Educational and Psychological
Measurement
71(3) 503–522
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164410381521
http://epm.sagepub.com
SAGE

Hypothetical Use of Multidimensional Adaptive Testing for the Assessment of Student Achievement in the Programme for International Student Assessment

Andreas Frey[1] and Nicki-Nils Seitz[1]

Abstract

The usefulness of multidimensional adaptive testing (MAT) for the assessment of student literacy in the Programme for International Student Assessment (PISA) was examined within a real data simulation study. The responses of $N = 14{,}624$ students who participated in the PISA assessments of the years 2000, 2003, and 2006 in Germany were used to simulate MAT with different restrictions (unrestricted, treatment of link items, treatment of open items, content balance, unitwise item selection, all restrictions). Compared with conventional testing based on the booklet design of PISA 2006, unrestricted MAT increases measurement efficiency by 74% and reduces the average number of presented items from 55 to 26 without a loss in measurement precision. The incorporation of restrictions reduces the advantages of MAT. MAT is recommended for the assessment of newly introduced constructs but not for the assessment of the literacy domains in PISA.

Keywords

computerized adaptive testing, multidimensional adaptive testing, item response theory, Programme for International Student Assessment, large-scale assessments

[1]Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany

Corresponding Author:
Andreas Frey, Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstrasse 62, 24098 Kiel, Germany
Email: frey@ipn.uni-kiel.de

The Programme for International Student Assessment (PISA) is an international large-scale assessment of student achievement developed jointly by member countries of the Organisation for Economic Cooperation and Development (OECD; (http://www.pisa.oecd.org). The objective of PISA is to assess the degree to which 15-year-old students have acquired skills and knowledge essential for successful participation in the modern knowledge society. Beginning in the year 2000, PISA is conducted every 3 years. The study focuses on measuring literacy in the three domains of reading, mathematics, and science (see OECD, 2006). In every assessment, one of the domains is treated as the major domain and is analyzed in depth (PISA 2000: reading, PISA 2003: mathematics, PISA 2006: science, PISA 2009: reading). The cognitive tests measuring the three literacy domains are administered in the paper-and-pencil format. PISA results have received a lot of attention and have often stimulated intensive and productive discussions about the effectiveness of educational systems. However, the valuable results come at a rather high price since large sample sizes of around 4,500 to 10,000 students are tested in each country. Moreover, the tests are rather time-consuming and require 120 minutes of testing time per student for the cognitive items and an overall testing time of about 220 minutes per student.

All in all, the testing load associated with PISA is high, thereby resulting in high costs. In the long run—especially if other large-scale assessments and tests are carried out at the same schools within a short time period—the willingness of schools and teachers to participate in PISA may decrease. For the students, long testing sessions may have a negative impact on their test-taking motivation. Thus, to ensure the cooperation of schools, teachers, and students in the long term and to limit costs, possibilities for increasing the efficiency of the testing procedures while maintaining the high level of precision and interpretability of the results should be examined. A promising testing procedure for achieving these objectives lies in computerized adaptive testing (CAT).

This article describes a real data simulation study that examines the usefulness of multidimensional adaptive testing (MAT) for the assessment of student achievement in PISA. The text is organized as follows: First, the concept of unidimensional CAT and its generalization to MAT are described. Subsequently, restrictions associated with the PISA assessments are depicted and the research questions are stated. Then, the method and the results of the simulation study are presented. The final section covers the implications of the results regarding the use of MAT in PISA.

Computerized Adaptive Testing

CAT is a special approach to the assessment of latent abilities in which the selection of the test items presented to the examinee is based on the responses given by the examinee to previously administered items. The aim of this selection procedure is to tailor the item presentation to the ability level of the examinee in order to maximize the information drawn from each response.

The main advantage of CAT is its capacity to substantially increase measurement efficiency. Since measurement efficiency is defined by the ratio of measurement precision to test length (Frey & Seitz, 2009; Segall, 2005), this gain in efficiency can be used either for reducing the number of items presented to an examinee or for increasing measurement precision if the number of items is held constant for all examinees. Compared to a conventional test with a fixed number of items in a fixed order (fixed item test, FIT), the number of items can typically be reduced by approximately half when CAT is used, without a loss in measurement precision (e.g. Frey, 2007; Segall, 2005).

Most CATs use item pools that have been calibrated with a unidimensional item response theory (IRT) model (e.g. van der Linden & Hambleton, 1997). Frequently applied unidimensional IRT models are logistic test models with either one, two or three parameters characterizing the items of a test. A general model is given by the three parameter logistic test model (3PL), which describes the probability of person j giving a correct answer $U_{ij} = 1$ to item i as a logistic function of the latent ability θ_j of the person, an item difficulty parameter b_i, an item discrimination parameter a_i, and an item-specific pseudo-guessing parameter c_i:

$$P(U_{ij} = 1|\theta_i, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\mathrm{e}^{a_i(\theta_j - b_i)}}{1 + \mathrm{e}^{a_i(\theta_j - b_i)}}. \tag{1}$$

The two-parameter logistic test model (2PL) and the one-parameter logistic test model (1PL or Rasch model) are special cases of the 3PL. The 2PL is derived from the 3PL if the pseudo-guessing parameter c_i is set to 0 for all I items of the item pool, c_1, c_2,..., $c_I = 0$. The 1PL in turn, is derived from the 2PL if the discrimination parameter a_i is constrained to be equal across items.

