discrimination a lower weight. Correspondingly, the weight of a subdimension also changes if the average discrimination of its items is above or below the average of the total test.

Reliability

In NAEP the IRT analyses always consider the subdimensions separately. Also, the item fit statistics calculated for the field trial, for example, are based on the results in each of the respective subdimensions. Hence, the psychometric item selection process is aiming at a maximization of the reliability for the measurement of the subdimensions. A possible disadvantage of this multidimensional test construction might be that the reliability of the overarching comprehensive score is reduced since the items are not fitted to be on a common scale. This reduction has a two-fold reason: first, it is easier to develop a scale with a high reliability if you can make use of more items (if the existing abilities differ strongly, it is also necessary to a certain extent to cover the whole range), and, second, the less the separate scales correlate, the less they measure a common construct and the lower the reliability of the measured overall scale. That is, the more the separate scales are actually able to identify and measure different constructs, the worse for the overall scale. These are probably the reasons why in PISA the unidimensional test construction approach is preferred. The more serious disadvantage, however, is probably that the corresponding approach does not allow the calculation of reliable individual scores, which is necessary in any entrance exam or similar high-stakes tests. The plausible values allow calculating reliable group-level results for the comprehensive scores, but the calculation of individual point estimates, such as the WLE, MLE, or EAP is not possible (for more details on these estimates see, e.g., Rost, 1996). The approach is therefore only appropriate if the calculation of individual estimates is not necessary. Furthermore, the approach is not appropriate if the test is analyzed using the Rasch model. In contrast to the 2-PL model, the calibration of the Rasch model does not allow constraining subdimensions to equal variances without imposing additional constraints on the data. Therefore, in the case of the Rasch model the plausible

values have to be standardized after the calibration of the model, using the estimated variances of the subdimensions. In doing so however, the standard error of the variance estimate of a subdimension will be added to each plausible value of that subdimension, and the plausible values lose their beneficial characteristic of being unbiased due to the estimation.

In PISA, TIMSS, and PIRLS the focus is clearly on a unidimensional test construction. The item selection process is based on the unidimensional Rasch model[3] and aimed at maximizing the reliability of the overarching comprehensive scores. The advantage of this approach is that it provides the option to calculate reliable individual scores. Furthermore, fitting the items to the comprehensive scale might yield a higher reliability for this scale. However, if the items attributed to a common subdimension have something particular in common—and hence include LID due to the subdimensions—, the estimated reliability will be biased and higher than it actually is.

Validity

Besides the technical requirements of measurement considered in the above sections, it is also important to consider to what extent the constructed measures are valid, that is, yield the intended use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The intended use for the subdimensions in mathematics, for example, is to differentiate the persons' achievements in specific areas of mathematics in order to identify weaknesses and strengths within the area of mathematics. Such a distinction is only useful, however, if the subdimensions actually differ. In NAEP several dimensionality analyses were conducted, and it was decided that it is reasonable to interpret mathematics as a multidimensional construct (Allen et al., 2001, pp. 155–156). The technical reports of PISA, TIMSS, and PIRLS do not include any results from analyses investigating the assumed dimensional structures of the mathematics, reading, or science scales. Here instead the reported results for the subdimensions can be taken as an indication for the given differences between the subdimensions. In the PISA 2012 mathematics achievement test, for example, the Netherlands achieved a mean score of 532 points for the subdimension *Uncertainty and Data* and a score of 507 points for *Space and Shape* (OECD, 2014). Besides statistical significance, which can be assumed as given considering the sample sizes in PISA, the difference corresponds to an effect size of 0.25 using Cohen's d[4], and can therefore also be considered as meaningful (Cohen, 1988). That is, in this case for the PISA mathematics test, the results indicate that the subdimensions in fact measure different aptitudes and, hence, are multidimensional. However, we have to consider that the test was constructed to be unidimensional; in fact, across all pilot studies, the field trial, and also the main survey, the items were constructed and selected in order to fit on a common unidimensional scale. The important question for the validity of the multidimensional results therefore is: Are the items selected to measure the subdimensions representative for the actually defined aptitudes? It might well be that the results for the subdimensions are biased due to the test construction and that therefore the validity of the subdimension results is reduced.

