

# Differentiated assessment of mathematical competence with multidimensional adaptive testing

*Anna Mikolajetz<sup>1</sup> & Andreas Frey<sup>2,3</sup>*

## **Abstract**

The theoretical frameworks of large-scale assessments (LSAs) typically describe complex competence constructs. However, due to restrictions in testing time, the complexity of these competence constructs is often reduced to one or a small number of dimensions in operational LSAs. Because of its very high measurement efficiency, multidimensional adaptive testing (MAT) offers a solution to overcome this shortcoming. The present study demonstrates the capability of MAT to measure the 11 subdimensions of mathematical competence that are described in the theoretical framework of the German Educational Standards in Mathematics with sufficient precision without increasing test length. The characteristics of an empirically derived 11-dimensional competence distribution of 9,577 students and the parameters for 253 operational test items were used to simulate the application of MAT. Typical restrictions such as the usage of testlets or the fact that items in an open response format are in the item pool were taken into account in the simulation. Although the used item pool was not constructed for adaptive testing, the results show substantially higher reliability estimates for MAT compared to non-adaptive testing, especially for the subdimensions of mathematical competence, which are not as yet reported in the assessment. The results underscore the capacity of MAT to precisely measure competence constructs with many dimensions without the need to increase test length. This research therefore closes the current gap between theoretical underpinnings and actual measures in LSAs.

Keywords: computerized adaptive testing, item response theory, multidimensional IRT models, large-scale assessment

---

<sup>1</sup> *Correspondence concerning this article should be addressed to:* Anna Mikolajetz, Dipl.-Psych., Friedrich Schiller University Jena, Department of Methodology and Evaluation Research, Am Steiger 3, 07737 Jena, Germany; email: [anna.mikolajetz@uni-jena.de](mailto:anna.mikolajetz@uni-jena.de)

<sup>2</sup> Friedrich Schiller University Jena, Germany

<sup>3</sup> Centre for Educational Measurement, University of Oslo, Norway

The improvement of educational processes is a primary aim of societies, governments, and research. In a variety of large-scale assessments of student achievement (LSAs), educational attainments are assessed and associated with different individual or contextual characteristics such as socio-economic background or the structural features of school systems. The results of LSAs provide valuable information for governments, and such information can be used to evaluate the extent to which educational goals are being achieved to monitor student achievement over time and to facilitate the making of decisions on reform measures. Well-known international LSAs include the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). These assessments are based on elaborate theoretical frameworks. In these frameworks, the structure of the competence constructs of interest is typically specified in terms of complex theoretical models. The development of these models was motivated by an increasing interest in measuring competencies with a strong reference to real-life tasks in specific contexts instead of measuring general cognitive abilities (e.g., McClelland, 1973; Shavelson, 2013). Typical examples of such complex competence constructs are the constructs of student literacy in mathematics, reading, and science, which are the focuses of the PISA (e.g., the Organisation for Economic Co-operation and Development [OECD], 2013a). For instance, the definition of mathematical literacy in the PISA is based on a theoretical model that differentiates between multiple cognitive processes, mathematical content areas, and situations in which mathematical literacy can be applied.

To adequately operationalize such complex definitions and thus the corresponding theoretical underpinnings, large item pools are required. The examinees' response behavior to the items that are included in such a pool is usually modeled by item response theory (IRT) models (e.g., de Ayala, 2009; Embretson & Reise, 2000). Generally, multidimensional IRT (MIRT) models can be used to model the complex structures of competence models and, hence, they make it possible to link task performance with the multiple aspects of the competence construct that are considered to affect performance (e.g., Ackerman, Gierl, & Walker, 2003; Adams, Wilson, & Wang, 1997; Hartig & Höhler, 2008; Walker & Beretvas, 2003). However, the potential of MIRT models is restricted by the resources that are available for testing. Higher demands in terms of precise measures for multiple dimensions imply higher testing effort, e.g., considerably longer testing times and/or a higher number of students being assessed. To limit testing effort and the associated costs, as well as to ensure the cooperation with the institutions that are involved, highly efficient testing procedures are needed. Multidimensional adaptive testing (MAT, e.g., Frey & Seitz, 2009; Segall, 2010) is a promising procedure for the measurement of complex competence constructs on a very high level of efficiency, while consuming considerably fewer testing resources than conventional testing with fixed test forms. MAT is a method that is used to simultaneously measure the standing of an examinee on several dimensions. In MAT, item selection is based on the previous responses of the examinee. Tailoring the item selection to the individual competence level makes it possible to substantially increase measurement efficiency. This can be used to either reduce test length or increase measurement precision (e.g., Frey & Seitz, 2011). Thus, MAT provides the necessary prerequisites to measure higher-dimensional competence constructs with a reasonable amount of testing effort and opens up the possibility of

reporting more differentiated results compared to conventional methods. This approach was supported by van der Linden in a review of the developments in the area of adaptive testing in 2008, and it was illustrated by Frey, Seitz, and Kröhne (2013) in the case of the PISA and by Yao in the case of the Armed Services Vocational Aptitude Battery (ASVAB), who compared different item selection algorithms (Yao, 2012, 2013) and two item exposure methods (Yao, 2014). However, until now, MAT has not been utilized in operational LSA programs, and the potential of MAT for the assessment of complex competence constructs has only been examined rudimentarily.

To demonstrate the benefits of MAT for the assessment of complex competence constructs, the present study investigates the use of the method in a typical LSA context. As a placeholder for other LSAs, the assessment that is used to measure the attainment of the German Educational Standards in Mathematics for Secondary Education is focused. This LSA is based on a theoretical framework that is similar to the mathematical literacy framework of the PISA, which differentiates between a content-related and a process-related view of mathematical competence. The differentiated conceptualization of mathematical competence that is described in the theoretical framework of the German Educational Standards in Mathematics is, however, reduced to one overall mathematical competence dimension and five subdimensions for content-related competencies in terms of the reporting of the test results (which is similar to the PISA's overarching ideas). No results are reported for the six process-related competencies that are also described in the theoretical framework. The present study investigates whether MAT can be used to successfully measure all of the content-related and process-related subdimensions that are described in the theoretical framework of the German Educational Standards in Mathematics for Secondary Education with acceptable reliability. The study should be of particular interest for researchers who are involved in the development and implementation of LSAs, considering the use of MIRT models within a computer-based assessment. Note that the focus of the study is not the examination of which model structure best fits with a particular LSA data set, but rather to the examination of the benefits of MAT in the measurement of already established complex MIRT model applications. Accordingly, we use an already established MIRT model and examine which improvements can be obtained by using MAT instead of conventional testing.