Multidimensional Adaptive Testing

Although unidimensional CAT has proven beneficial in many simulation studies and empirical applications, large-scale assessments of student achievement such as PISA include additional requirements stemming from complex theoretical underpinnings that cannot be tackled optimally with a unidimensional approach. The theoretical frameworks of these studies often conceptualize multidimensional constructs or multiple unidimensional constructs (e.g., reading literacy, mathematical literacy, and scientific literacy in PISA). To reflect this theoretical complexity directly within the measurement process, MAT can be applied. In MAT, multidimensional item response theory (MIRT; e.g., Reckase, 2009) models are used as measurement models. Similar to unidimensional IRT, a general form of frequently used MIRT models is given by the multidimensional three-parameter logistic test model (M3PL). The M3PL specifies the probability of a person j to correctly answer an item i as a function of m latent abilities of the person, $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_m$, an $1 \times m$ item discrimination vector $\mathbf{a}'_i$, an item difficulty parameter b_i, an item-specific pseudo-guessing parameter c_i, and an $m \times 1 -$ vector $\mathbf{1}$, consisting of 1s expanding the item difficulty to the multidimensional space:

$$P(U_{ij} = 1|\boldsymbol{\theta}_j, \mathbf{a}'_i, b_i, c_i) = c_i + (1 - c_i)\frac{\mathrm{e}^{\mathbf{a}'_i(\boldsymbol{\theta}_j - b_i\mathbf{1})}}{1 + \mathrm{e}^{\mathbf{a}'_i(\boldsymbol{\theta}_j - b_i\mathbf{1})}}. \tag{2}$$

The multidimensional two-parameter logistic test model (M2PL) is derived from the M3PL if c_i is set to 0 for all items, and the item discrimination vector $\mathbf{a}_i'$ is estimated from the data. However, within the area of large-scale assessments of student achievement, it is more common to fix the elements of $\mathbf{a}_i'$ to the values 0 or 1 to define item loadings on one (within-item multidimensionality) or more (between-item multidimensionality) measured dimensions (see Hartig & Hoehler, 2008, for a distinction between within-item multidimensionality and between-item multidimensionality). Models with fixed values of $\mathbf{a}_i'$ are also called multidimensional Rasch-models (Reckase, 2009). Especially the multidimensional Rasch-model with between-item multidimensionality is very useful for large-scale assessments of student achievement as it allows for unequivocal interpretations of the measured dimensions. Although many other MIRT models can be used for MAT besides those mentioned, applications of more complex models are still rare. One exception is given in Segall (2001), describing the use of a hierarchical MIRT model with three levels comprising seven dimensions in a MAT framework.

The two major MAT approaches were introduced by Segall (1996), who describes a Bayesian as well as a maximum likelihood approach, and by van der Linden (1999), who uses maximum likelihood for item selection and ability estimation. In both, the item parameters are assumed to be known. The Bayesian approach of Segall (1996) is especially appealing since—besides providing the possibility to reflect on multidimensional theoretical assumptions in terms of the measurement model—even higher measurement efficiency than in CAT is achieved if correlated dimensions are measured. The increase in measurement efficiency is caused by using knowledge about the multivariate prior distribution of the measured dimensions to optimize both the estimation of the latent ability vector, and the item selection process. For the estimation of the latent abilities, Segall (1996) proposes a multidimensional Bayes modal estimate using Fisher scoring and prior knowledge of the variance–covariance matrix $\mathbf{\Phi}$. Regarding item selection, he suggests selecting the item for presentation that maximizes the quantity

$$|\mathbf{W}_{t+i^*}| = \left|\mathrm{I}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j\right) + \mathrm{I}(\boldsymbol{\theta}, u_{i*}) + \mathbf{\Phi}^{-1}\right|. \qquad (3)$$

Thus, the item i^* is selected, which results in the largest determinant of the matrix $\mathbf{W}_{t+i^*}$, based on the information matrix of the previously t administered items: $\mathbf{I}\left(\theta, \hat{\theta}_j\right)$, the information matrix of a response u_{i*} to item i^*: $\mathbf{I}(\theta, u_{i*})$, and the inverse of the variance–covariance matrix of the prior distribution of the measured dimensions $\mathbf{\Phi}^{-1}$. This item provides the largest decrement in the volume of the credibility ellipsoid around the vector of latent abilities $\hat{\boldsymbol{\theta}}_j$.

The sketched Bayesian MAT approach showed very high measurement efficiency in simulation studies. Segall (1996), for example, compared Computerized Adaptive Testing and MAT using *measurement precision* as well as the *number of items* as termination criteria with existing item parameters of the CAT testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB; Sands, Waters, &

McBride, 1997). For each of the nine content areas covered by the CAT-ASVAB, one dimension was specified within the multidimensional model given by Equation (2).[1] MAT was shown to require about one third fewer items than CAT to reach an equal or a higher reliability. If the number of items was used as termination criterion and, therefore, held constant between MAT and CAT for most of the nine content areas, substantially higher reliability coefficients were observed for MAT than for CAT.

The results of this initial study were underlined by the extensive simulation study of Wang and Chen (2004). They examined the effect of the five independent variables *testing algorithm* (Segall's Bayesian MAT approach, CAT, random item selection), *multidimensionality* (between-item, within-item), *scoring mode* (dichotomous, polytomous), *correlation between measured dimensions* (moderate, high), and the *number of administered items* on the dependent variable *measurement efficiency*. Generally, MAT proved to be considerably more efficient than CAT, which in turn showed much higher measurement efficiency than a random item selection. For a typical large-scale assessment situation with five highly correlated dimensions, a dichotomous scoring mode and a test reliability of .90, MAT needed only 48% of the number of items selected in CAT, and 31% of the number of items needed within the random item selection condition. The authors concluded that the higher the correlation between measured dimensions, the more categories the scoring mode has, and the less reliability is requested, the more efficient MAT will be compared with CAT and the random item selection.

The results of the simulation study of Frey and Seitz (2010) confirmed these findings. They compared the effect of the three independent variables *testing algorithm* (Segall's Bayesian MAT approach, CAT, FIT), *number of measured dimensions* (2, 3, 4, 5) and *correlation between measured dimensions* (.00, .50, .85) relating to *measurement efficiency*. Measurement efficiency was calculated by the ratio of the inverse of the mean squared error to the number of items used. CAT and MAT were specified as flexible length tests taking the mean standard errors of the person parameter estimates for the single dimensions that were observed in the FIT condition as termination criteria. The authors report that the advantage of MAT compared with CAT and FIT regarding measurement efficiency strongly depends on the magnitude of the correlation between the measured dimensions. Under conditions typical for large-scale assessments (five dimensions, mutual correlation of .85), the measurement efficiency for MAT (0.87) was 27% higher than for CAT (0.68) and 273% higher than for FIT (0.23).

