Estimation of a Rasch model including subdimensions

Steffen Brandt
Leibniz Institute for Science Education, Kiel, Germany

Many achievement tests, particularly in large-scale assessments, deal with measuring abilities that are themselves assumed to be composed of other more specific abilities. A common approach to obtaining all necessary ability estimates, therefore, is to analyze the same data once using a unidimensional model and once using a multidimensional model. This approach not only poses a theoretical contradiction but also means neglecting the assumed local dependencies between the items of the same “specific” ability within the unidimensional model. This paper presents the application of a Rasch subdimension model that explicitly considers local item dependence (LID) due to specific abilities and thereby yields more adequate estimates. In addition to providing a short theoretical presentation of the model, the paper focuses on making it easier for researchers to apply the model. The paper accordingly uses an empirical example to show how results using the subdimension model differ from results arising out of the unidimensional model, the multidimensional model, and the Rasch testlet model (as an alternative model that models LID). It also offers an explicit description of how ConQuest software can be used to define and calibrate the models.

INTRODUCTION

Many of today's achievement tests, in particular those used within large-scale assessments, deal with measuring abilities that are themselves assumed to be composed of other more specific abilities. As such, a common approach that researchers involved in large-scale cross-national assessments such as TIMSS and PISA take when endeavoring to yield the necessary ability estimates is to analyze the same data-set once using a unidimensional model and once using a multidimensional model (cf. Martin, Mullis, & Chrostowski, 2004; Organisation for Economic Co-operation and Development/OECD, 2005). However, this approach has two major downsides. First, from a theoretical point of view, the assumption that the data fit both unidimensional and multidimensional models seems to make model-fit tests obsolete and the application of a particular model somewhat arbitrary—or simply determined by pragmatic needs. Second, and this time from a practical point of view, neglecting the assumed local dependencies among the items of the same subtest (or subdimension) that measure a more specific ability means accepting the negative impacts of local item dependence (LID).

As shown by many authors (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wang & Wilson, 2005a; Yen, 1993), an inappropriate assumption of LID results in an overestimation of test information and reliability and an underestimation of the measurement error. Furthermore, because LID influences item discriminations, items showing LID also show lower discriminations than is the case with items showing no LID (Yen, 1993). Finally, the variance of the estimated parameters decreases for items with LID.

Yen and Thissen and his colleagues have examined these effects of LID through the use of "testlets"—a subset of items in a test that have a common structural element. An example is bundles that have a common stimulus (Wainer & Kiely, 1987). More recently, Wang and Wilson (2005a, 2005b) showed that it is possible to model LID in relation to testlets by using a Rasch testlet model and thereby obtaining more precise and adequate estimates. The Rasch testlet and the Rasch subdimension model proposed below are special cases of the group of so-called bi-factor models. These models are characterized by the fact that each item loads on at least two dimensions, on a general factor, and on one or more group—or method-specific—factors, such that the loading on the general factor is non-zero (Holzinger & Swineford, 1937).

In order to analyze these types of models, Gibbons and Hedeker (1992) developed a full-information item bi-factor analysis for binary item responses. The development of appropriate models and estimation procedures relative to graded response data has, however, been less successful (cf. Muraki & Carlson, 1995). The additional computational complexity associated with graded response data leads to the introduction of additional model constraints in order to estimate the model. One restriction commonly applied involves constraining the method-specific—or group-specific—factors (denoted in the case of the model presented in the following section as the latent traits of the *subdimensions*) so that the factors are independent from the general factor (termed the "main dimension" in the following). While this constraint is appropriate in that the specific factors measure only the residual associations of the items beyond those due to the general latent trait, and although this constraint is a common feature of bi-factor models, the computational complexity seems to make a second model constraint necessary for graded response data. As a consequence, an additional assumption in regard to the Rasch testlet model, as well as in regard to the recently proposed full-information item bi-factor analysis for graded response data (Gibbons et al., 2007), is that the specific factors are also independent of one another.

The model that I propose in this paper tries to loosen this latter—rather strong—constraint through application of a different constraint but one that still allows for correlation of the specific factors. I discuss the possible consequences of these different assumptions in somewhat more detail after defining the model in the next section. I then show how to calibrate the model using the software ConQuest. This is followed by an empirical example that depicts the differences between the unidimensional, the (unrestricted) multidimensional, and the subdimension models.

A RASCH MODEL THAT INCLUDES SUBDIMENSIONS

To resolve the theoretical problem of unidimensionality versus multidimensionality and to reduce negative impacts on measurement precision due to LID, the model proposed here is a Rasch subdimension model (Brandt, 2007a, 2008b). The model extends the standard Rasch model (Rasch, 1980) by using an additional set of parameters for subdimensions, and it is based on the assumption that each person has a general ability in the measured dimension (which in the subdimension model is denoted as the *main* dimension) as well as strengths and weaknesses (to be defined *ex ante*) in the subdimensions that measure specific abilities within the measured main dimension. This way, the model is able to yield person parameters that account for existing LID among the items of the same subdimension. The model is also a special case of the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997), and so can be directly estimated through use of the software ConQuest (Wu, Adams, & Wilson, 1998).