The text is organized as follows: First, the theoretical framework for mathematical competence according to the German Educational Standards for Secondary Education, its assessment, and psychometric modeling are described. Second, the concept of MAT is outlined, including the empirical research on the application of MAT to measure multi-dimensional constructs. These two theoretical sections are then combined, and research questions for the present study are formulated. Subsequently, the methods and the results of the simulation study are presented, followed by a section that offers implications for the application of MAT in the differentiated measurement of the competence construct that underlies the German Educational Standards in Mathematics for Secondary Education.

## Mathematical competence according to the German educational standards for secondary education

### Theoretical framework

In 2003, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Kultusministerkonferenz) decided on the implementation of national educational standards in mathematics, German, and the first foreign language for the secondary school qualification, which is usually attained at the end of grade ten. In the subsequent years, the standards for additional subjects and other graduation levels were introduced. These educational standards define competencies in different subjects that should have been acquired by students by a certain point in their educational career (Kultusministerkonferenz, 2003). These standards serve as a reference for competence-oriented teaching, and they provide a basis for evaluations at various levels: the educational system as a whole, schools, and classes (Blum, Drücke-Noe, Leiß, Wiegand, & Jordan, 2005). The underlying theoretical frameworks were developed based on a combination of already established models and practical school experience.

The theoretical framework of the German Educational Standards in Mathematics for Secondary Education is based on two major sources: the Standards of the National Council of Teachers of Mathematics (NCTM, 2000) and the theoretical framework of the PISA 2003 for mathematics (OECD, 2003; see Hartig & Frey, 2012 for a conceptual and psychometric comparison of the framework of the German Educational Standards in Mathematics and the mathematical literacy framework of the PISA). Both of the source frameworks, the NCTM and the PISA 2003, distinguish between a content-related and a process-related view of mathematical competence. The former represents different content areas in mathematics; the latter refers to cognitive processes and activities while addressing mathematical content. This distinction was adopted in the theoretical framework of the German Educational Standards in Mathematics for Secondary Education (Kultusministerkonferenz, 2004). The content-related view is reflected by the five mathematical content areas: 1) *numbers*, 2) *measurement*, 3) *space and shape*, 4) *functional relationships*, and 5) *data and chance*. The cognitive processes that are involved in mathematical activities are described by six general mathematical competencies: 1) *reasoning and argumentation*, 2) *problem solving*, 3) *modeling*, 4) *using representations*, 5) *applying techniques and formal procedures*, and 6) *communicating*. It is assumed that solving a mathematical task requires sufficient ability level in two areas, mathematical content and one or several general mathematical competencies.

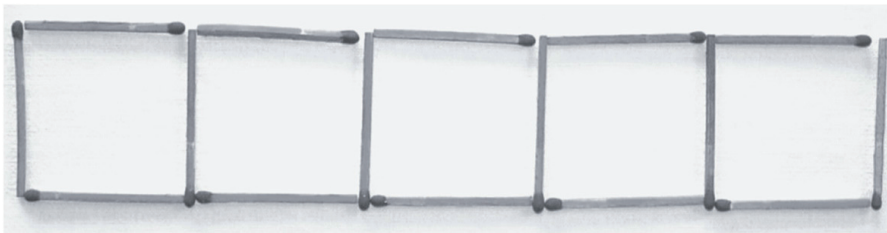
### Assessment and psychometric modeling

The degree to which the German Educational Standards in Mathematics for Secondary Education are attained is empirically assessed with typical LSA methods. To operationalize the theoretical framework that is described above, an extensive item pool was devel-

oped that was used to assess the degree to which standards are attained. To illustrate the operationalization, Figure 1 contains two items from the development of the test to measure the attainment of the German Educational Standards in Mathematics for Secondary Education (Prenzel & Blum, 2007). They are both assigned to mathematical content area 4) functional relationships and general mathematical competency 4) using representations. Additionally, a second item is assigned to measure general mathematical competencies 2) problem solving and 5) using mathematical symbols and techniques.




**CHAIN OF MATCHES**

A chain of squares can be formed by using matches.



**Item 1**

Write the required number of matches in the empty boxes.

	Number of squares	Number of matches
	1	<input type="text" value="4"/>
	2	<input type="text" value="7"/>
	3	<input type="text"/>
	4	<input type="text"/>

**Item 2**

Specify an equation that describes the relationship between the number  $s$  of squares and the number  $m$  of matches in a general manner.

$m =$  \_\_\_\_\_

**Figure 1:**  
Two example items (translated and adopted from Katzenbach et al., 2009)

The first assessment of the German Educational Standards for reporting purposes was carried out in 2012 (Pant et al., 2013). The results regarding mathematical competence were reported by means of an overall score for mathematical competence and by scale scores for the five mathematical content areas. The scaling of the assessed data was conducted in two steps (Hecht, Roppelt, & Siegle, 2013). In the first step, item parameters were estimated by using five separate unidimensional Rasch models – one for each of the five content-related competencies. Equation 1 shows the Rasch model (Rasch, 1960) where the probability of a person  $j$  to answer an item  $i$  correctly ( $P(U_{ij} = 1)$ ) is a function of the ability  $\theta_j$  of the person and the item difficulty  $b_i$ .

$$P(U_{ij} = 1 | \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (1)$$

In the second step, person parameters were estimated with conditioning and item parameters that were fixed to the values that were obtained in step one. Two models were used in this step. A unidimensional Rasch model (Equation 1) with all of the items of the pool loading on the same dimension was fitted to obtain an overall score for mathematical competence. Additionally, a five-dimensional Rasch model was used to obtain scale scores for the five mathematical content areas. For both models, the item parameters were fixed to the values from step one. Equation 2 shows the used multidimensional extension of the Rasch model from Equation 1, where the probability of person  $j$  to answer item  $i$  correctly is defined as a function of a set of item parameters  $\mathbf{a}_i$ , and  $b_i$  and an ability vector  $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jk})$  for  $k$  dimensions.  $b_i$  specifies the item difficulty. It is multiplied by the  $k \times 1$  vector  $\mathbf{1}$  which is filled with ones to be used for all measured dimensions.  $\mathbf{a}_i$  is a  $k \times 1$  vector filled with zeros and equal, nonzero values which indicate whether a dimension is measured by the item ( $a_{ik} \neq 0$ ) or not ( $a_{ik} = 0$ ).

$$P(U_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, b_i) = \frac{\exp(\mathbf{a}_i'(\boldsymbol{\theta}_j - b_i \mathbf{1}))}{1 + \exp(\mathbf{a}_i'(\boldsymbol{\theta}_j - b_i \mathbf{1}))} \quad (2)$$

In the assessment of the Educational Standards in Mathematics for Secondary Education,  $\mathbf{a}_i$  contains only one element with  $a_{ik} \neq 0$  for each item in the item pool. Hence, the performance on each single item is predicted by exactly one dimension that is the proficiency related to one of the five mathematical content areas. Models with this type of multidimensionality are also referred to as models with between-item multidimensionality or as between-item models (Adams, Wilson, & Wang, 1997).