The gains in measurement efficiency found in simulation studies propose MAT as a testing algorithm that has great potential to substantially increase the measurement efficiency of PISA. This may be used to cut costs or to gather additional or more precise information in the same amount of time.

Restrictions in PISA

Although the results from simulation studies are very promising for MAT, they cannot be directly transferred to an application in PISA for two major reasons. First, large item pools well suited for MAT are used in the simulation studies. These item pools include a large number of items covering a broad difficulty range. This enables MAT to work at, or at least close to, its maximum performance level over a broad ability range. The PISA item pool is less optimal for MAT. The number of items is comparably small, the majority of items have a medium item difficulty, and only some items have very high or very low item difficulties. Thus, MAT based on the PISA items will most likely work below its optimum performance level in case of extreme abilities. The second—and even more important—reason is PISA's association with a couple of restrictions. The most prominent restrictions that need to be taken into account are as follows:

- A total of 54% of the cognitive items used in PISA 2006 are so-called link items that had already been used in previous PISA assessments. The link items are presented to the assessed student sample with a fixed relative frequency. This allows a stable linking of the literacy scales over different assessments. An unrestricted adaptive algorithm may not result in the desired relative frequencies of presented link items and, therefore, may jeopardize trend reporting.
- The item pool used in the PISA assessments from 2000 to 2006 contains 49% items in *open response* or *short response* format. Many items of these types cannot be directly scored by a computer. Hence, the responses given cannot be used within an adaptive testing procedure to revise the provisional ability vector.
- Only 13% of the items used in the PISA assessments from 2000 to 2006 are single items. All other items are grouped in so-called units (testlets). All items of one unit are connected to the same stimulus. Adaptively selecting single items may result in a multiple presentation of the same stimulus to one student. This can be problematic regarding acceptance by the student and may invalidate item parameters.
- *Content balancing* is incorporated in PISA regarding the number of items used to measure literacy in reading, mathematics, and science. More items are presented for the major domain of one assessment than for the minor domains to report on the subdimensions of the major domain. An unrestricted adaptive algorithm is likely to miss the desired proportions of items per dimension and, therefore, may make it hard to reach an acceptable level of precision for subdimensional reporting.

Research Questions

All restrictions mentioned above can be accounted for by modified MAT algorithms. Nevertheless, the incorporation of the restrictions and the use of the PISA item pool will decrease the measurement efficiency of MAT to a certain degree. Whether or not MAT still leads to a relevant increase in measurement efficiency is not yet known. Thus, the present study examines the following three research questions:

1. Which gains in measurement efficiency can optimally be achieved if MAT is used instead of FIT for the assessment of reading, mathematics, and science in PISA?

2. How much can the testing sessions optimally be shortened if MAT is used instead of FIT for the assessment of reading, mathematics, and science in PISA without losing measurement precision?
3. Which gains in measurement efficiency can be expected by using MAT instead of FIT for the assessment of reading, mathematics, and science if the typical restrictions of the PISA assessments are taken into account?

The research questions are answered by a real data simulation. To allow direct interpretations regarding an application of MAT for the assessment of students' literacy in PISA, the simulation design is specified to match the conditions of the PISA 2006 assessment as closely as possible.

Method

Sample

The study is based on the responses of 14,624 fifteen-year-old students who participated in the PISA assessments during the years 2000 ($n = 5{,}073$), 2003 ($n = 4{,}660$), and 2006 ($n = 4{,}891$) in Germany. The answers of these students were used for the international PISA reports (OECD, 2001, 2004, 2007). Further descriptions of the samples can be found in the respective technical reports (Adams & Wu, 2002; OECD, 2005, 2009). The responses were used to estimate the item parameters and the multidimensional ability distribution. Both were needed to simulate an application of MAT in PISA. The details are given in the procedure section.

Design

Seven testing algorithms were compared with regards to three dependent variables. The reference condition FIT was contrasted with MAT without restrictions, MAT taking link items into account, MAT taking items in open response format into account, MAT including content balance, MAT taking the grouping of items to units into account, and MAT with all the restrictions mentioned before.

The first dependent variable is the *mean squared error* (*MSE*) of the ability estimates. It was calculated by the mean squared difference of the true person parameters θ and the person parameters $\hat{\theta}$ estimated after the simulation of the testing procedure for every dimension l (i.e., reading, mathematics, and science):

$$MSE_l = \frac{1}{N}\sum\nolimits_{j=1}^{N}\left(\hat{\theta}_{jl} - \theta_{jl}\right)^2. \tag{4}$$

The overall mean squared error, MSE_{all}, was calculated by averaging the dimension-specific *MSE* values:

$$MSE_{\text{all}} = \frac{1}{m}\sum\nolimits_{l=1}^{m} MSE_l. \tag{5}$$

Second, the *proportion of items*, *PCT*, for each dimension l was calculated by the average number of items presented for this dimension, T_l, and the average number of items presented for all three dimensions multiplied by 100:

$$PCT_l = \frac{100}{N} \sum\nolimits_{j=1}^{N} \frac{T_{jl}}{\sum\nolimits_{l=1}^{m} T_{jl}}. \tag{6}$$

Third, *measurement efficiency*, *ME*, was calculated by the ratio of measurement precision, defined as the inverse of the mean squared error to the mean number of presented items (see Frey, 2007; Segall, 2005) for every dimension l,

$$ME_l = \frac{\text{Measurement precision}}{\text{Test length}} = \frac{\frac{1}{MSE_l}}{T_l}, \tag{7}$$

as well as jointly for all dimensions,

$$ME_{\text{all}} = \frac{mN}{\sum\nolimits_{l=1}^{m} \sum\nolimits_{j=1}^{N} \left(\hat{\theta}_{lj} - \theta_{lj}\right)^2 \sum\nolimits_{l=1}^{m} T_l}. \tag{8}$$

Procedure

The simulation was accomplished in three steps: the generation of item and person parameters, the generation of responses, and the actual simulation of the testing procedure. Details of the three steps are provided in the following sections.