3 TIMSS and PIRLS use the 2-PL model to calibrate the final results, the analysis of the item characteristics and the item selection process, however, is based on the Rasch model (Martin & Mullis, 2012).

4 The PISA scales are standardized to a standard deviation of 100; the difference of 25 points therefore corresponds to a Cohen's d of 0.25.

For a different reason also the validity of the unidimensional results might be affected if the subdimensions actually differ (and only then it makes sense to report them separately). All considered large-scale assessments use a rotated booklet design, in which not all items are administered to all persons but each person answers only a sample of items. In PISA 2012, for example, the main survey includes 13 different booklets, each composed of four, so-called, item clusters, which include items for 30 minutes of testing time (that is, each booklet includes items for 2 hours of testing time). For mathematics the test includes seven different clusters distributed across the 13 booklets, with some booklets including just one of these clusters and others including up to three (OECD, 2012b, p. 31). The booklets 2, 8, 12, and 13 each include only one cluster. Table 3 shows the weights of the subdimensions within these clusters. The weights vary from 7,7% to 40%, which makes clear that a person, for example, with a relative strength in *Change and Relationships* and a relative weakness in *Uncertainty and Data* will receive a different score depending on whether he or she completed booklet 8 or booklet 13. If the test of mathematics includes multidimensionality due to the subdimensions, it will therefore not yield valid results for the comprehensive achievement in mathematics on the individual level. Considering group-level results these can still be considered as valid since the booklet differences will be equaled out if the groups are sufficiently large. However, since the unidimensional calibration assumes that the individual scores in mathematics are independent of the booklets' weightings according to the subdimensions, it is plausible to assume that the rotated booklet design leads to an overestimation of the unidimensional reliability estimate (additional to the overestimation of the reliability due to LID described in the previous section). A more detailed consideration of this aspect is beyond the scope of this work, though.

Table 3
Weights of the Mathematics Subdimensions in the Booklets 2, 8, 12, and 13 in PISA 2012

Booklet	Change and Relationships	Quantity	Space and Shape	Uncertainty and Data
2	30.8%	30.8%	7.7%	30.8%
8	15.4%	23.1%	23.1%	38.5%
12	18.2%	36.4%	27.3%	18.2%
13	40.0%	20.0%	26.7%	13.3%

Note. The weights were calculated based on the score and item classifications given in Annex A of the PISA 2012 Technical Report (OECD, 2012b).

Merging the Currently Used Approaches: The Generalized Subdimension Model

In order avoid the above described disadvantages of the respective approaches currently used in large-scale assessments, the GSM was developed. It combines the two approaches by restricting a multidimensional IRT model to yield an additional weighted comprehensive score. This way, the model allows calculating reliable individual scores and at the same time avoids the described problems by the inappropriate assumption of unidimensionality. The development of the model was conducted in two steps. In the first step, the subdimension model was developed. This earlier version of the GSM also allows calculating a weighted mean score based on a restriction of the multidimensional model, however, it includes a hidden constraint on the variances of the subdimensions. Therefore, the Subdimension Model only shows a fit equal to the multidimensional model if the variances of the subdimensions are equal (cf. Brandt, 2008, 2010, 2012b). In the second

step the Subdimension Model was generalized to the GSM by incorporating an additional parameter type considering the subdimensions' variance differences (cf. Brandt, 2012a, 2016; Brandt, Duckor, & Wilson, 2014). After providing its definition, the following sections consider the characteristics of the GSM and provide information on how to classify it in comparison to other currently existing models. The paper then concludes with some final remarks on current and possible future research.