Model Definition

Assuming we have a single measured main dimension (e.g., mathematics) that is composed of a number of defined subdimensions (e.g., differently defined areas of mathematics), and assuming that we can characterize each person's ability in a subdimension according to a strength or weakness relative to his or her ability in the measured main dimension, we end up with three different sorts of parameters to consider when modeling the answers based on a Rasch model approach. The first two sorts of parameters are analogous to the Rasch model, the item parameters b_i (with i=1,...,I and I the total number of items) that describe the item difficulties, and the person parameters θ_v (with v=1,...,V and V the total number of persons) that describe persons' abilities on the measured main dimension. In addition to these parameters, we need parameters γ_{vd} that describe persons' strengths (or weaknesses) in the measured subdimensions. A person's actual ability parameter to solve an item from subdimension *d* (with *d*=1,...,*D* and *D* the total number of subdimensions) is thus defined by

$$\theta_{vd} = \theta_v + \gamma_{vd}. \tag{1}$$

While the parameter θ_v denotes the *overall* ability parameter across subdimensions, the parameter γ_{vd} denotes the *specific* ability parameter for the subdimension. If we use the definition in Equation (1), the probability P_{vi1} of person *v* producing a correct response to a dichotomous item *i* in a Rasch subdimension model is then given as

$$P_{vi1} = \frac{\exp\,(\theta_v + \gamma_{vd(i)} - b_i)}{1 + \exp\,(\theta_v + \gamma_{vd(i)} - b_i)} \tag{2}$$

where $\gamma_{vd(i)} = d(i) \cdot \gamma_{vd}$ and $d(i)$ is equal to 1 when item *i* measures subdimension *d*, and 0 otherwise. To ensure that the parameters have the needed properties, further restrictions of the parameters have to be introduced (cf. Brandt, 2007a, 2008b):

Restriction 1: $\sum_{d=1}^{D} \gamma_{vd} = 0$ for all $v = 1, \ldots, V$ (3)

Restriction 2: $\text{cov}\,(\theta_v, \gamma_{vd}) = 0$ for all $d = 1, \ldots, D$ (4)

Restriction 3: $\sum_{v=1}^{V} \theta_{vd} = 0$ (5)

Restriction 1 is equivalent to $\frac{\sum_{d=1}^{D} \theta_{vd}}{D} = \theta_v$; that is, it assures that θ is the average of persons' absolute abilities in the subdimensions (θ_{vd}). This restriction is essential for correctly identifying the model. Restriction 2, however, is not necessary in this respect; rather, it specifies the composition of the estimate for the main dimension. By constraining all subdimension-specific factors to have the same covariance with the main dimension (namely zero), the subdimensions are defined to be equally weighted for the composition of the main dimension. This practice accords with the common assumption inherent within the bi-factor models described above. It also accords with Humphreys' (1962, 1970, 1981, 1986) recommendation to control differential item functioning or DIF (which arises in the considered case here via the subtests measuring different specific abilities) by balancing across items. Humphreys is supported in this opinion by Wainer, Sireci, and Thissen (1991), who also address the difficulty of this task. Finally, Restriction 3 is one of the common restrictions that ensure correct identification of the model. The shown restriction in this case represents the constraint of the mean of the person parameters of the main dimension to zero. However, as an alternative to this restriction, we can constrain the item parameters to have a mean of zero, or we can anchor one or more of the item parameters.

By using Equation (2), we can also formulate the log-odds form of the subdimension model. This results in

$$\log(p_{v1}/p_{vi0}) = \theta_v + \gamma_{vd(i)} - b_i\,, \tag{6}$$

where p_{vi0} denotes the probability of person v giving an incorrect answer to item *i*, and requires application of Restrictions 1 to 3, described above. Furthermore, the equations stated above for dichotomous items can be extended to

$$\log(p_{vij}/p_{vi\,(j-1)}) = \theta_v + \gamma_{vd(i)} - b_{ij}\,, \tag{7}$$

for polytomous items, where p_{vij} and $p_{vi\,(j-1)}$ are the probabilities of scoring *j* and *j*-1 (where $j = 1,\ldots, K_i\text{-}1$ and K_i is the number of categories for item (*i*) to item *i* for person *v*, respectively, and b_{ij} is the *j*th step difficulty of item *i*. By introducing a parameter b_i, called overall item difficulty, and a parameter τ_{ij}, called *j*th threshold of item *i*, where

$$b\sigma_{ij} = b_i + (b_{ij} - b_i) = b_i + \tau_{ij}\,, \tag{8}$$

we can express Equation (7) as

$$\log(p_{vij}/p_{vi\,(j-1)}) = \theta_v + \gamma_{vd(i)} - (b_i + \tau_{ij})\,, \tag{9}$$

which reduces to the partial credit model (Masters, 1982) when $\gamma_{vd(i)} = 0$. Extending other Rasch models to include a subdimension component, such as the rating scale model (Andrich, 1978) or the linear logistic test model (Fischer, 1973), is straightforward.