Generation of item and person parameters. The complete item pool of the present study consisted of all 348 items used in the assessments of PISA 2000, PISA 2003, and PISA 2006. The items are divided into 129 reading items, 95 mathematics items, and 124 science items. To obtain a common set of item parameters, the responses of the complete sample of 14,624 students were scaled with the Rasch-model for the dichotomously scored items and the partial credit model (Masters & Wright, 1997) for items with multiple score categories. In accordance with the international procedures of PISA, a separate unidimensional model was fitted for each content domain using ACER ConQuest (Wu, Adams, Wilson, & Haldane, 2007). In the following, this initial scaling is referred to as *Scaling 1*. The resulting set of item parameters was used in all conditions.

In *Scaling 2*, the responses of the subsample of $n = 4{,}891$ students who enrolled in PISA 2006 in Germany were scaled with the three-dimensional Rasch model for the dichotomously scored items and the partial credit model for the polytomously scored items. The item parameters were anchored at the values retrieved from Scaling 1. Thereby, the results of the present simulation study can be interpreted regarding the PISA 2006 assessment in Germany. The resulting means and the variance–covariance matrix of the multidimensional latent distribution were used for the generation of the responses and the simulation of the testing procedure.

Table 1. Booklet Design of PISA 2006

	Booklet												
Position	1	2	3	4	5	6	7	8	9	10	11	12	13
1	S1	S2	S3	S4	S5	S6	S7	M1	M2	M3	M4	R1	R2
2	S2	S3	S4	M3	S6	R2	R1	M2	S1	M4	S5	M1	S7
3	S4	M3	M4	S5	S7	R1	M2	S2	S3	S6	R2	S1	M1
4	S7	R1	M1	M2	S3	S4	M4	S6	R2	S1	S2	S5	M3

Note. PISA = Programme for International Student Assessment; R1 to R2 = reading clusters; M1 to M4 = mathematics clusters; S1 to S7 = science clusters.

Generation of Responses. In MAT, every item of the complete item pool theoretically could be presented to every student. Thus, for simulating MAT, a response of every student to every item is needed. As each student who participated in one of the PISA assessments only answered to a subset of the entire item pool, responses were generated. Therefore, individual values in reading, mathematics, and science were randomly drawn for 4,891 simulees from the multidimensional latent distribution derived from Scaling 2 under the assumption of multivariate normality. These person parameters were considered as true ability parameters $\boldsymbol{\theta}_j$. In conjunction with the item parameters from Scaling 1, $\boldsymbol{\theta}_j$ was used to generate a response for each simulee to each item of the item pool based on the three-dimensional Rasch model. The mean item parameter was used for the items with multiple score categories. The resulting 4,891 × 348 complete response matrix served as basis for the simulation of the testing procedure.

Simulation of the testing procedure. The actual testing procedure was simulated using the statistical package SAS 9.2. The following seven testing conditions were specified.

Condition 1: Fixed item testing. Within the reference condition, *FIT,* the characteristics of the PISA 2006 assessment were rebuilt. The PISA 2006 booklet design (see Frey, Hartig, & Rupp, 2009; OECD, 2009), comprising 13 booklets, each with a testing time of 120 minutes, was used to assign the items to the simulees. The item pool of the PISA 2006 assessment consisted of 179 items measuring reading, mathematics, and science. Each item was assigned to so-called clusters, which were systematically assigned to booklets and positions in booklets (Table 1).

In the present study, the booklet design was used to select a set of responses from the complete response matrix for each simulee. The responses to all other items were treated as not administered for this simulee. To arrive at the final statistics, the selected responses were scaled with a three-dimensional Rasch model with the item parameters anchored at the values from Scaling 1.

Condition 2: Unconstrained MAT. The booklet design was only used in the condition FIT. In all other conditions, the items were selected adaptively using MAT with none,

one, or several restrictions. The first item to be presented was randomly chosen from the complete item pool. Adaptive item selection started with the second item. Item selection and ability estimation were based on the Bayesian approach of Segall (1996), making use of the item parameters (from Scaling 1) and the variance–covariance matrix of the prior distribution $\mathbf{\Phi}$ (from Scaling 2). In the condition unconstrained MAT, Equation (3) served as item selection criterion. The responses to the selected items were taken from the complete response matrix, not selected items were treated as not administered. The test was terminated when the next item would have exceeded the maximum testing time of 120 minutes. The calculation of the testing time was based on the testing time scheduled for item delivery in PISA 2006, which was 2.14 minutes for reading items, 2.50 minutes for mathematics items, and 2.05 minutes for science items. For the calculation of the final results, the selected responses were scaled with a three-dimensional Rasch model with the same specifications as in the condition FIT.

Condition 3: MAT with link items. The same procedure as described for unrestricted MAT was applied for Condition 3 and extended by the additional restriction of presenting a sufficient number of link items. The aim was to present the minimum number of 9 link items measuring reading literacy, 15 link items measuring mathematical literacy, and 6 link items measuring scientific literacy to each simulee. To achieve this, item selection was restricted to the subset of link items after a testing time of 40 minutes had elapsed. At this stage, only the link items were considered as candidate items. After the desired number of link items was presented to the simulee, items were again selected from the complete item pool.

Condition 4: MAT with open items. In this condition, the procedure described for unrestricted MAT was supplemented by the restriction of accounting for items that cannot directly be scored by a computer. A qualitative analysis of the complete item pool revealed that 236 of the items (68%) can directly be scored by a computer. A human coder must score the remaining 112 items (32%). Only the 236 items of the first group were used to revise the provisional ability vector $\hat{\theta}$ in the condition MAT with open items. The other items were presented to the simulees as well. The responses to these items were not used to revise $\hat{\theta}$ but were considered in the final scaling.

Condition 5: MAT with units. In this condition, the item selection was restricted to the selection of complete units instead of single items. All other specifications were the same as in unrestricted MAT. The size of the units in the PISA item pool ranges from one to seven items. The mean item difficulty of units containing several items is predominantly around 0. Therefore, the summed or averaged item information would not have been a good criterion for selecting units. To enable a better adjustment of the selected units to the provisional ability vector, the unit including the item with the highest information was selected for presentation.