Model Definition

The definition of the GSM (in its partial credit extension) is as follows:

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = d_{k(i)}\left(\theta_n + \gamma_{nk(i)}\right) - b_{ij} \quad , \quad (1)$$

where p_{nij} is the probability of person n giving an answer corresponding to answer category j of item i; p_{ni0} the corresponding probability of giving an answer matching category (j-1); b_{ij} is the difficulty of step j of item i; θ_n is person n's ability on the constructed unidimensional dimension (denoted as main dimension); $\gamma_{nk(i)}$ is the person's subtest specific ability for (sub-) dimension k (with item i referring to dimension k) relative to the ability on the main dimension; and $d_{k(i)}$ is the translation parameter that translates the different multidimensional (or subdimensional) scales to a common one. Corresponding to hierarchical models, it is assumed that each item loads on exactly one subdimension. To identify the model several restrictions on the parameters have to be applied. First, the mean of the ability estimates θ and γ_k have to be constrained to zero, and the correlations between the main dimension and the K subdimensions have to be set to zero. Further, for each person the sum of the subtest specific parameters has to be constrained to zero ($\sum_k \gamma_{nk} = 0$; Constraint I), and the square of the parameters d_k are constrained to the sum of K with each d_k additionally constrained to be positive ($\sum_k d_k^2 = K$; Constraint II).

The latter two constraints result from the characteristics of a mean score, and it can be shown that the given definition results in the main ability estimate to be the (equally weighted) mean of the specific abilities.

The Testlet, the Higher Order, and the Hierarchical Model

The problem of calculating unidimensional scale scores for data assumed to be multidimensional has received substantial attention and had led to the formulation of a growing variety of IRT models to cope with this issue. In this section, the different groups of existing IRT models will therefore, at first, shortly be characterized before the special characteristics of the GSM and its relationship to these models are considered in more detail in the following sections.

Depending on whether an assumed LID originates in the test construction or in the psychological construct that is to be measured, the models are typically denoted as testlet models, or as hierarchical or higher-order models, respectively.

Testlet models (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) assume that the answers on a test depend on a single psychological construct. Additionally though, they assume that, due to a testlet based test construction, the test includes multidimensionality that corresponds

to each person's familiarity with the stimulus given for a testlet. From the perspective of the unidimensional latent trait that is to be measured this multidimensionality corresponds to local item dependence (LID).

Hierarchical or higher-order models (de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sheng & Wikle, 2008), on the other hand, typically assume that the answers in a test depend on multiple psychological constructs that are related by a common underlying construct. That is, from a unidimensional perspective, the test includes local item dependencies not due to the test construction but due to the nature of the psychological construct.

While the reasons for the necessity to model multidimensionality (or LID) are different for the two mentioned groups of models, the statistical assumptions made are very similar. Yung, Thissen, and McLeod (1999) and Li, Bolt, and Fu (2006) have shown that both the higher-order model and the testlet model are restrictions of the hierarchical model. Furthermore, Rijman (2010) has shown that the testlet model is formally equivalent to a second-order model (i.e., a higher-order model with a second order as the highest order). A general assumption of hierarchical models (i.e., as well of testlet and of higher-order models) is that existing correlations between the subtest (or testlet) factors fully originate in the underlying common latent trait they measure (Holzinger & Swineford, 1937). That is, constrained on the common latent trait, often denoted as g-factor and in the GSM denoted as main dimension, the subdimension factors are independent (see Rijmen, 2010; Yung et al., 1999). To what extent this theoretical requirement of hierarchical models and the resulting constraints applied for the estimation lead to a significantly worse model fit in comparison to the common multidimensional model depends on the given multidimensional construct that is measured (or on the stimuli used for the testlets). In chapter 3, it is shown that for the TIMSS 2003 mathematics achievement test, for example, the difference in the deviance of the testlet model and the multidimensional model is about 50% of the difference in deviance between the multidimensional model and the unidimensional model. That is, the testlet model is only partially able to model the multidimensionality in the given data set. The ability of testlet models, or more generally hierarchical models, to model multidimensionality will be different for each data set depending on the given covariance structure of the subdimensions. Due to the restrictive assumptions, however, it is very unlikely that they will be able to fully model the multidimensionality, and the resulting model fit will therefore be significantly worse, as in the analysis using the TIMMS data.