By defining weights

$$q_{id} = \frac{u_{id}}{\sum_{d=1}^{D} u_{id+}}\,, \tag{10}$$

where u_{id} is an indicator variable that is 1 if item *i* is within dimension *d* and is zero otherwise, and by inserting Equation (1), we can further express Equation (2) as

$$p_{vi} = \frac{\exp\left(\left(\sum_{d=1}^{D} q_{id}\ \theta_{vd}\right) - b_i\right)}{1 + \exp\left(\left(\sum_{d=1}^{D} q_{id}\ \theta_{vd}\right) - b_i\right)} \tag{11}$$

thereby matching the multidimensional Rasch model (Carstensen, 2000; Rost, 1996). We can, in fact, see the subdimension model as a reparameterized multidimensional model, somewhat similar to Masters' partial credit model, which reparameterizes the Rasch model for polytomous items. Note, however, that in the case of the subdimension model, it is not the item parameters but the person parameters that are reparameterized.

Discussion of the Model

To provide more insight into the subdimension model, I now discuss Restriction 1 and Restriction 2 of the model in more detail.

As I mentioned above, the subdimension model allows for correlations between specific abilities, in contrast to (for example) the Rasch testlet model. Instead, the model incorporates a restriction on the sum of the estimates for the specific abilities (Restriction 1)—a characteristic that can constrain the size of the measured variances for the subdimensions, particularly if the differences in the measured variances are very large. For tests with subdimensions of equal variance, however, it has been shown that the subdimension model provides results equivalent to those of the unrestricted multidimensional model (Brandt, 2007a, 2008b), so allowing the subdimension model to be derived through variable transformation.

This attribute is particularly noteworthy in relation to large-scale assessments such as PISA and TIMSS that utilize detailed background information on the students to impute values for the proficiency variable even though a large portion of item responses are missing due to the matrix-sampling of items administered to each student. Within the calibration process of the person parameter estimates, analysts can use this background information as regression parameters on the estimated latent traits, a process that leaves the residual variances of the latent traits reflecting only those parts of the variances that are not attributable to the regression parameters. This situation, in turn, leads to a decrease in the size of the residual (conditional) variance[1] for the latent traits. It also typically results in variances that are closer to one another in size. Use of the subdimension model can therefore be particularly beneficial in these cases.

An important difference between the subdimension model and other bi-factor models such as the Rasch testlet model is evident in the assumptions each holds about the covariances between specific abilities. To make the resulting differences more obvious, let us consider an example of a science assessment consisting of four testlets, each with five items that refer to a common stimulus, and let us additionally assume that the single measurement of each testlet results in the same variance for the distribution of the measured latent trait. (In other words, the subdimension model will yield results equivalent to the unrestricted multidimensional model.) Let us further assume that the stimuli relate to the following different application areas of science—agriculture, medicine, electronics, and environmental pollution. And then let us suppose, for the purposes of this test, that application of the Rasch testlet model leads to variances of var_{T1} to var_{T4} for the testlet-specific dimension and that application of the subdimension model leads to variances of var_{S1} to var_{S4}.

While, for a given test, we may not find it easy to recognize how these two differently measured variances for the same testlet differ, the difference becomes transparent

[1] The actual variances for the latent trait, including the explained variances due to the regression parameters, are calculated *post hoc*.

when we assume that a fifth testlet has been added to the test and that this testlet has a higher correlation with one of the already existing testlets than with the remaining three, perhaps because the stimulus relates to the same application area of (say) agriculture as Testlet 1 does. As a consequence, that part of the testlet-specific variance var_{T1} attributable to the application area is equivalent to that of the new Testlet 5. However, because the Rasch testlet model assumes these variances are independent, applying the Rasch testlet model to the test that has all five testlets will not model the specific variance attributable to the application area of agriculture because of the need to comply with the independence assumption concerning the testlet-specific effects. In short, the model will not account for LID because of the application area in question. Hence, in the Rasch testlet model, var_{T1} will be smaller in the test with all five testlets than in the test with just four testlets. An analysis of item-bundle effects for the mathematics achievement test of PISA 2003 showed that the size of the testlet-specific variances for the item bundles included in all tests differed to a considerable extent according to whether certain item bundles were included or excluded from the analysis (Brandt, 2006).

In the subdimension model, however, the variance of var_{S1} will be the same in the test with four and five testlets (subdimensions) if the testlets yield variances of equal size. When the testlets do not show equal variances, the model usually becomes less capable of modeling the local dependencies among the items of the same testlet.[2] The measured specific effects are comparatively stable, however, and do not depend on the content of the other subdimensions (testlets) in the test. Rather, they depend solely on the size of the variances of these subdimensions, a situation that essentially is due to a normalization problem. Because the subdimension model does not assume the independence of the subdimension-specific factors, the model is less sparse than the testlet model.