Condition 6: MAT with content balance. In condition 6, unrestricted MAT was constrained to achieve the same proportions of items per dimension as specified by the booklet design of PISA 2006, which is 15.64% items measuring reading literacy, 26.82% items measuring mathematical literacy, and 57.54% items measuring scientific literacy. If the proportion of presented items exceeded these proportions within the testing process for one dimension, no additional items were presented for the respective dimension. When later the proportion of presented items had fallen below the upper limit, items from this dimension were reconsidered as candidate items within item selection. Thus, content balance was used here on the level of dimensions and not on the level of content areas within a dimension, which is the conventional application of content balance in unidimensional CAT.

Condition 7: MAT with all restrictions. In this condition, all restrictions were used simultaneously. All other specifications were the same as in unrestricted MAT. Content balance was only active until a testing time of 40 minutes had passed and after the required number of link items had been presented. Hence, reaching the required number of link items was given higher priority in this condition than content balance. In all testing phases, complete units were selected instead of items and open items were not used to revise the provisional ability vector.

Results

First, the psychometric properties of the literacy scales in reading, mathematics, and science will be described. Then, results regarding the mean squared error, the proportions of presented items per dimension, and the measurement efficiency observed in the seven conditions will be presented. Based on this information, the research questions will be answered.

Scaling Outcomes

The characteristics of the item pool strongly influence the performance of adaptive tests (e.g., Veldkamp & van der Linden, 2010). The item parameters derived from Scaling 1, show a smaller variation for the science items ($SD = 0.98$) than for the reading items ($SD = 1.07$) and the mathematics items ($SD = 1.24$). The mean of the item parameters was constrained to 0 for each dimension. For each of the three scales, the frequency of items with a medium difficulty is high, whereas the frequency of items decreases toward the extremes of the ability scale (Figure 1).

The item parameter distributions are well suited for FIT but are far from optimal for MAT. Only 10 items of the complete item pool (2.87%) have an item parameter smaller than -2.00 logits. Only 13 items (3.74%) have an item parameter larger than 2.00 logits, seven of them are reading items. For MAT, the availability of more items of extreme difficulty would be favorable to facilitate the selection of items at an appropriate difficulty level for students with very low or very high ability. Nevertheless, even though the item pool is not optimal, it can be used for MAT and should lead to increases in terms of measurement efficiency.

The multidimensional person parameter distribution obtained from Scaling 2 has a mean of 0.48 for reading, 0.00 for mathematics, and 0.32 for science. The variance–covariance matrix $\mathbf{\Phi}$ is shown in Table 2.

The comparably small variance for science reflects that $\mathbf{\Phi}$ is based on the booklet design of PISA 2006 in which more than half of the presented items were science

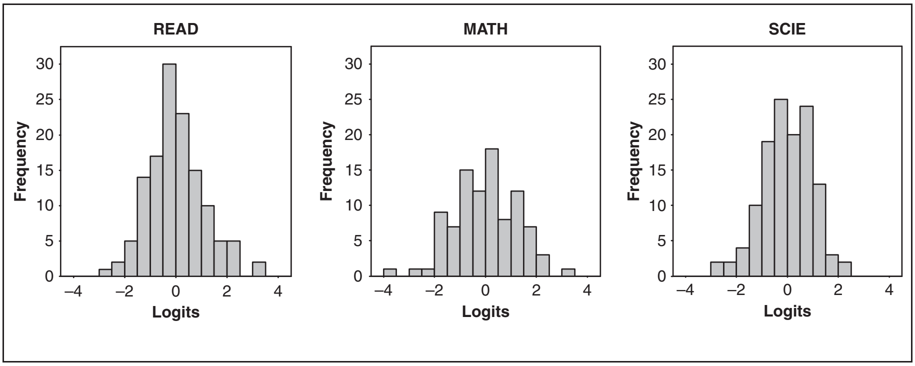

Figure 1. Item parameter distribution of 129 items measuring reading literacy (READ), 95 items measuring mathematical literacy (MATH) and 124 items measuring scientific literacy (SCIE).

items. The many science items permit a more precise measurement of sciences compared with reading and mathematics. This also results in the EAP/PV (expected a posteriori/plausible values) reliability (Adams, 2005) obtained from Scaling 2, which is higher for science (.92) than for reading (.81) and mathematics (.86).

Mean Squared Error, Proportions of Items per Dimension, and Measurement Efficiency

The mean squared error, the proportions of presented items per dimension, and the measurement efficiency observed in the seven conditions are shown in Table 3.

In all MAT conditions, the mean squared error for reading and mathematics, as well as the average mean squared error, is smaller than for FIT. The smallest mean squared error is observed for unrestricted MAT. This is depicted in Figure 2, which shows scatterplots of the true values and the estimated values in reading, mathematics, and science for FIT and unrestricted MAT. The spread of the marks for reading and mathematics is obviously smaller for unrestricted MAT than for FIT. For science, however, this is not the case. Here, only for MAT with content balance the mean squared error is smaller compared with FIT.

The comparably small mean squared error for science in FIT and MAT with content balance is achieved by administering a lot of items for this dimension. With 57.52% (FIT) and 56.36% (MAT with content balance), the majority of administered items are science items. While showing that the used content balance method worked well in achieving the desired proportions of items per dimension, the results also show that the administration of many science items leads to comparably lower values in measurement efficiency. Subsequently, in the condition MAT with content balance, the measurement efficiency for science is the lowest of all MAT conditions. When looking at the overall measurement efficiency calculated by Equation (8), however, all MAT conditions outperform FIT. Based on these general observations, the three research questions will subsequently be answered in the following sections.

Table 2. Variance–Covariance Matrix of Reading Literacy, Mathematical Literacy, and Scientific Literacy

	Dimension		
Dimension	READ	MATH	SCIE
READ	1.89	1.42	1.35
MATH	1.42	1.57	1.23
SCIE	1.35	1.23	1.22

Note. READ = reading literacy; MATH = mathematical literacy; SCIE = scientific literacy.