Estimated Parameters

A basic characteristic of all IRT models is the number of parameters they use. The more parameters a model comprises, the better it will typically be able to model a given data set. A higher number of parameters, however, usually means a more complicated interpretation, whereas models with fewer parameters often provide clearer interpretations. On the other hand, fewer parameters typically correspond to stronger assumptions, that is, more constraints for the calibration of the model. Comparing the number and types of estimated parameters is therefore a useful way to discuss the characteristics of different models.

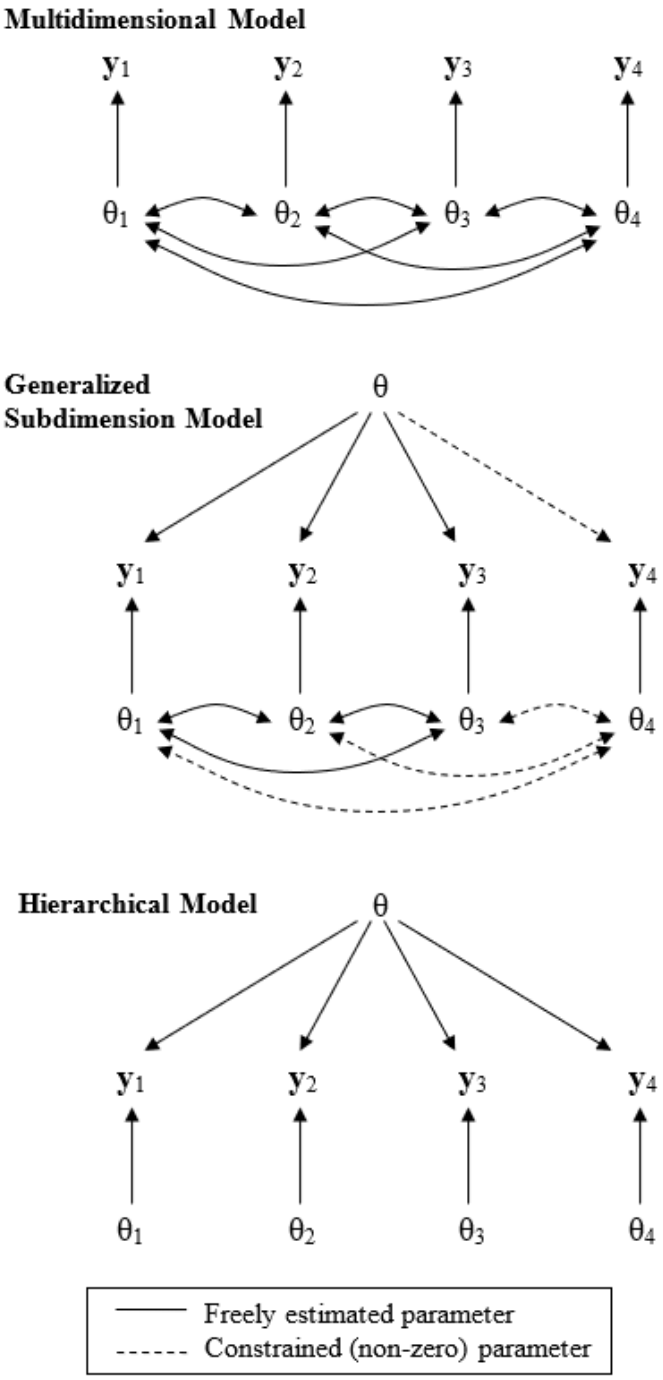

Figure 1. Graphical representations of the estimated parameters of the multidimensional model, the generalized subdimension model, and the hierarchical model.