In the Rasch testlet model, only one parameter (the variance of the testlet-specific effects)[3] has to be estimated, but in the subdimension model, the covariances for all other existing subdimensions have to be estimated as well. Therefore, it is possible to calibrate the Rasch testlet model for even large numbers of testlets. The number of parameters to be estimated for the subdimension model, however, increases in the same way as occurs with those within the unrestricted multidimensional model. In fact, for both models, the same numbers of parameters always have to be estimated given that the subdimension model is essentially a variable transform of the unrestricted multidimensional model. Comparison of the unrestricted multidimensional model and the subdimension model shows that the number of estimated dimensions is equivalent in both models (see the definition of the scoring matrix above). So, if we assume that the ability estimates in both models are constrained to a mean of

[2] In certain cases, the subdimension model might be able to yield results equivalent to those of the unrestricted multidimensional model when the variances of the testlets/subdimensions differ (cf. Brandt, 2007a, 2008b).

[3] Because ConQuest uses a marginal maximum likelihood approach for the parameter estimation, and assuming a standard normal distribution for the measured latent traits, only the mean and the variance of the distribution are estimated.

zero for a test with *n* dimensions, we will not need to estimate the parameters for *n*-1 covariances in the subdimension model due to Restriction 2. However, unlike the situation with the unrestricted model, we would have to estimate the *n*-1 additional parameters for the means of the subdimension-specific latent traits. This is because, in the subdimension model, only the ability estimates of the main dimension are constrained to a mean of zero. Thus, the number of parameters to be estimated is always the same for both models.

ESTIMATION USING CONQUEST

Because the MRCMLM includes the subdimension model as a special case (Brandt, 2007a, 2008b), the software ConQuest (Wu et al., 1998) can be used to estimate the model. Although the mathematical definition of the subdimension model given via the MRCMLM is provided in the proof that the subdimension model is a special case of the MRCMLM, it is still necessary to understand the notations used for the definitions of the scoring and design matrices and, furthermore, to be able to define the resulting constraints using ConQuest syntax. Given the complexity of the MRCMLM notation, as well as the need for knowledge about ConQuest, models like those described above, and also the Rasch testlet model (Wang & Wilson, 2005b), are barely accessible to people interested only in applying adequate models to their data and who are less interested in understanding the theoretical definitions and concepts of particular models. Therefore, a main goal of this paper is to fill in this gap relative to the subdimension model and to give a detailed description for calibrating it. To do this, I begin by briefly describing the different given ways of defining or constraining a model via ConQuest.

Basically, ConQuest offers five different ways of defining or constraining a specific model within the MRCMLM:

1. Through the definition of the design matrix, which describes the linear relationship among the items;
2. Through the definition of a scoring matrix, which assigns the items to specific ability dimensions and assigns scores to their response categories;
3. Through the anchoring of the item-difficulty parameters, which can be used not only for linking to other tests but also for identification purposes;
4. Through specification of mean abilities for the population distribution[4] (within ConQuest denoted as regression parameters); and
5. Through anchoring of the variance–covariance matrix.

Although the definition of the scoring matrix is embedded in ConQuest's command language, the remaining four types of specifications are done via imported text files.

[4] Because ConQuest uses a marginal maximum likelihood estimation method, the ability distributions for a given population are assumed to be normal.

For standard unidimensional or multidimensional calibrations, ConQuest's command language provides the means by which the analyst can automatically generate the appropriate design matrices and anchorages. It is through the command `model` that the simple Rasch model (`model item;`), the rating scale model (`model item + step;`), the partial credit model (`model item + item*step;`), and other multifaceted models can be defined, and it is through the set constraint command that identification of the model can be set to `items` (i.e., the mean of the item difficulties is set at zero) or `cases` (i.e., the mean of the ability distribution is set at zero).

On completion of other necessary commands concerning the data file to be calibrated and the output files that are to be generated, a ConQuest command file for the calibration of a unidimensional partial credit model looks like this:

```
datafile estimation.dat;       /* Definition of the data file with
                               the students' answers */

format responses 1-40;         /* Columns in the data file that
                               represent the students' answers
                               */

codes 0,1,2;                   /* Definition of valid answer
                               codes–all other codes will be
                               interpreted as missing by design
                               */

score (0,1,2) (0,1,2)!items    /* Definition of the scoring
(1-40);                        matrix (here, according to a
                               unidimensional model)*/
model item + item*step;        /* Definition of the design matrix
                               (here, according to the partial
                               credit model) */

set constraints = items;       /* Constraint for the identification
                               of the model */

export designmatrix >>         /* Export of the design matrix
estimation.dsm;                to a data file with the given name
                               */

estimate;                      /* Start of the calibration using
                               the standard settings /*

show >> estimation.shw;        /* Export of the calibration
                               results to a data file with the
                               given name  */
```

This example assumes that the answer data provided by the data file has already been scored and that a single digit represents each coded answer. Thus, each column represents the students' answers to a particular item, scored with 0, 1, or 2 credits (cf. also the **code** statement above).[5] In the unidimensional case, the scoring matrix reduces to a simple vector (with 40 elements) that maps all scores on all items to the same dimension.