Table 3. Mean Squared Error, Proportion of Presented Items per Dimension, and Measurement Efficiency of Seven Testing Algorithms

	MSE				*PCT*			*ME*			
Condition	READ	MATH	SCIE	*M*	READ	MATH	SCIE	READ	MATH	SCIE	All
FIT	0.38	0.25	0.13	0.25	15.65	26.83	57.52	0.31	0.27	0.25	0.22
Unrestricted MAT	0.16	0.15	0.15	0.15	40.78	38.32	20.90	0.29	0.33	0.62	0.38
MAT + link items	0.16	0.15	0.15	0.15	41.58	39.03	19.39	0.29	0.33	0.64	0.37
MAT + open items	0.14	0.17	0.16	0.15	50.92	33.11	15.97	0.27	0.34	0.73	0.36
MAT + units	0.20	0.17	0.15	0.17	37.40	34.65	27.95	0.25	0.31	0.45	0.32
MAT + content balance	0.25	0.17	0.09	0.17	16.36	27.27	56.36	0.44	0.39	0.36	0.32
MAT + all	0.25	0.17	0.13	0.18	28.01	33.85	38.14	0.27	0.32	0.37	0.30

Note. *MSE* = mean squared error; *PCT* = percentage of presented items per dimension; *ME* = measurement efficiency; READ = reading literacy; MATH = mathematical literacy; SCIE = scientific literacy; FIT = fixed item testing; MAT = multidimensional adaptive testing.

Maximum Gain in Measurement Efficiency

The first research question asks which gains in measurement efficiency can optimally be achieved if MAT is used instead of FIT for the assessment of reading, mathematics, and science in PISA. The optimal measurement efficiency for MAT is achieved by an unrestricted item selection. The measurement efficiency observed for unrestricted MAT is 0.38 compared with 0.22 for FIT. Even though the measurement efficiency of unrestricted MAT is 74% higher than for FIT, the difference is relatively small compared with the differences found in simulation studies. Nevertheless, the observed gain in measurement efficiency may well be of practical importance as it can be used to reduce the number of presented items without losing measurement precision.

Maximum Reduction of Administered Items

The second research question asks for the number of items to which the test can be optimally shortened if MAT is used for the assessment of reading, mathematics,

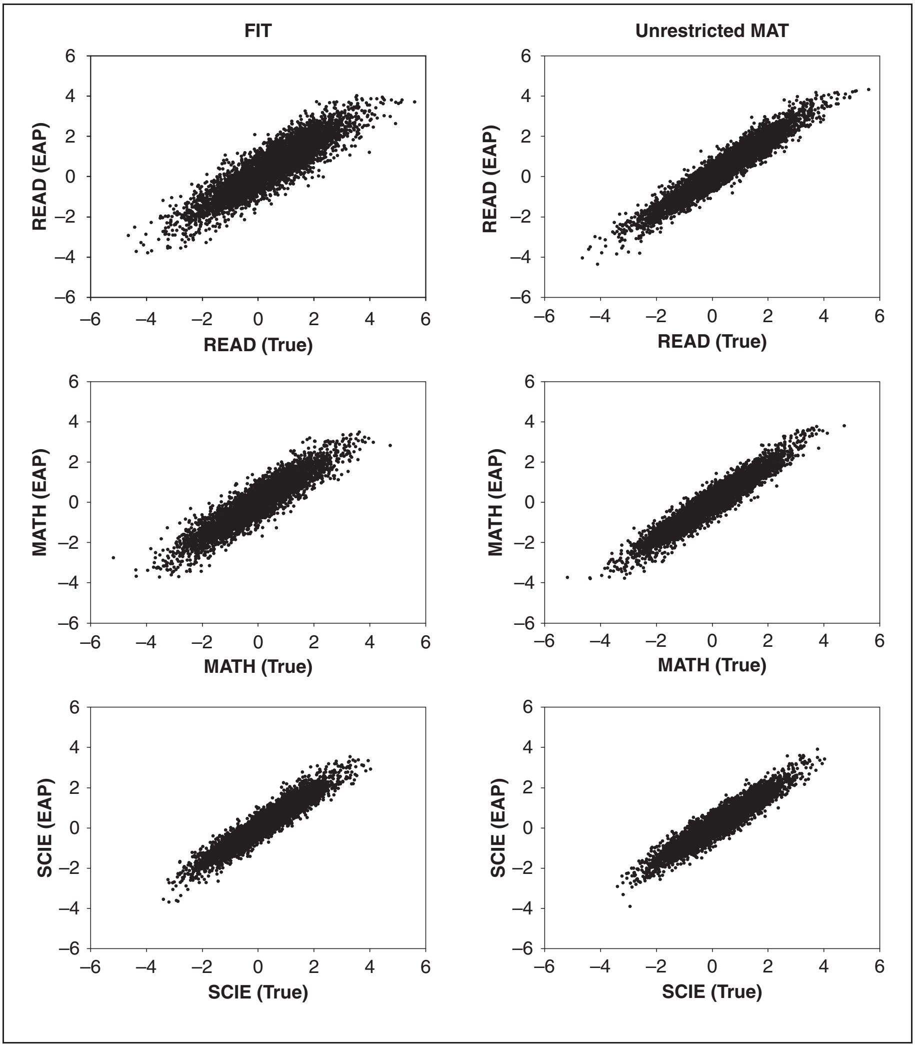

Figure 2. Scatter plots of true values and estimated values in reading, mathematics and science for fixed item testing (FIT) and unrestricted multidimensional adaptive testing (Unrestricted MAT).

and science in PISA instead of FIT. Figure 3 shows the average mean squared error of unrestricted MAT as a function of the number of administered items. The average mean squared error of FIT with a mean number of 55 items is indicated by the dotted line. The point of intersection indicates the test length at which the average mean squared error of unrestricted MAT equals the average mean squared error of FIT. Beginning with the 26th item, the average mean squared error of unrestricted MAT becomes smaller than for FIT. Thus, the test can be optimally reduced to a length of 26 items if MAT is used instead of FIT, without losing measurement precision.