The GSM's constrains the means of the person abilities to zero, which is a standard constraint in IRT models and is necessary to fix the scales on the latent continuum. Furthermore, it is a characteristic of mean scores calculated from distributions (here dimensions) with equal variances to yield a covariance of zero between the mean scores and the distributions of the difference values, that is the distribution values minus the respective mean scores. Hence, constraining the covariance of the main dimension and the subdimensions to zero does not constrain the estimation of the mean score. This constraint is still in accordance with the hierarchical models, the remaining constraints of the GSM are different. They are necessary to allow for correlations between the subtest specific parameters, which are constrained to be independent in hierarchical models. The differences between these assumptions of the GSM and the hierarchical model are depicted in Figure 1. Here, $\mathbf{y}_k$ denotes the response vector pertaining to subtest *k*, θ the latent variable of the main dimension, and θ_k the latent variable of subdimension *k*. Additionally, the characteristics of the two models are contrasted to the multidimensional model that is depicted as well, and it can be shown that the number of (freely) estimated parameters in the GSM equals that of the multidimensional model.

The multidimensional model (cf. Rost, 1996) is defined as

$$\log\left(\frac{p_{ni1}}{p_{ni0}}\right) = \mathbf{a}_i\left(\boldsymbol{\theta}_n - b_i\mathbf{1}\right), \quad (2)$$

where $\boldsymbol{\theta}_n = \left(\theta_{n1},\ldots,\theta_{nK}\right)^T$ is the ability vector of person *n* for his/her abilities in the *K* dimensions; $\mathbf{a}_i = \left(a_1,\ldots,a_K\right)^T$ is a vector with values of only 0 and 1, indicating whether an item loads on a dimension or not; $\mathbf{1}$ is the unity vector with *K* elements; and the remaining variables are defined as above. Without loss of generality, we assume that the variances in the multidimensional model are constrained to a mean of zero to identify the model. That is, *K* parameters for the estimation of the dimensions' variances must be estimated. In the generalized subdimension model as well *K* variances have to be estimated: *K*-1 subdimension specific variances (one subdimension specific variance, typically that of subdimension *K,* is not estimated since its parameters result via the constraint $\sum_k \gamma_{nk} = 0$) and the variance for the main dimension (cf. Equation 5.1 in the previous chapter). That is, as well *K* parameters for the dimensions' variances have to be estimated.

Considering the covariance estimates, these are all estimated freely in the multidimensional model, resulting in $\sum_{k=1}^{K}(k-1)$ covariance parameters to be estimated. In the GSM the covariance matrix includes the main dimension, of which the covariances are constrained to zero, and *K*-1 subdimensions (one subdimension results from constrained parameters; cf. above) with freely estimated covariances; that is, here $\sum_{k=1}^{K-1}(k-1)$ covariance parameters have to be estimated, which are *K*-1 covariance parameters less than in the multidimensional model. However, there are exactly *K*-1 additional parameters to be estimated for the translation parameters d_k (cf. definition above). Since the number of estimated parameters for the items' difficulties are equal as well, the GSM and the multidimensional model hence comprise an equal number of parameters.

Equivalence With the Multidimensional Model

The equivalence of the GSM and the multidimensional model is provided via the definition of $\theta_{nk} = d_k(\theta_n + \gamma_{nk})$, where θ_{nk} equals the ability estimate for the corresponding (sub-)dimension k in the multidimensional model. To show the equivalence, it is demonstrated in the following that the estimation of the means, variances, and covariances for the distributions of the parameters θ_{nk} are equal to those of the multidimensional model.

Without loss of generality, it is again assumed that both models are estimated using constraints on the cases. That is, the distributions of the parameters θ_n and γ_{nk} are constrained to a mean of zero. Hence, the K distributions of the θ_{nk} are trivially constrained to a mean of zero as well, and their means correspond to those in the multidimensional model.

The independence of the variance estimation for the subdimensions within the GSM is not as straightforward. By constraining the subdimension specific parameters to a sum of zero for each person (Constraint I), the estimations of the ability parameters for the different subdimensions are dependent on each other; because of this, the standard subdimension model includes an implicit variance restriction for the estimation of the subdimensions by applying this constraint. To neutralize this implicit variance constraint, the additional introduction of the translation parameters d_k is necessary. They yield that each variance for the K subdimensions is estimated independently. However, since the main dimension variance is also estimated (resulting in a total number of K+1 estimated variances) the variance parameters need a further constraint in order to be identified, which is yielded by constraining the square of the parameters d_k to the sum of K (Constraint II).