Note that no results for the additional components of the theoretical framework, i.e., the general mathematical competencies, have yet been included in the standard reporting process because it was not possible to constitute respective scales with sufficient precision (Carstensen & Frey, 2007; Roppelt, Blum, & Pöhlmann, 2013). Mikolajetz and Frey (2014) attempted to overcome this issue. They hypothesized that the problems that are encountered in the fitting of psychometric models that distinguish between the general mathematical competencies may be due to the unexamined theoretical assignments of the

items to the six general mathematical competencies that were carried out by subject matter experts. To solve this problem, they used empirical data to refine the subject matter experts' assignments of the items to the general mathematical competencies by using statistical criteria. They specified an 11-dimensional model according to Equation 2, where the response to an item was a function of both the proficiency that is related to a mathematical content area and at least one general mathematical competency. Because in this case more than one element of  $\mathbf{a}_i$  is nonzero, so that responding correctly to an item is a function of more than one latent variable, this model represents a model with within-item multidimensionality (Adams et al., 1997). For the mathematical content areas, the elements of  $\mathbf{a}_i$  were constrained to an equal, nonzero value for all items that loaded on the respective dimension – which corresponds to the modelling approach in the original assessment. For the general mathematical competencies, the elements of  $\mathbf{a}_i$  did not have an equality constraint. The refined assignment of items to the six general mathematical competencies resulted from the elimination of the loadings of items on competencies with low or negative discrimination parameter estimates. More information regarding the item pool and the multidimensional structure can be found in the method section. The authors showed that using the refined scales makes it possible to report results for both mathematical content areas and general mathematical competencies on a group level. They found that gender-related results regarding the general mathematical competencies with the established scales were comparable to the few findings on similar constructs. While boys outperformed girls on problem solving, which was also found, for example, by Hyde, Fennema and Lamon (1990) and Lindberg, Hyde, Petersen and Linn (2010), girls showed better results in applying techniques and formal procedures, which corresponds to findings on equal or better results by girls in the computation and application of standard procedures (e.g., Hyde et al., 1990; van den Heuvel-Panhuizen, 2004). The EAP/PV reliability (Adams, 2005) for content-related competencies (without conditioning) varied between .63 and .79 and between .39 and .72 for process-related competencies. The authors noted that improving reliability for the subdimensions would be necessary to fully meet the standards of LSA.

## Multidimensional adaptive testing

In MAT, the multidimensionality of constructs is incorporated directly into the measurement process by using MIRT models. MAT requires an item pool whose item parameters have already been estimated in a previous calibration study. The two major approaches that are used for ability estimation and item selection in MAT are the maximum likelihood approach (Segall, 1996; van der Linden, 1999) and the Bayesian approach by Segall (1996). In the simultaneous measurement of multiple correlated dimensions, the Bayesian approach is particularly beneficial. By using knowledge about the multivariate distribution of the dimensions, a substantially higher measurement efficiency can be achieved compared to non-adaptive fixed-item testing, multiple unidimensional adaptive tests, and MAT using the maximum likelihood approach (Frey & Seitz, 2010). In the Bayesian approach, the ability vector  $\boldsymbol{\theta}$  can be estimated by the mode of the posterior density function, which comprises both the prior density function that is based on the

variance-covariance matrix of the measured dimensions  $\Phi$  and on the likelihood function  $L(\mathbf{u}|\boldsymbol{\theta})$ , based on the responses  $\mathbf{u}$  given by the examinee (Segall, 2010). The items are selected by the maximum information criterion, which is quantified by the maximum of the following expression:

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

According to this criterion – which is also referred to as the  $D$ -optimality criterion – the item is selected that provides the maximum determinant of the matrix  $\mathbf{W}_{i+i^*}$ , which consists of the information matrix of previously administered items  $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)$ , the information matrix for the candidate item  $i^*$ ,  $\mathbf{I}(\boldsymbol{\theta}, u_{i^*})$ , and the inverse of the variance-covariance matrix  $\Phi^{-1}$ . In addition to the  $D$ -optimality criterion, other item selection methods have been proposed, such as maximizing the posterior expected Kullback-Leibler information (Veldkamp & van der Linden, 2002) or maximizing the mutual information between the current posterior distribution of  $\boldsymbol{\theta}$  and the response distribution on the candidate item (Mulder & van der Linden, 2010). Compared to other item selection methods, the  $D$ -optimality criterion performed well across a broad range of MAT configurations in terms of accuracy and precision of ability estimates (e.g., Mulder & van der Linden, 2009; Veldkamp & van der Linden, 2002; Wang & Chang, 2011; Wang, Chang, & Boughton, 2011; Yao, 2012, 2013, 2014). As these different selection methods are expected to perform similarly, and a comparison of item selection methods is not the subject of the present study, we chose the  $D$ -optimality criterion for the MAT in the current study.

In most of the studies that demonstrate the advantages of MAT in terms of measurement efficiency for both simulated and empirical data, the multidimensional constructs were modeled by using between-item models (e.g., Frey & Seitz, 2011; Frey et al., 2013; Haley, Pengsheng, Ludlow, & Fragala-Pinkham, 2006; Makransky, Mortensen, & Glas, 2012). Wang and Chen (2004) conducted a simulation to study the effect of several independent variables, e.g., multidimensionality (between-item, within-item) and correlation between dimensions (moderate, high), on measurement efficiency and item selection in MAT compared to unidimensional computerized adaptive testing (UCAT) and random item selection. The results revealed that the more dimensions that are measured simultaneously and the higher the dimensions are correlated with each other, MAT is found to be more efficient compared to UCAT or random item selection. In the within-item MAT condition, two-dimensional items were selected more often than one-dimensional items because they provided information about two abilities. Frey et al. (2013) used a between-item model to measure the 10 subdimensions of students' literacy in reading, mathematics and science that were assessed in the PISA within a real-data simulation. They showed that MAT can produce sufficiently reliable scores for all subdimensions, which even holds for 9 of the 10 subdimensions with the incorporation of the restrictions that are associated with the PISA (link items, open response format, item units, distinction between major and minor domains).