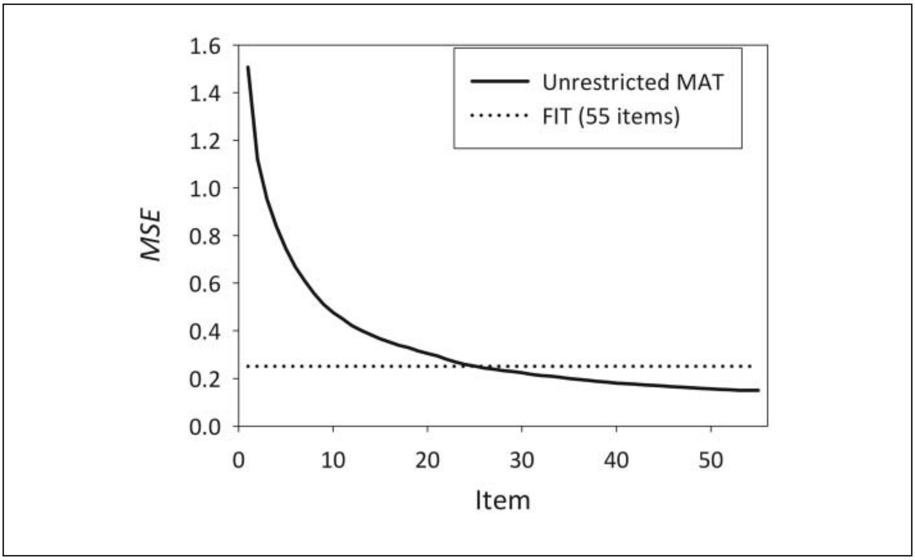

Figure 3. Mean squared error (*MSE*) of unrestricted multidimensional adaptive testing (MAT) as a function of test length compared to the mean squared error of fixed item testing (FIT, dotted line) using PISA-items.

Nevertheless, because different mean squared errors were observed for FIT for the three dimensions, the number of items to which a test can be shortened with unrestricted MAT also varies between dimensions. Although unrestricted MAT resulted in a smaller mean squared error for reading, after presenting an average of 15 items, and for mathematics, after an average of 22 items, the mean squared error in science observed for FIT was not reached until the end of the test. Only in the condition of MAT with content balance, a smaller mean squared error in science was achieved than for FIT (after Item 36).

Measurement Efficiency Taking Restrictions Into Account

The third research question asks which gains in measurement efficiency can be expected if MAT is used instead of FIT for the assessment of reading, mathematics, and science if the restrictions that are associated with PISA are taken into account. A comparison of the measurement efficiency between the seven conditions is shown in Figure 4.

As expected, all restrictions lead to a decrease in measurement efficiency compared with unrestricted MAT. Whereas taking link items and open items into account only induces small reductions in measurement efficiency, larger reductions result for the other restrictions. If all restrictions are included in MAT, the measurement efficiency is 40% higher than for FIT. Nonetheless, the combination of the four

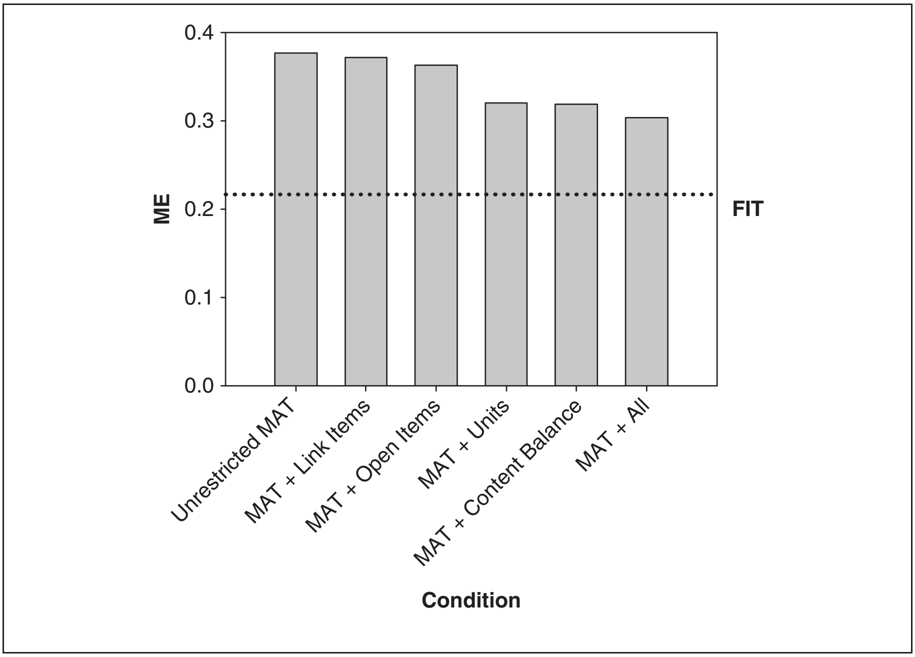

Figure 4. Measurement efficiency of multidimensional adaptive testing (MAT) with different restrictions compared to fixed item testing (FIT, dotted line) using PISA-items.

restrictions led to the case in which not all of them were completely successful. Although with 9.09 reading items, 15.08 mathematics items, and 6.03 science items the average numbers of link items are reached as requested, the desired proportions of items per dimension are not reached satisfactorily in every case.