The correspondence of the covariances in the GSM and the multidimensional model is shown by Equation 3. For any two distributions of θ_{n1} and θ_{n2}, it is true that

$$\begin{aligned}
\mathrm{Cov}(\theta_{n1}, \theta_{n2}) &= \mathrm{Cov}\left(d_1(\theta_n + \gamma_{n1}), d_2(\theta_n + \gamma_{n2})\right) \\
&= d_1 d_2 \mathrm{Cov}(\theta_n + \gamma_{n1}, \theta_n + \gamma_{n2}) \\
&= d_1 d_2 \left(\mathrm{Cov}(\theta_n, \theta_n) + \mathrm{Cov}(\theta_n, \gamma_{n2}) + \mathrm{Cov}(\gamma_{n1}, \theta_n) + \mathrm{Cov}(\gamma_{n1}, \gamma_{n2})\right) \\
&= d_1 d_2 \left(\mathrm{Var}(\theta_n) + \mathrm{Cov}(\gamma_{n1}, \gamma_{n2})\right) \\
&\qquad (\text{since } \mathrm{Cov}(\gamma_{n1}, \theta_n) = \mathrm{Cov}(\gamma_{n2}, \theta_n) = 0)
\end{aligned} \tag{3}$$

That is, the existing covariance structure between the dimensions in the multidimensional model can be fully recovered by the generalized subdimension model even though the underlying parameters θ_{nk} are split into the parameters γ_{nk}, θ_n, and d_k, and only the covariance structure of the parameters γ_{nk} is estimated.

Correspondence with the Unidimensional Model

If the assumed multidimensional structure does not exist but the subdimensions in fact measure the same construct, the variances of the subdimension specific parameters γ_{nk} reduce to zero, and, hence, it is true that all subdimension variances are equal (namely zero). In order to yield Constraint II of the GSM, the parameters d_k then are all equal to 1, and the variance of the main dimension equals the variance of the ability estimates in the unidimensional model. Due to the definition of Constraint II, it is generally yielded that the variance of the main dimension is the mean of the unidimensional variance components of the subdimensions, that is, the mean of the total subdimension variances less the mean of the subdimension specific variances (see Equation 4).

$$\begin{aligned}
\mathrm{Var}\,(\theta_n) &= \mathrm{Var}\,(\theta_n)\frac{\sum_k d_k^2}{K} \qquad \text{(due to Constraint II)} \\
&= \frac{\sum_k d_k^2 \mathrm{Var}\,(\theta_n)}{K} + \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= \frac{\sum_k \mathrm{Var}_k\left(d_k\left(\theta_n+\gamma_{nk}\right)\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= \frac{\sum_k \mathrm{Var}_k\left(\theta_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= M\left(\mathrm{Var}_k\left(\theta_{nk}\right)\right) - M\left(\mathrm{Var}_k\left(d_k\gamma_{nk}\right)\right)
\end{aligned} \tag{4}$$

Furthermore, as shown by Equation 5, each person's ability in the main dimension is the mean of his/her abilities in the subdimensions considered on the common scale they are translated to. As noted before, this translation is yielded by the parameters d_k. A multiplication using the reciprocal of d_k therefore translates the subdimension ability estimates onto the scale in which their variances are of equal extent. That is, while θ_{nk} is equivalent to the multidimensional ability estimate, $\frac{1}{d_k}\theta_{nk}$ is the corresponding ability estimate translated to the scale in which the variances are of equal extents

$$\begin{aligned}
\theta_n &= \theta_n + \frac{\sum_k \gamma_{nk}}{K} \qquad \text{(due to Constraint I)} \\
&= \frac{K\cdot\theta_n}{K} + \frac{\sum_k \gamma_{nk}}{K} \\
&= \frac{\sum_k \left(\theta_n+\gamma_{nk}\right)}{K} \\
&= \mathrm{M}\left((\theta_{n1}+\gamma_{n1}),\ldots,(\theta_n+\gamma_{nK})\right) \\
&= \mathrm{M}\left(\frac{1}{d_1}\theta_{n1},\ldots,\frac{1}{d_K}\theta_{nK}\right)
\end{aligned} \tag{5}$$