In regard to the definition of the design matrix via the **model** command, note that if the model is constrained to have a mean item difficulty of zero (as in the above case), the design matrix will have to be changed accordingly. Therefore, the design matrix will not be generated until the start of the calibration in order to comply with the given **set constraint** command. As for the unidimensional calibration conducted by the above command file, the **export designmatrix** command is not necessary. Nevertheless, this statement is valuable here because the design matrix generated for the calibration is exactly the design matrix needed in order to define the subdimension model presented below.[6] Finally, the **estimate** command starts the calibration with the standard algorithm and convergence criteria of ConQuest, and the show command generates a standard output for the results of the calibration, written to a text file named "estimation.shw".

Defining models like the subdimension model requires somewhat more effort since there is no ConQuest command to automatically generate and set the necessary constraints according to the model definitions. Instead, this has to be done manually by providing appropriate import files. The main focus of this section, therefore, is to describe the construction and definition of these import files as well as the definition of the specific scoring matrix needed for the model.

```
datafile estimation.dat;     /* See above */
format responses 1-40;       /* See above */
codes 0,1,2;                 /* See above */

score (0,1,2) (0,1,2) (0,1,2) () () !items (1-10);
score (0,1,2) (0,1,2) () (0,1,2) () !items (11-20);
score (0,1,2) (0,1,2) () () (0,1,2) !items (21-30);
score (0,1,2) (0,1,2) (0,-1,-2)(0,-1,-2) (0,-1,-2)!items (31-40);
                             /* Definition of the scoring matrix */

model item + item*step;      /*  Pseudo-definition  of  the  design
                             matrix */

import designmatrix <<       /* Actual definition of the design
estimation.dsm;              matrix */

import anchor_covariance     /* Setting of the constraints for
<< estimation.cov;           the variance-covariance matrix */

estimate !method=             /* Start of the calibration using a
montecarlo,nodes=2000;       Monte  Carlo  method  with  2000  nodes
                             and standard convergence criteria */
```

[5] ConQuest also provides a way of scoring the data via the command language; more information about these commands can be found in the ConQuest manual.

[6] If the design matrix has not been generated at the time the command is processed, ConQuest exports the file as soon as the design matrix is generated; that is, after the start of the calibration.

```
show >> estimation.shw;     /* Export of the calibration results
                            to a data file with the given name
                            */
```

The first three commands of the command file correspond to those of the unidimensional calibration above; that is, the same data-set as the one above is calibrated. Here, it is assumed that the test includes four subtests with 10 items each, with Items 1 to 10 referring to Subtest 1, Items 11 to 20 to Subtest 2, Items 21 to 30 to Subtest 3, and Items 31 to 40 to Subtest 4. In order to account for the assumed local dependencies between the items of the same subtest, the subdimension model is used for estimation. As the definition of the scoring matrix above shows, the subdimension model is a multidimensional model; in the above example, it has four dimensions. The first dimension, comparable to the unidimensional case above, refers to the unidimensional latent trait that all 40 items commonly measure. The second and fourth dimensions, however, refer to the specific parameters of the subdimensions that are to be estimated. According to Restriction 1 of the definition of the subdimension model, the subdimension-specific parameters must add up to zero for each student. In order to comply with this restriction, the parameter estimates of the fourth subdimension cannot directly be estimated but rather defined as constrained parameters. When the sum of the four subdimension-specific parameters is zero, then each person's fourth parameter (or any single other of the four) always equals the negative of the sum of the other three parameters. What this means, in essence, is that the subdimension model actually contains only *d-1*-estimated *specific* dimensions, and one final specific dimension, which is totally determined by the negative sum of the previous *d-1*. Therefore, Items 1 to 10 load (in addition to the main dimension) on Dimension 2, Items 11 to 20 load on Dimension 3, Items 31 to 40 load on Dimension 3, and Items 31 to 40 load negatively on Dimensions 2 to 4.

The model statement follows the process involved in defining the scoring matrix. This statement has only a dummy function, which exists for programming reasons, given that ConQuest's `estimate` command must always be preceded by a `model` command. The design matrix generated according to this standard statement cannot be used because it is problematic in two ways. First, for items that load on more than one dimension, ConQuest adjusts the design matrix in order to keep the difficulty estimates of the items in proportion to the size of the ability estimates. In the case of the subdimension model, this step simply results in estimates that are exactly half the size of the unidimensional estimates, thereby making comparisons just that little bit more difficult. Secondly, ConQuest does not adjust the design matrix according to the necessary constraint of the item parameters needed for the subdimension model. The needed constraint is correctly defined, though, in the design matrix generated for the corresponding unidimensional calibration. Furthermore, by using this design matrix, the software renders the parameter estimates of the calibration for the subdimension model comparable to those of the unidimensional model, and it does this without any further linear transformation. Therefore, the easiest and (probably) least error-inducing way to obtain the correct design matrix is to generate it with the corresponding unidimensional model, as shown in the example above.