## Research questions

The capability of MAT to enhance measurement precision can be used to measure higher-dimensional competence constructs in an appropriate amount of time and to report more differentiated information. The potential of MAT to measure complex competence constructs with MIRT models with within-item multidimensionality has not as yet been thoroughly examined. The present study thus aims to answer the following main research question:

1. Is MAT capable of enhancing the measurement precision for all 11 subdimensions of mathematical competence that are described by the theoretical framework of the German Educational Standards in Mathematics for Secondary Education up to a range that makes reporting possible in a typical LSA setting? Here, a reliability value of at least .50 (without conditioning) is considered to be a minimum for LSA test score reporting and reliabilities of above .85 as an optimum. The threshold values of .50 and .85 are oriented on the PISA assessments in 2006, 2009, and 2012 (OECD, 2009, 2012, 2014), where the reliabilities (which are based on uni-dimensional scalings without conditioning) were in a range from .85 to .86 for the major domains and from .43. to .61 for the minor domains.

Further, the application of MAT can result in an undesired variation in the number of items that are presented per subdimension when the items are selected according to the statistical criterion of maximal information (e.g., Frey et al., 2013). This can lead to low reliabilities for subdimensions for which few items have been presented. In this study, we examine the severity of this potential issue with the use of a within-item MIRT model in MAT. The corresponding second research question is:

2. Does the application of MAT in the context of the German Educational Standards in Mathematics result in undesired proportions of items being presented per subdimension? The results will make it possible to decide whether a content management method is needed.

Lastly, a preferred selection of higher-dimensional items can be expected (Wang, Chen, & Cheng, 2004). Therefore, the following additional research question is examined:

3. Are items that load on several subdimensions selected more often in MAT than in a non-adaptive test?

## Method

The research questions were answered by means of a simulation study. The simulation study was configured under the objective of matching the conditions of the regular assessment of the German Educational Standards in Mathematics as closely as possible. The simulation was accomplished in several steps. First, items that are applicable for a computer-based assessment were selected from the paper-and-pencil test item pool that was developed in the course of the development of a testing procedure for the assessment of the German Educational Standards in Mathematics for Secondary Education (Prenzel

& Blum, 2007). Subsequently, the item parameter values and the multidimensional competence distribution that were estimated from a sample of students were used for the generation of person parameters and responses which – in turn – were used to simulate the testing procedure. The following subsections outline the design, the data basis, and the procedure of the simulation study.

## Design

Two different testing algorithms were compared in the study: the random administration of items (RA) and the MAT based on the Bayesian approach of Segall (1996). A detailed description of the testing algorithms can be found in the last subsection below.

As a dependent measure for precision (Research question 1), the reliability per subdimension was calculated. The reliability was estimated by the squared correlation of the true person parameters  $\theta$  and estimated person parameters  $\hat{\theta}$  for every subdimension  $k$  :

$$REL_k = r_{\hat{\theta}_k, \theta_k}^2. \quad (4)$$

In an additional condition (MAX), all items that were available for one subdimension were administered to obtain an upper bound for the subdimension-specific reliabilities to make it easier to gauge the effect of MAT on reliability.

For the conditions RA and MAT, the proportion of items that were selected for each subdimension ( $PCT_k$ ) was calculated to assess whether the proportions that resulted from the item selection algorithm in the MAT condition would differ substantially from the proportions in the RA condition (Research question 2). The  $PCT_k$  was computed by the average number of items that were presented for this subdimension,  $T_k$ , and the average number of items that were presented in total,  $T$ , multiplied by 100:

$$PCT_k = \frac{100}{N} \sum_{j=1}^N \frac{T_{jk}}{T_j}. \quad (5)$$

Further, the proportion of items depending on the assigned number of process-related competencies  $l$  ( $PCT_l$ ), with  $l$  ranging from 0 to 4, was calculated to examine whether the MAT algorithm would lead to a preferred selection of items that are assigned to several subdimensions compared to RA (Research question 3). The  $PCT_l$  was computed by the average number of the presented items that were assigned to  $l$  process-related competencies,  $T_l$ , and the average number of items that were presented in total,  $T$ , multiplied by 100:

$$PCT_l = \frac{100}{N} \sum_{j=1}^N \frac{T_{jl}}{T_j}. \quad (6)$$

## Data basis

The present study is based on a pool of 290 items and the responses of 9,577 students to these items, gathered in the course of the development of the test to measure the attainment of the German Educational Standards in Mathematics for Secondary Education (Prenzel & Blum, 2007). The MIRT model, the item parameters, and the 11-dimensional latent multidimensional distribution that were used for MAT in the current study were adopted from the study of Mikolajetz and Frey (2014). The mean vector  $\boldsymbol{\mu}$  of the latent multidimensional distribution was set to zero for all of the means, whereas the variance-covariance matrix  $\boldsymbol{\Phi}$  was estimated from the response data. Table 1 contains the variances, covariances, and correlations for the 11 sub dimensions. As seen from the table, the correlations between the subdimensions with regard to mathematical content areas C1 to C5 were higher – which ranged from .76 to .84 – than the correlations between the general mathematical competencies G1 to G6 – which ranged from .19 to .67.

**Table 1:**

Variances, covariances and correlations for the five mathematical content areas (C1 – C5) and the six general mathematical competencies (G1 – G6) as subdimensions of mathematical competence

	Subdimension										
	C1	C2	C3	C4	C5	G1	G2	G3	G4	G5	G6
C1	1.02	0.77	0.62	0.88	0.71	0.71	0.56	0.85	0.61	0.72	0.29
C2	.83	0.86	0.62	0.76	0.55	0.57	0.50	0.69	0.58	0.52	0.21
C3	.76	.84	0.65	0.68	0.55	0.62	0.59	0.58	0.47	0.57	0.26
C4	.84	.79	.81	1.08	0.71	0.69	0.55	0.77	0.59	0.69	0.26
C5	.78	.67	.76	.76	0.80	0.55	0.48	0.58	0.44	0.47	0.22
G1	.77	.68	.84	.73	.67	0.83	0.48	0.61	0.54	0.62	0.41
G2	.54	.53	.72	.51	.52	.51	1.07	0.64	0.43	0.43	0.16
G3	.84	.74	.72	.74	.64	.67	.61	1.01	0.65	0.55	0.23
G4	.63	.64	.60	.59	.50	.61	.43	.66	0.94	0.47	0.23
G5	.69	.54	.68	.64	.51	.65	.40	.52	.47	1.07	0.37
G6	.36	.28	.39	.30	.31	.55	.19	.29	.30	.44	0.66

*Note.* Values below the main diagonal are correlations, values above the main diagonal are covariances.