Relationship to the Hierarchical Model

The restriction of the translation parameters d_k via Constraint II in the GSM makes clear that these cannot be interpreted as regression coefficients. This is in contrast to the hierarchical models, in which the corresponding parameters are equivalent to the loadings or regression coefficients of the subdimensions on the main dimension (cf. de la Torre & Song, 2009). However, the estimation of these coefficients in the hierarchical models relies on the assumption that the subtest specific abilities are independent and will therefore correspond to the correlation of the subtest ability and the overall test ability only if this independence assumption holds. The GSM on the other hand allows for a correlation of the subdimension specific abilities. Following the definition of Holzinger and Swineford (1937), the GSM, thereby, corresponds to a modified hierarchical model, which denotes a hierarchical model with overlapping specific factors.

Besides the differences in the assumed covariance structure, the difference between the hierarchical model and the GSM is also depicted by the different levels on which the constraints of the models are applied. While the main constraint of the hierarchical model yields a characteristic on the level of the test, or the latent trait (the independence of the distributions for the specific factors), the main constraint of the GSM yields a characteristic on the level of the individual person, namely, that the sum of the (translated) specific ability estimates for each person is zero.

Conclusion

Considering the multidimensional model, it has been shown above that the GSM yields equivalent parameter estimates and that the unidimensional parameter estimates are constructed as means of the translated multidimensional parameters. The application of the GSM thereby yields important advantages for the calculation of unidimensional achievement scores for multidimensional tests. First, the resulting unidimensional estimate is free of any negative impact of LID due to the subdimensions; second, the estimation is conducted within the common framework of IRT, allowing the calculation of reliable individual estimates for the comprehensive scores; and third, being defined as a mean score the interpretation of the estimate is clear and transparent, which is particularly important in high-stakes testing.

A further advantage of the GSM is based on its characteristic to directly yield standardized estimates and posterior distributions for the difference scores, that is, the differences in the achievements in the subdimensions. Often researchers are not interested in the absolute abilities in the subdimensions but in the differences between these. This might be in order to investigate if students differ in their subdimension abilities, to differentiate types of students via their ability profiles, or to consider trends in longitudinal studies (where the subdimensions represent measurements at different points in time). Brandt, Duckor, and Wilson (2014), for example, presented an approach based on the GSM's difference scores in order to investigate the dimensionality of a given test.

Prospect of Current and Future Research

The given definition of the GSM relates to the Rasch Model (Rasch, 1980). However, its extension to a corresponding 2-PL model (Birnbaum, 1968) is straightforward, and an application of a 2-PL variant of the GSM to NAEP data in order to compare the results of the unidimensional scores from the GSM with those based on the mean scores from the multidimensional plausible value estimates will be of great interest. Hitherto, the analysis of large-scale data using the GSM was difficult, though, since a calibration of the model was only possible using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; cf. chapter 5)[5]. Besides the complexity of correctly defining models, estimations in WinBUGS are, even for small data samples, extremely time consuming. Since an update in the R software package TAM (Kiefer, Robitzsch, & Wu, 2015), however, the GSM can be calibrated using TAM and can be estimated as efficient as any standard multidimensional IRT model. Furthermore, TAM also allows the calibration of the GSM including regression models. Future applications of the GSM to large scale assessment data are therefore now unproblematic.

Besides a broader application of the GSM in order to further explore the statistical differences of the unbiased and weighted unidimensional GSM estimates and the standard unidimensional IRT estimates, it is hoped that the GSM might also add to the discussion and determination of the dimensionality of data sets. The opportunity of selecting an IRT model that provides a unidimensional score but does not assume unidimensionality might help test developers in more easily accepting given multidimensionality and allowing for the construction of more multidimensional tests. Furthermore, the reliable difference scores yielded by the GSM might help in guiding the discussion of dimensionality from a decision based on model fit towards a decision based on utility. Typically, the dimensionality of a test is defined via model fit comparisons, even though, often enough, the results are contradicting and depend on the chosen fit criterion. In a utility-based approach a very clear question is asked: Is there a relevant number of persons that actually differ on the given dimensions, so a separate interpretation of the dimensions is useful? Being based on the utility, the approach thereby also yields a direct relation to the validity of the assumed dimensionality (American Educational Research Association et al., 2014; cf. Brandt et al., 2014).