Once the correct design matrix is imported, all that remains to do is correctly anchor the variance–covariance matrix according to Restriction 2 of the model. This step requires creation and importation of an appropriate text file. For the above example, the import file "estimation.cov" has the following format:

```
1        2     0.0000
1        3     0.0000
1        4     0.0000
```

The first two figures in a row define which covariance is to be set. Thus, in the first row above (the covariance of Dimensions 1 and 2), the third figure sets the value for the given covariance. Here, all listed covariances are set at zero, a practice that aligns with the definition of the subdimension model, which requires the covariances between the main dimension and the subdimensions to be constrained to zero.

The empirical example presented in the next section was calibrated using ConQuest, as described in this section.

AN EMPIRICAL EXAMPLE

The empirical example given here is based on data taken from the mathematics achievement test used for TIMSS 2003 (Mullis, Martin, Gonzales, & Chrostowski, 2004; Mullis et al., 2003). This test was developed according to two different aspects—a content domain and a cognitive domain. While the latter domain consisted of *knowing facts and procedures, using concepts, solving routine problems,* and *reasoning*, the analysis presented in the following refers to the five defined content domains, which were *number, algebra, measurement, geometry*, and *data*. To select appropriate items for the main study, the TIMSS researchers conducted a full-scale field trial. They then used the results of this trial to determine which items would be used in the main study. During this selection process, the researchers took care not only to distribute the items across the four cognitive and five content domains according to the proportions defined in the assessment framework but also to ensure that the psychometric characteristics of the items were sufficient, particularly in relation to DIF effects and discrimination power (Martin et al., 2004).

Because the psychometric criteria chosen refer to a unidimensional analysis of the data, we can consider the test to have been constructed as multidimensional from a qualitative point of view, via the *ex ante* defined domains, and as unidimensional from a quantitative measurement point of view. The described test construction displays the dilemma of TIMSS and other large-scale assessments associated with lack of appropriate models (cf. the test construction for the PISA study, for example; OECD, 2005). The resulting data-sets, therefore, are not good examples of true multidimensionality. Despite this, the assessment results are publicly reported and interpreted. With these considerations in mind, the following analysis shows the extent to which the subdimension model can still help provide more appropriate measures by modeling the five content domains defined for the mathematics test.

Data and Analysis

The analyzed test used data obtained from the United States sub-sample of students who participated in TIMSS 2003. This sub-sample consisted of 8,912 students in total, and the test included 194 mathematics items: 47 items for the content domain *algebra*, 28 for *data*, 32 items for *geometry*, 31 for *measurement*, and 56 for *number*. Nineteen of the 194 items were partial-credit items, each with three score categories. In order to compare and discuss the results obtained via the subdimension model (more precisely its extension to the partial credit model), I also analyzed the data using the unidimensional model, the testlet model, and the (unrestricted) multidimensional model.

Results and Discussion

Table 1 summarizes the results of the estimated means[7] and variances for the distributions, their reliabilities, the correlations, and the -2 log likelihoods for the different models. The index M (main dimension) refers to the unidimensional latent trait; the indices 1 to 5 refer to the content domains algebra, data, geometry, measurement, and number, respectively.

On comparing the variance obtained for the main dimension of the subdimension model with the variance obtained via the unidimensional model, we find that the actual variance is underestimated in the unidimensional case because of the local dependencies of the items of the same content domain. Although the test was constructed to be unidimensional, the subdimension model shows an increase in measured variance. The variance rises from 1.19 to 1.25, which is equivalent to an increase of about 5%. The increase in variance accords with findings by other authors (e.g., Sireci et al., 1991; Wang & Wilson, 2005b; Yen, 1993). If we look at the given reliabilities for the main dimensions, it becomes even clearer that the reliability given for the ability estimates is overestimated in the unidimensional case. Despite the subdimension model allowing for a gain in measured variance, the given reliability of its estimates is still lower than that of the unidimensional estimates. Essentially, the true reliability of the ability estimates calculated via the unidimensional model is *smaller* than that given for the main dimension of the subdimension model.

A difference between the multidimensional model and the subdimension model that becomes apparent on looking at the results is that the absolute variances of the latent traits measured by the subtests are closer to one another when the subdimension model is used than when the multidimensional model is used. While use of the multidimensional model shows subtest variances ranging from 0.86 to 1.74, the (absolute) variances obtained using the subdimension model range from only 1.35 to 1.40. This difference reflects the inability of the subdimension model to fully model the differences between the subtests due to their different variances. The estimated likelihoods of the two models provide a further indication of the extent to which the

[7] The means and correlations for the testlet model given in Table 1 are not estimated but instead display the anchor values of the parameters; the testlet model is constrained on the cases, given that this constraint is the only one that yields an optimum model fit.