Discussion

In an *optimal world* where an unrestricted MAT algorithm is applicable, the results are promising for MAT: Compared with FIT, measurement efficiency can be increased by about 74% and the mean number of items that need to be presented to each student can be reduced from 55 items to 26 items without a loss in average measurement precision. Testing time could, therefore, theoretically be shortened from 120 to 57 minutes if the item pool of PISA is used. Still, the observed gain in measurement efficiency is rather small compared with the gain reported in simulation studies. For example, in a similar case with five highly correlated dimensions, Frey and Seitz (2010) found the measurement efficiency of MAT to be 273% higher than for FIT. The comparatively small gain in measurement efficiency of unrestricted MAT in the present study is mainly due to using the PISA item pool which is not very well suited for MAT and to not using knowledge about the multivariate prior distribution for ability estimation. If an item pool better suited for adaptive testing than the PISA item pool is used and knowledge about the prior distribution is used for ability estimation, further increases up to the level reported in simulation studies (overview in Frey & Seitz, 2009) can be expected. Then, measurement efficiency can be increased by MAT by more than 250%, which translates to a reduction of the number of items that need to be presented to less than one third while keeping measurement precision constant. Nevertheless, it has to be noted that only items with an individual solution probability of about 50% are presented in MAT. These items may require a longer time to answer than typical FIT items that cover a broader range of individual solution probabilities. Items with both a very high and a very low solution probability are presumably answered rather quickly. However, whether these assumptions regarding the answering time are true for PISA is an empirical question not yet answered.

In a *real world*, when PISA-specific restrictions are taken into account, the measurement efficiency of MAT is lower than in the case of unrestricted MAT. It is only 40% higher compared with FIT. Thus, it is questionable whether it suffices for a change of the testing algorithm from FIT to MAT. Furthermore, the desired average measurement precision for the major domain science is not reached within the given testing time when all restrictions are used in MAT. Thus, prioritizing content balance seems to be necessary if subdimensional reporting is the goal. Additionally, the present simulation study only highlights the most important formal restrictions of PISA. Other possible restrictions as well as the psychological effects of the testing algorithm on students' response behavior are not modeled. These may also affect the multidimensional ability distribution. Since the differences interpreted in PISA are often rather small, systematic effects due to a change of the testing algorithm may easily lead to invalid inferences. Therefore, we suggest not applying MAT for the assessment of the already established literacy scales in reading, mathematics, and science in PISA.

Nevertheless, we propose to consider MAT as an alternative to FIT whenever new constructs are introduced to PISA. With a suitable item pool and a well-planned MAT algorithm, very high measurement efficiency can be expected. Other advantages of MAT are, for example, the possibility to accomplish linking and content balance on an individual level and not on the level of larger groups or the whole sample and the possibility to stabilize the mean squared error of the ability estimates over the ability scale, whereas for FIT the mean squared error typically increases toward the extremes of the ability scale. When a test is constructed anew, most of the PISA-specific restrictions need not be taken into account. For example, single items or units optimized for adaptive testing could be used in the item selection process. One option to optimize units for adaptive testing is to compose them in a way that the item difficulties vary only slightly within units, and the mean item difficulties of complete units vary strongly between units. Thus, the units could be better tailored to the response patterns of the students. This should lead to more satisfying results than using the original PISA units in MAT, whose item difficulties vary strongly within units but only slightly between units. When implementing MAT, another objective should be to minimize the number of restrictions that need to be taken into account when selecting items. The results found in the present study for MAT with all restrictions indicate that combinations of multiple restrictions can quickly lead to inferences between the restrictions. The number of restrictions that can be dealt with in MAT may be increased by using methods proposed for severely constrained computerized adaptive tests, such as the shadow testing approach of Veldkamp and van der Linden (2002) or the maximum priority index method of Cheng and Chang (2009). Of course, an operational MAT will almost always include restrictions, but one should keep in mind that the less restrictive the item selection process is, the higher the measurement efficiency will be. Finally, very high measurement efficiency can be fostered by the application of an item pool whose item parameter distribution is well suited for adaptive testing. The item pool should entail enough items over the whole ability range.

When considering the usage of MAT, we propose to precisely cost out all possible alternatives. Obviously, at this point one must take into account the fact that computers must be available at the testing location when MAT is used. It should be shown beforehand that the high measurement efficiency of MAT and possible other advantages of a computer-based test delivery will outweigh the costs induced by MAT. Other general advantages and disadvantages of computer-based testing compared with paper-and-pencil testing have already been discussed in detail and are thus not repeated here (see Bartram & Hambleton, 2005; Parshall, Spray, Kalohn, & Davey, 2002).

In conclusion, the present real data simulation illustrates that MAT can be advantageous for the measurement of student achievement even under constrained conditions. We suggest that assessment developers and measurement specialists consider this highly efficient way of testing whenever new assessments are set up or new constructs are introduced to ongoing assessments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The preparation of this article was supported by grant FR 2552/2-2 from the German Research Foundation (DFG) in the Priority Programme ''Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes'' (SPP 1293).'' and from the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany.

Note

1. More precisely, in the used model the constant $D = 1.7$ was multiplied with $\mathbf{a}_i'$ in the exponent to produce equivalence to the normal ogive model. This does not affect the reported results.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

Adams, R. J., & Wu, M. (Eds.). (2002). *PISA 2000 technical report.* Paris, France: OECD.

Bartram, D., & Hambleton, R. K. (2005). *Computer-based testing and the Internet: Issues and advances*. New York, NY: Wiley.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369-383.

Frey, A. (2007). Adaptives Testen [Adaptive testing]. In H. Moosbrugger & A. Kelava (ed.), *Testtheorie und Fragebogenkonstruktion* [Test theory and construction of questionnaires].(pp. 261-278). Berlin, Germany: Springer.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice,28*, 39-53.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*, 89-94.

Frey, A., & Seitz, N. N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz [Multidimensional adaptive testing of competencies: Results regarding measurement efficiency]. *Zeitschrift für Pädagogik, 56,* 40-51.

Hartig, J., & Hoehler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie, 216*, 89-101.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-122). New York, NY: Springer.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris, France: Author.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris, France: Author.

OECD. (2005). *PISA 2003 technical report*. Paris, France: Author.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris, France: Author.

OECD. (2007). *PISA 2006. Science competencies for tomorroẁs world: Vol. 1. Analysis*. Paris, France: Author.

OECD. (2009). *PISA 2006 technical report.* Paris, France: Author.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. Dordrecht, Netherlands: Springer.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

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

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

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York, NY: Academic Press.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398-412.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575-588.

Veldkamp, B. P., & van der Linden, W. J. (2010). Designing item pools for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 231-245). New York, NY: Springer.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 450-480.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne, Victoria, Australia: ACER Press.