## Procedure

### *Item pool compilation*

To make it possible to draw direct conclusions from the simulation study about a real MAT-based assessment, only the items that can be presented on a computer were selected. For example, items that require geometrical constructions with a set square or a compass were excluded. In total, 253 items were selected, including 141 items in multiple-choice format and 112 items in short or open response format. All of the items in the short or open response formats were examined to determine whether they could be scored directly by a computer and, hence, provide information for the adaptive testing algorithm during a testing session. A total of 59 of the items require entering just a digit or marking the answer on the stimulus (e.g., area or point in a coordinate system) that would make feasible direct scoring by the computer with an acceptable amount of programming. In contrast, 53 items require longer written responses. Although these items, which make up 21 % of the MAT item pool, can easily be presented to the participants during an adaptive test on the computer, they must be scored manually after a testing session.

The number of items that were assigned to mathematical content areas and general mathematical competencies that were used for the simulation is presented in Table 2. Each item is assigned to only one of the five mathematical content areas and to one or more of the six general mathematical competencies that are described in the theoretical framework – except for 18 items which are assigned to only a mathematical content area.

Regarding the distribution of item difficulties, the frequency of items with a medium difficulty is high, whereas the frequency of items decreases toward the extremes of the difficulty scale. For the mathematical content areas C1 to C5, the mean discrimination

**Table 2:**  
Assignment of the 253 items in the pool to the mathematical content areas and the general mathematical competencies

General mathematical competency	Mathematical content area					Items
	C1	C2	C3	C4	C5	
G1	12	2	4	3	10	31
G2	5	9	4	6	3	27
G3	28	7	1	26	23	85
G4	7	5	13	19	11	55
G5	21	17	10	26	7	81
G6	17	7	3	16	7	50
None	2	3	8	2	3	18
Items	71	40	35	69	38	

values – which ranged from 0.70 to 1.10 – are higher than for those for the general mathematical competencies G1 to G6 – which ranged from 0.50 to 0.67. The standard deviation for C1 to C5 is zero because the discrimination parameters within the mathematical content areas were set to be equal. The discrimination values for G1 to G6 vary, with standard deviations from 0.30 to 0.78.

#### *Generation of person parameters and responses*

Person parameters were generated for the 11 subdimensions for 1,000 simulees using the estimated latent multidimensional distribution  $\theta \sim MNV(\mu, \Phi)$  that is described above. In both conditions, MAT and RA, a response is needed from every person to every item because, theoretically, every item could be administered to every person. Therefore, item and person parameters were used to generate responses for all of the simulees to all of the items using the 11-dimensional model that is described above. To account for the statistical uncertainty of the simulated answering process, 100 replications of a complete response matrix were generated, which were then used for the simulation of the testing procedure. The final statistics were calculated by averaging the statistics of the replications.

#### *Simulation of the testing procedure*

In the RA condition, the procedure of the non-adaptive paper-and-pencil test administration was simulated. The items were selected at random from the item pool, whereby the proportions of the items that were presented per content-related competency were adopted from the original assessment by Prenzel and Blum (2007). The person estimates for the 11 subdimensions were calculated after the test administration.

In the MAT condition, the items were adaptively selected using the Bayesian MAT approach of Segall (1996). For each person, the first item to be administered was chosen randomly from the item pool. The adaptive item selection subsequently started by making use of the item parameters, the provisional ability estimates and the variance-covariance matrix of the latent multidimensional distribution  $\Phi$ . At every step of the test, the item was selected that maximized Equation 3.

In both conditions, the estimation of provisional and final ability parameters was based on the responses to the administered items and the latent multidimensional distribution  $\Phi$ . The testing procedures in RA and MAT were realized by taking into account the restrictions of the original assessment, such as the grouping of items to units (so-called testlets) and the presence of items that cannot be scored directly by a computer. If an item that had been selected either randomly or adaptively was grouped with other items, the complete testlet was presented to the examinee. Items with a response format that makes a direct scoring by the computer impossible were also included but not used for provisional person parameter estimates in the MAT condition. The responses to those items were incorporated in the final ability estimation that was carried out after the test administration. In both conditions, the test administration was terminated after 60 items had been administered, which corresponds with a testing time of approximately 120 minutes for the original assessment.

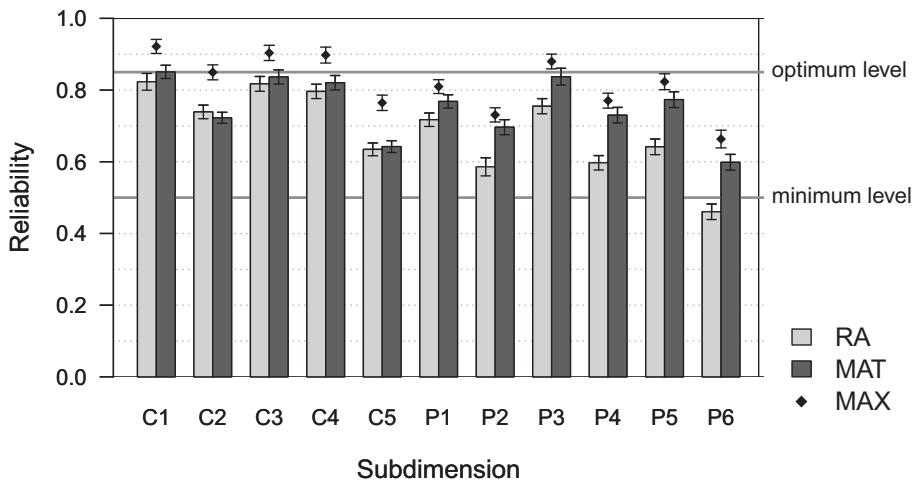
In the MAX condition, the final ability parameters were estimated based on the responses to all of the items of the item pool and the latent multidimensional distribution  $\Phi$ .

## Results

Our main research question asks whether the reliability of all 11 subdimensions of mathematical competence that are considered by the German Educational Standards in Mathematics for Secondary Education can be increased to an optimum value of at least .85 or – if the latter is not the case – up to a minimum value of .50. The reliability estimates for the 11 subdimensions that resulted from the simulation study are shown in Figure 2.

For 5 of the 11 subdimensions (C1, C2, C3, C4, and P3), the upper bound reliability values that were obtained by administering all of the items that were available in the item pool for the respective subdimension (MAX) the optimum level of .85 was reached, whereas for the other subdimensions, the values remained under the optimum value. However, the minimum level of .50 could be reached with the items at hand for all of the subdimensions.