Table 1: Results of the reanalysis of the US TIMSS 2003 mathematics achievement test

	Unidim.		Testlet		Subdim.		Multidim.	
Parameter	*Estimate*	*Reliability*	*Estimate*	*Reliability*	*Estimate*	*Reliability*	*Estimate*	*Reliability*
σ^2_M	1.19	0.820	1.22	0.812	1.25	0.816		
$\sigma^2_1 / \sigma^2_{S1}$			0.21	0.141	0.14	0.148	1.45	0.767
$\sigma^2_2 / \sigma^2_{S2}$			0.19	0.103	0.13	0.113	1.74	0.757
$\sigma^2_3 / \sigma^2_{S3}$			0.15	0.096	0.15	0.145	0.86	0.722
$\sigma^2_4 / \sigma^2_{S4}$			0.08	0.053	0.10	0.107	1.48	0.781
$\sigma^2_5 / \sigma^2_{S5}$*			0.07	0.062	0.05		1.41	0.800
μ_M	0.02		0.00		0.00			
μ_1 / μ_{S1}			0.00		0.12		-0.04	
μ_2 / μ_{S2}			0.00		0.29		0.43	
μ_3 / μ_{S3}			0.00		-0.29		-0.15	
μ_4 / μ_{S4}			0.00		-0.20		-0.29	
μ_5 / μ_{S5}*			0.00		0.09		0.12	
r_{12} / r_{S12}			0.00		-0.32		0.88	
r_{13} / r_{S13}			0.00		-0.26		0.85	
r_{14} / r_{S14}			0.00		-0.49		0.87	
r_{15} / r_{S15}*			0.00		-0.03		0.92	
r_{23} / r_{S23}			0.00		-0.36		0.84	
r_{24} / r_{S24}			0.00		-0.26		0.90	
r_{25} / r_{S25}*			0.00		-0.14		0.90	
r_{34} / r_{S34}			0.00		-0.14		0.90	
r_{35} / r_{S35}*			0.00		-0.51		0.89	
r_{45} / r_{S45}*			0.00		0.09		0.95	
Estimated Param.	214		219		228		228	
-2 Log Likelihood	275738.4		275380.2		275311.4		275022.6	

Note: * Calculated via plausible values.

subdimension model is capable of modeling the differences between the subtests. The likelihood deviances (-2 log likelihood) of using the multidimensional and the subdimension models are 275,022.6 and 275,311.4, respectively. The likelihood deviance for the unidimensional model, however, is 275,738.4; its difference of just 715.8 within the multidimensional model reflects the unidimensional construction of the measure. Nevertheless, the subdimension model does close the gap between the unidimensional and the multidimensional by about 50%.

Besides the differences in measurement precision and model fit, the interpretational differences of the measures provided by the two models should be of particular interest to test developers and analysts. In the case of the subdimension model, it is not the reliabilities of the *total* subtests that are measured but the reliabilities of the differences *between* the subtests. This is particularly interesting if the tests are being used to analyze, for example, student profiles constructed via the subtests.

Here, the reliabilities provide a measure of how reliable differentiating these students according to these profiles will be and so help develop tests that provide especially reliable measures in these terms.

For the given empirical example, the results with subtest-specific reliabilities ranging from 0.107 to 0.148 indicate that an interpretation of the subtest-specific variances—that is, of differences between the subtests—need to be interpreted with caution. On the other hand, the correlation estimates for the subtest-specific variances provided by the subdimension model bring greater transparency to the differences between the subtests. In the multidimensional model, the large proportion of common variance dominates the correlation estimates, and these differ, at most, by 0.11 (from 0.84 to 0.95), and the estimated correlations for the subdimension model differ by up to 0.60 (from -0.51 to 0.09). Nevertheless, the interpretation of the usually negative correlations provided by the subdimension model (resulting from the applied constraint) is not as intuitive. This is because a correlation of close to 0 for the relative subdimension-specific parameters is usually equivalent to a very high correlation of the corresponding absolute ability estimates. An example of this relationship is provided via Dimensions 4 and 5 above. Although their estimated correlation in the multidimensional model is given as 0.95, the corresponding correlation in the subdimension model is 0.09. This example is a very unusual case of positive correlation. Moreover, when compared with the other subtest correlations within the test, it represents a particularly high correlation between the two dimensions.

As a further comparison, and in order to show other differences, I also applied the testlet model to the data. The comparison of the likelihood deviances showed that, even given the very unfavorable conditions for the subdimension model due to the large difference between the smallest and the largest estimated variances, the model under discussion outperformed the testlet model. The difficulties for the testlet model to appropriately model the given data are best displayed by the relationship between Dimensions 4 and 5. As the results of the multidimensional analysis show, the correlation between these two dimensions is, on average, over 0.05 higher than the correlations between the remaining dimensions. While the subdimension model allows for any specific variance the dimensions have in common, the testlet model constrains the covariance of the respective testlet dimensions to zero. In other words, the large common part of their specific variances is not modeled and, in turn, the modeled variance is comparatively small; in the above example, it is less than half that modeled for the other dimensions.

In summary, the results of the reanalysis show that the application of the subdimension model allows for an increase in measurement precision for the students' unidimensional parameter estimates despite the very unfavorable conditions. Furthermore, the above results indicate that, for the analyses conducted above, the parameter estimates from the multidimensional model yield higher measurement precision for researchers endeavoring to interpret a person's abilities relative to the subtests.