The original article was received on December 27th, 2016
The revised article was received on October 22 nd, 2017
The article was accepted on October 27th, 2017

5 In contrast to the Subdimension Model, a calibration of the GSM using ConQuest is not possible.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1–23.

Allen, N. L., & Carlson, J. E. (1987). Scaling Procedures. In A. E. Beaton (Ed.), The NAEP 1983-1984 technical report. Princeton, NJ: Educational Testing Service.

Allen, N. L., Carlson, J. E., & Donoghue, J. R. (2001). Overview of part II: the analysis of 1998 NAEP data. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), The NAEP 1998 Technical Report (pp. 143–160). Washington, D. C.: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Beaton, A. E. (1987). Implementing the new design: The NAEP 1983-84 technical report. Princeton, NJ: Educational Testing Service.

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

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

Brandt, S. (2006). Exploring bundle dependencies for the embedded attitudinal items in PISA 2006. Presented at the 13th meeting of the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments (Vol. 1, pp. 51–70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. Journal of Applied Measurement, 11, 352–367.

Brandt, S. (2012a). Definition and Classification of a Generalized Subdimension Model. Presented at the 2012 annual conference of the National Council on Measurement in Education (NCME), Vancouver, BC.

Brandt, S. (2012b). Robustness of multidimensional analyses against local item dependence. Psychological Test and Assessment Modeling, 54, 36–53.

Brandt, S. (2016). Unidimensional Interpretation of Multidimensional Tests (Doctoral dissertation). University of Kiel, Kiel. Retrieved from http://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00006439/Brandt_Dissertation_v1.00.pdf

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. Psychological Test and Assessment Modeling, 55, 148–161.

Brandt, S., Duckor, B., & Wilson, M. (2014). A utility-based validation study for the dimensionality of the performance assessment for california teachers. Presented at the 2014 annual conference of the American Educational Research Association (AERA), Philadelphia, PA.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. Applied Psychological Measurement, 33, 620–639.

Donahue, P. L., & Schoeps, T. L. (2001). Assessment frameworks and instruments for the 1998 national and state reading assessments. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), The NAEP 1998 Technical Report (pp. 255–268). Washington, D. C.: National Center for Education Statistics.

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.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: test analysis modules (Version 1.3) [R]. Retrieved from http://cran.r-project.org/package=TAM

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

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs - A Bayesian modelling framework: Concepts, structure, and extensibilty. Statistics and Computing, 10, 325–337.

Martin, M. O., & Mullis, I. V. S. (2012). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. IERI Monograph series–Issues and Methodologies in Large Scale Assessments, 4, 131–158.

Moosbrugger, H., & Kelava, A. (2007). Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]. Heidelberg: Springer Medizin Verlag.

OECD. (2005). PISA 2003 technical report. Paris: OECD.

OECD. (2012a). PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy. OECD Publishing.

OECD. (2012b). PISA 2012 technical report. Paris: OECD.

OECD. (2014). PISA 2012 results: what students know and can do (Vol. 1, revised edition). Paris: Organisation for Economic Co-operation and Development.

OECD, O. for E. (2004). The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills. OECD Publishing.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

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

Rost, J. (1996). Lehrbuch Testtheorie, Testkonstruktion [Textbook test theory, test construction]. Bern; Göttingen; Toronto; Seattle: Verlag Hans Huber.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. Educational and Psychological Measurement, 68, 413–430.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychological Methods, 6, 181–195.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. IERI Monograph Series, 2, 9–36.

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

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement, 29, 126–149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ACER ConQuest: generalized item response modeling software. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125–145.

Yen, W. M. (1993). Scaling performance assessments - strategies for managing local item dependence. Journal of Educational Measurement, 30, 187–213.

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