As a general trend, the reliability estimates for MAT were higher ( $M = .76$ ,  $SD = 0.19$ ) than those that were obtained by RA ( $M = .70$ ,  $SD = 0.21$ ). The only exception was the subdimension C2, for which the reliability in the MAT condition was slightly lower, compared to RA. In particular, the mean reliability for the general mathematical compe-



**Figure 2:**

Reliability coefficients  $\pm 1$  SE for the five mathematical content areas (C1 to C5) and the six general mathematical competencies (G1 to G6) as subdimensions of mathematical competence for random administration of items (RA), multidimensional adaptive testing (MAT), and the maximum possible reliability by administering all of items that are available in the item pool (MAX)

tencies (G1 to G6) improved from .64 ( $SD = 0.17$ ) to .74 ( $SD = 0.18$ ) by using MAT, and the reliability values were very close to the upper bound reliability values that resulted from the MAX condition. While in the RA condition, the reliability of one subdimension did not exceed .50, and this minimum requirement for test score reporting was reached by MAT for all of the subdimensions. Nevertheless, although the reliability for some of the subdimensions came close to .85, this optimal value was not reached.

The second research question asks whether the application of MAT results in undesired proportions of items that are presented per subdimension. To answer this question, please review Table 3, which contains the proportions of the items that were presented per subdimension for the two conditions.

For the mathematical content areas, the numbers of items that were presented in MAT were comparable to those for RA, except for C2, where the number of administered items decreased substantially from 15.8 % in RA to 8.6 % in MAT, and for C4, where the number increased substantially from 27.0 % in RA to 36.9 % in MAT. Note that the reliability remained largely unaffected by the changes in the number of items that were presented per subdimension (Figure 2). Regarding the number of administered items for the six general mathematical competencies G1 to G6, the number of items that were administered per subdimension increased substantially for all of the competencies except for G6. Thus, using MAT instead of RA only slightly affected the proportions of the items that were presented for the 11 subdimensions. For all of the subdimensions, a reasonable average number of items was presented if the adaptive item selection is solely based on the criterion of maximum statistical information.

The third research question asks whether the items loading on several subdimensions are selected more often in MAT than in RA. It was assumed that higher-dimensional items are more frequently selected in MAT compared to a random selection. The results that are shown in Table 4 support this assumption. Items with two or three assignments to the general mathematical competencies, in addition to their loading on a mathematical content area, were selected more often in MAT than in RA. In contrast, the items with no or one assignment to the general mathematical competencies were selected more often in RA than in MAT.

**Table 3:**

Proportions of items per subdimension ( $PCT_k$ ) for random administration of items (RA) and multidimensional adaptive testing (MAT)

Condition	Proportion of items per subdimension										
	Mathematical content area					General mathematical competency					
	C1	C2	C3	C4	C5	G1	G2	G3	G4	G5	G6
RA	23.9	15.8	17.4	27.0	15.8	12.7	13.0	29.4	24.1	28.6	23.2
MAT	23.8	8.6	15.8	36.9	15.0	14.5	15.3	33.8	26.2	32.7	21.9

**Table 4:**

Proportions of administered items according to the number of assigned subdimensions for the general mathematical competencies G1 to G6 (*PCT*)

Condition	Number of loadings				
	0	1	2	3	4
RA	7.3	60.7	26.4	4.6	0.9
MAT	3.5	56.2	33.6	5.8	0.9

*Note.* RA = random administration of items; MAT = multidimensional adaptive testing.

## Discussion

The present simulation study aimed to answer the main question of whether MAT is capable of enhancing the measurement precision for all 11 subdimensions of mathematical competence that are considered by the German Educational Standards in Mathematics for Secondary Education up to a range that is comparable to the reliability values that are usually deemed to be appropriate for test score reporting in LSAs. Therefore, a reliability of at least .50 was regarded as the minimum, while a reliability of at least .85 was regarded as being optimal. If these goals are reached, the results that are related to the mathematical content areas could be complemented by the results that are related to the general mathematical competencies by using a more complex MIRT model within MAT. According to the results, when the available item pool is used, MAT produces reliability coefficients with  $M = .76$  and with values of at least .50 for all of the 11 subdimensions in contrast to  $M = .70$  in the RA condition, with 10 of 11 subdimensions having values of at least .50. Hence, the standard that was set by the PISA assessments regarding the reporting for minor domains and that were considered to be the minimum level that was necessary for reporting was reached for every subdimension. However, neither MAT nor random item administration (RA) would produce reliability coefficients of at least .85 for any of the 11 subdimensions. To assess whether .85 was a realistic target with the available item pool, we estimated the upper bound reliability from the responses to all of the items of the item pool. Only for five subdimensions did the upper bound reliability values reach the optimum level of .85. For four of those subdimensions, MAT produced values that were close to the optimum level. Gains in measurement precision were observed, especially for the general mathematical competencies. Here, the reliability was increased for all six subdimensions, on average, from .64 for RA to .74 for MAT, and the reliability values were close to the upper bound reliability values. This can be attributed to some specific characteristics that were related to the general mathematical competencies. As the mean reliability in RA for the general mathematical competencies ( $M = .64$ ) is substantially lower than the mean reliability for the mathematical content areas ( $M = .79$ ), there is greater potential for improvement in the MAT condition regarding the reliability coefficients of the general mathematical competencies. Further, some properties of the modeling approach enhance the potential of MAT to select particularly informative items regarding the general mathematical competencies. First, the MIRT model in this study includes equal discrimination parameters for the mathematical content



areas, whereas the discrimination parameters for the general mathematical competencies were not constrained to an equal value. For the latter, the variation of discrimination parameters resulted in the availability of items with discrimination parameters that were significantly above the average and hence particularly informative. Second, the gains in reliability can be partly explained by the preference of the MAT algorithm for items that measure the standing of the examinees on several general mathematical competencies (i.e., with several loadings on the general mathematical competencies), which was the subject of the third research question. This finding is straightforward because every loading of an item on the subdimensions contributes to the item information that is used for item selection. Finally, the second research question asked whether the application of MAT would result in undesired proportions of the presented items per subdimension. If the items are only selected based on the criterion from Equation 3, large variations in the number of items that are presented per subdimension might be the consequence. Here, especially the proportions of items that are presented per mathematical content area showed more variation in MAT (e.g., 8.6 % for C2 and 36.9 % for C4) than in RA (15.8 % for C2 and 27 % for C4), which, however, did not noticeably affect measurement precision. Even when a lower proportion of items was presented, as with the content-related competency C2, this had no substantial effect on the reliability. Because a reasonable average number of items was presented per subdimension, no content management method seems to be necessary in the present case.