CONCLUSION

The subdimension model offers test developers and analysts a way of handling the common conflict between theory and practice that arises whenever both unidimensional and multidimensional ability estimates of the same test are needed. Hitherto, tests were usually constructed in a unidimensional manner even if they included subtests that supposedly incorporated different characteristics. This practice meant that expected differences between these subtests due to test construction were minimized. Thus, any items particularly adept at showing differences between the subtests would probably not comply with the (unidimensional) psychometric criteria used within the selection process after field trial of the items. Therefore, in order to gain interpretational value for the analysis of the subtests, psychometric criteria need to be based on a model that explicitly accounts for the differences in the subtests. The subdimension model provides exactly this opportunity. By allowing for correlations between the subtest-specific factors, the model is particularly effective in accounting for differences and is able to outperform more restrictive models, like the testlet model (see discussions above).

Due to the restriction of the subdimension-specific parameters to yield a mean of zero, the correlations obtained under the subdimension model cannot be compared directly with those of the multidimensional model. For tests with large differences in subtest variances, this restriction also hinders the ability of the subdimension model to model, to full extent, the LID brought about by the different subtests. The advantages of the model become particularly apparent, however, when the variances of the measured subdimensions are approximately equal. In these cases, the subdimension model yields results almost equivalent to those of the unrestricted multidimensional model. With large-scale assessment studies that use matrix-sampling for administering the items and detailed background information for estimating person parameters, the chances of obtaining favorable conditions for the subdimension model are particularly high. Additionally, and/or in other cases, it might also be possible to provide more favorable conditions by adjusting the subtests for the differences in variances apparent after the field trial and by, for example, using different numbers of items for each subtest.

Another benefit of the subdimension model becomes apparent in regard to large-scale assessments when the matrix sampling for items is used. In these cases, each student receives only one booklet containing a subset of items, which means that several different booklets are needed to administer all items. (TIMSS 2003 used 12 different booklets.) The construction of these booklets typically endeavors to link the items that measure the same construct and to balance item-difficulty differences due to positional effects. An additional balancing of the booklets according to the number of items from the same subtest is usually not feasible. In this instance, the various booklets frequently end up including more items of a particular subtest and fewer of another. A student who performs particularly well in one subtest and poorly in another will effectively get different scores for the overall test depending on the booklet he or she completed. More specifically, this is because the common unidimensional Rasch model does not account for differences in sets of items due to different subtests.

The subdimension model, however, accounts for these differences, and thereby yields more adequate individual measures. Although researchers conducting large-scale assessments are usually not interested in achievement scores for single students, estimation of adequate ability estimates for single students is important because the calculation of adequate correlations (e.g., between a person's achievement score and his or her socioeconomic background) depends on adequate scores at the single-person level.

In addition to its ability to analyze tests measuring a general domain and multiple sub-domains at the same time, the subdimension model seems to provide benefits for other applications. The application for vertical scaling, for example, is straightforward, with the subtests representing tests given at different points in time. However, future research in this area needs to investigate how results arising from use of the subdimension model relate to other models used for vertical scaling. Furthermore, and beyond its application to empirical data, the subdimension model could be usefully employed in simulation studies because of its ability to provide additional and more subtle information, as some of my recent work shows (Brandt, 2007b, 2008a).

Finally, another way of using the subdimension model could be to adjust Restriction 2 of the model so that the measured subtests are not balanced within the overall measure but instead are "assigned" (per definition) more weight—or relevance—than others, which means the characteristics of such models would have to be investigated as well. By providing a detailed description on how to calibrate the subdimension model using ConQuest, I hope that the gap between the development of new models and their application in practice becomes somewhat smaller and that a larger community than at present finds conducting research and practice via a model like the subdimension model a considerably more accessible proposition.

References

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Brandt, S. (2006). *Exploring bundle dependencies for the embedded attitudinal items in PISA 2006*. Paper presented at the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2007a). *Applications of a Rasch model with subdimensions.* Paper presented at the 2007 annual conference of the American Educational Research Association (AERA), Chicago, IL.

Brandt, S. (2007b). *Item bundles with items relating to different subtests and their influence on subtests' measurement characteristics*. Paper presented at the 2007 annual conference of the American Educational Research Association (AERA), Chicago, IL.

Brandt, S. (2008a). *The impact of local-item dependence on multidimensional analyses.* Paper submitted for publication.

Brandt, S. (2008b). *Modeling tests with subtests.* Paper submitted for publication.

Carstensen, C. H. (2000). *Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik (Multidimensional test models with applications from educational and psychological diagnostics).* Kiel: Leibniz Institute for Science Education (IPN).

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement, 31*(4), 4–19.

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.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 22–32). Seattle, WA: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Social Psychology, 71*, 327–333.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report.* Chestnut Hill, MA: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mullis, I. V., Martin, M. O., Gonzales, E. J., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 international mathematics report*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzales, E. J., et al. (2003). *TIMSS assessment frameworks and specifications 2003* (2nd ed.). Chestnut Hill, MA: Boston College.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73–90.

Organisation for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris: Author.

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

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

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247–260.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–202.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197–219.

Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296–318.

Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.