The differentiated results that were achieved with MAT could be used to report on both the content-related and process-related subdimensions of mathematical competence and provide a more comprehensive coverage of the theoretical construct of mathematical competence. Reporting not only on the five mathematical content areas but also on the six general mathematical competencies provides more information for policy-makers and researchers regarding the degree to which educational objectives have been fulfilled. The present study is the first to show an applicable method that is capable of measuring all of the 11 subdimensions within a reasonable testing time.

Although the results regarding reliability are generally promising, it should be noted that additional improvements in measurement precision are possible. Three ways to further increase measurement precision should be mentioned here. The first addresses the fact that the results of this study were obtained by using an item pool that was not explicitly developed for adaptive testing. The frequency distribution of item difficulties with a strong decrease towards the extremes is not optimal for a computerized adaptive test. To tap the full potential of MAT, more items with extreme difficulties are required. If the reporting on all 11 subdimensions is desired, future item development should aim to produce sufficient items on the entire range of abilities for all of the subdimensions. Second, additional gains in reliability could be expected if an item pool that contains fewer or at least smaller testlets as well as more items that could be directly scored on a computer were to be used. Third, the presented results are based on a MIRT scaling without conditioning. In the case that results are needed to be reported only at the group level, a background model could be added, as is typically done for LSAs, and additional gains in reliability could be achieved.

Although the simulation is based on the conditions of one LSA, similar improvements can be assumed when other complex competence constructs are measured (i.e., an increase in the number of subdimensions that can be measured with sufficient reliability). Because the gains in measurement efficiency that can be achieved by using MAT are related to the size of the correlation between the measured subdimensions, these gains will be higher the stronger the measured subdimensions are correlated. Similarly, this beneficial feature can be used in the measurement of multiple correlated competence constructs and the administration of items across content domains as in the PISA, which was already demonstrated by Frey and Seitz (2011) and Frey et al. (2013). Another issue that is associated with the generalizability of the findings to other LSAs is that the results of the present paper are based on the item pool of one specific LSA. The results regarding the efficiency of MAT will most probably differ with the use of a different item pool. However, the introduction of MAT in a LSA program will result in the development of an item pool that will be more suitable for adaptive testing – and will lead to improved measurement efficiency. To pinpoint the exact effect of the item pool characteristics on measurement efficiency with MAT in the planning phase of a LSA, pre-operational simulation studies are recommended.

Another issue to be addressed is the preferred selection of higher-dimensional and highly discriminating items by the MAT algorithm that will inevitably lead to higher exposure rates for these items. First, this can lead to undesired shifts in the construct coverage. In this study, the items that were assigned to the content-related competencies C1 and C2 were more likely to be excluded in the MAT condition, while the items that were assigned to the remaining eight subdimensions were equally or more likely to be selected (see Figure 3 in the appendix), thus a slight shift in the construct coverage can be stated. Second, high exposure rates can jeopardize test security. To avoid shifts in the construct coverage and to preserve test security, exposure control strategies should be considered (cf. Leroux, Lopez, Hembry, & Dodd, 2013 for an overview of exposure control strategies) in cases where test security and comparable item exposure rates are considered to be crucial for a study. Because the introduction of exposure control imposes an additional constraint on item selection, losses in measurement precision may result. However, in conjunction with an optimized item pool for MAT, as described above, the negative consequences of exposure control for measurement precision could most probably be held at an acceptable level.

In this study, a content balancing mechanism was not implemented in the item selection process of MAT. Nonetheless, such a mechanism might be desirable to ensure that a certain proportion of items per subdimension should be presented to each examinee. Cheng and Chang (2009) proposed a heuristic approach, i.e., the maximum priority index (MPI) method, to include non-statistical constraints, such as content balancing, in computerized adaptive testing. The MPI is designed to maintain the proportions of administered items with specific characteristics – for instance, the subdimension that is measured by an item – in accordance with the proportions that are prespecified by the test developer. The generalization of this approach to multidimensional testing, i.e., the multidimensional MPI (MMPI), was introduced by Frey, Cheng, and Seitz (2010) and applied successfully by Frey et al. (2013) to reduce variations in the reliability coefficients for all 10

subdimensions of students' literacy in reading, mathematics and science that is considered in the PISA (see also Born & Frey, 2016 for a comparison of the MMPI with other content management methods). The MMPI, however, is only optimal for balancing subdimensions in between-item multidimensionality applications. Future work should focus on developing a content balancing mechanism for applications with MIRT models with within-item multidimensionality.

Finally, another challenge lies in the extension of the item pool because in many applications items must be added to the item pool over time for several reasons (e.g., replacing items with outdated content, which are overexposed, or which show item parameter drift). For cases in which such problems occur, new items can be placed on the already established MIRT scales by using MAT online-calibration methods (e.g., Chen & Wang, 2016; Chen & Xin, 2013). With these methods, new items are seeded in the operational testing phase to avoid expensive separate calibration studies.

In conclusion, the present simulation study illustrates the potential of MAT for the modeling of complex competence constructs as they are often measured in LSAs. Although some methodological challenges should be addressed in the future, MAT could be regarded as a highly promising and efficient methodology to conduct differentiated assessments of complex competence constructs.

## References

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37–51. doi:10.1111/j.1745-3992.2003.tb00136.x
- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, 31, 162–172. doi:10.1016/j.stueduc.2005.05.008
- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–23. doi:10.1177/0146621697211001
- Blum, W., Drüke-Noe, C., Leiß, D., Wiegand, B., & Jordan, A. (2005). Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht [The role of educational standards for quality development in mathematics instruction]. *ZDM – The International Journal on Mathematics Education*, 37, 267–274. doi:10.1007/BF02655814
- Born, S., & Frey, A. (2016). *Heuristic constraint management methods in multidimensional adaptive testing*. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164416643744.
- Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, 81, 674–701. doi:10.1007/s11336-015-9482-9
- Chen, P., & Xin, T. (2013, July). *Developing online calibration methods for multidimensional computerized adaptive testing*. Paper presented at the 78th Meeting of the Psychometric Society, Arnhem, the Netherlands.

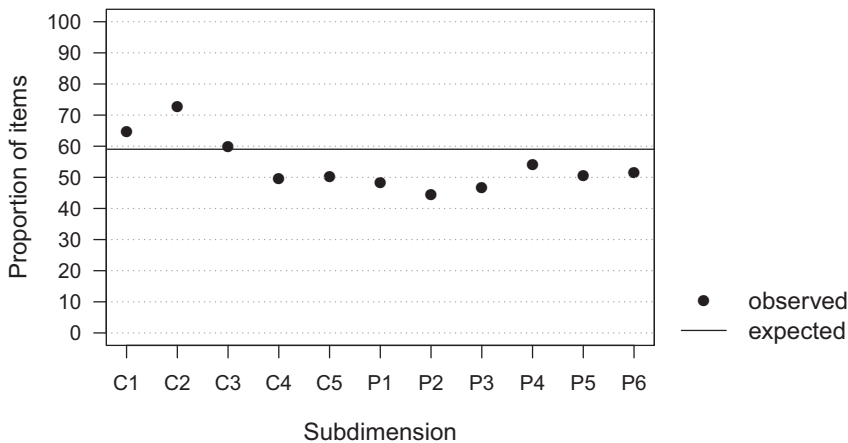
- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369–383. doi:10.1348/000711008X304376
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Frey, A., Cheng, Y., & Seitz, N. N. (2010, June). *Content balancing with the maximum priority index method in multidimensional adaptive testing*. Paper presented at the conference of the International Association for Computerized Adaptive Testing (IACAT), Arnhem, the Netherlands.
- 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. doi:10.1016/j.stueduc.2009.10.007
- Frey, A., & Seitz, N. N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz [Multidimensional adaptive testing of competencies: Results regarding measurement efficiency]. *Zeitschrift für Pädagogik, Beiheft*, *56*, 40–51.
- Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*, *71*, 503–522. doi:10.1177/0013164410381521
- Frey, A., Seitz, N. N., & Kröhne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA* (pp. 103–120). Dordrecht: Springer.
- Haley, S. M., Pengsheng, N., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation*, *87*, 1223–1229. doi:10.1016/j.apmr.2006.05.018
- Hartig, J., & Frey, A. (2012). Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik [Validity of a standard-based test for mathematical competencies]: Zusammenhänge mit den bei PISA gemessenen Kompetenzen und Varianz zwischen Schulen und Schulformen [Relations with the competencies assessed in PISA and variance between school and school tracks]. *Diagnostica*, *58*, 3–14. doi:10.1026/0012-1924/a000064
- Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs [Test design and analysis for the comparison of the German federal states]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [Comparison of the German federal states 2012. Competencies in mathematics and science at the end of secondary education] (pp. 391–402). Münster: Waxmann.
- Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139–155. doi:10.1037/0033-2909.107.2.139

- Katzenbach, M., Blum, W., Druke-Noe, C., Keller, K., Köller, O., Leiss, D., ... Roppelt, A. (2009). *Bildungsstandards: Kompetenzen überprüfen. Mathematik Sekundarstufe I*. Handreichung. [Educational standards: Assessing competencies. Mathematics in secondary education. A guidance]. Berlin: Cornelsen.
- Kultusministerkonferenz. (2003). *Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10)* [Agreement on educational standards for secondary education]. Retrieved from [http://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2003/2003\\_12\\_04-Bildungsstandards-Mittleren-SA.pdf](http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mittleren-SA.pdf)
- Kultusministerkonferenz. (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss* [Educational standards in mathematics for secondary education]. Neuwied: Luchterhand.
- Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, 73, 857–874. doi:10.1177/0013164413486802
- Lindberg, S. M., Hyde, J. S., Petersen, J. L. & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136, 1123–1135. doi:10.1037/a0021276
- Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the Neo Pi-R. *Assessment*, 20, 3–13, doi:10.1177/1073191112437756
- McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, 28, 1–14. doi:10.1037/h0034092
- Mikolajetz, A., & Frey, A. (2014, March). *Simultane Messung von mathematischen Leitideen und Kompetenzen* [Simultaneous measurement of content-related and process-related competencies in mathematics]. Paper presented at the meeting of the Gesellschaft für Empirische Bildungsforschung [Society for empirical educational research], Frankfurt, Germany.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Organisation for Economic Co-operation and Development. (2003). *The PISA assessment framework: mathematics, reading, science and problem solving knowledge and skills*. Paris.
- Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris.
- Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report*. Paris.
- Organisation for Economic Co-operation and Development. (2013a). *PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. Paris.
- Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. Paris.

- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [Comparison of the German federal states 2012. Competencies in mathematics and science at the end of secondary education]. Münster: Waxmann.
- Prenzel, M., & Blum, W. (2007). *Entwicklung eines Testverfahrens zur Überprüfung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss: Technischer Bericht* [Development of a test for the assessment of the German Educational Standards in Mathematics for Secondary Education: Technical report]. Kiel: Leibnitz-Institut für Pädagogik der Naturwissenschaften und Mathematik (IPN).
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Roppelt, A., Blum, W., & Pöhlmann, C. (2013). Beschreibung der untersuchten mathematischen Kompetenzen [Specification of the assessed mathematical competencies.]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [Comparison of the German federal states 2012. Competencies in mathematics and science at the end of secondary education] (pp. 23–37). Münster: Waxmann.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi:10.1007/BF02294343
- Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 57–75). New York, NY: Springer.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, *48*, 73–86. doi:10.1080/00461520.2013.779483
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- van den Heuvel-Panhuizen, M. (2004). Girls' and boys' problems: Gender differences in solving problems in primary school mathematics in the Netherlands. In B. Clarke, D. M. Clarke, G. Emanuelsson, B. Johansson, D. V. Lambdin, F. K. Lester, . . . K. Wallby (Eds.), *International perspectives on learning and teaching mathematics* (pp. 237–252). Göteborg: National Center for Mathematics Education.
- van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412. doi:10.3102/10769986024004398
- van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Zeitschrift für Psychologie/Journal of Psychology*, *216*, 3–11. doi:10.1027/0044-3409.216.1.3
- Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. doi:10.1007/BF02295132
- Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing – gaining information from different angles. *Psychometrika*, *76*, 363–384. doi:10.1007/S11336-011-9215-7

- Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13–39. doi:10.1007/s11336-010-9186-0
- Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multi-dimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295–316. doi:10.1177/0146621604265938
- Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116–136. doi:10.1037/1082-989X.9.1.116
- Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: theory and applications. *Psychometrika*, *77*, 495–523. doi:10.1007/s11336-012-9265-5
- Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3–23. doi:10.1177/0146621612455687
- Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement*, *51*, 18–38. doi:10.1111/jedm.12032

## Appendix



**Figure 3:**

Proportion of items per subdimension that were observed in the MAT (multidimensional adaptive testing) condition relative to the total number of items that are assigned to the respective subdimension for items with an exposure rate < 15 %. The expected proportion corresponds to the average proportion of items with an exposure rate < 15 %