NEPS SURVEY PAPERS

Luise Fischer, Theresa Rohm, and Timo Gnambs

# NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 4 FOR GRADE 12

NEPS Survey Paper No. 12
Bamberg, January 2017

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

**Editor-in-Chief**: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

**Contact**: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg − Germany − contact@lifbi.de

# NEPS Technical Report for Mathematics:

# Scaling Results of Starting Cohort 4 for Grade 12

*Luise Fischer, Theresa Rohm, and Timo Gnambs*

*Leibniz Institute for Educational Trajectories, Bamberg, Germany*

**E-mail address of lead author:**

luise.fischer@lifbi.de

**Bibliographic data:**

Fischer, L., Rohm, T., & Gnambs, T. (2017). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 12). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP12:1.0

**Acknowledgements:**

This report is an extension to NEPS working paper 22 (Duchhardt & Gerdes, 2013) that presents the scaling results for mathematical competence of starting cohort 4 for grade 9. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (Duchhardt & Gerdes, 2013) to facilitate the understanding of the presented results.

We thank Anna Scharl and Micha Freund for their assistance in scaling the data.

# NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 12

## Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the mathematical competence test in grade 12 of starting cohort 4 (ninth grade). The mathematical competence test contained 31 items (distributed among an easy and a difficult booklets containing 21 items each) with different response formats representing different cognitive requirements and different content areas. The test was administered to 5,733 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that all items but one fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the number of items targeted toward a lower and higher mathematical ability as well as the large percentage of items in the difficult booklet at the end of the test that were not reached due to time limits. Further challenges related to the dimensionality analyses based on the four content areas. Overall, the mathematics test had acceptable psychometric properties that allowed for a reliable estimation of mathematics competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the ConQuest-syntax for scaling the data.

## Keywords

item response theory, scaling, mathematical competence, scientific use file

# Content

1. Introduction ........ 4
2. Testing Mathematical Competence ........ 4
3. Data ........ 5
3.1 The Design of the Study ........ 5
3.2 Sample ........ 6
4. Analyses ........ 7
4.1 Missing Responses ........ 7
4.2 Scaling Model ........ 8
4.3 Checking the Quality of the Test ........ 8
4.4 Software ........ 9
5. Results ........ 10
5.1 Missing Responses ........ 10
5.1.1 Missing responses per person ........ 10
5.1.2 Missing responses per item ........ 12
5.2 Parameter Estimates ........ 15
5.2.1 Item parameters ........ 15
5.2.2 Test targeting and reliability ........ 17
5.3 Quality of the test ........ 19
5.3.1 Fit of the subtasks of the complex multiple choice item ........ 19
5.3.2 Item fit ........ 19
5.3.3 Distractor analyses ........ 19
5.3.4 Differential item functioning ........ 19
5.3.5 Rasch-homogeneity ........ 24
5.3.6 Unidimensionality ........ 24
6. Discussion ........ 26
7. Data in the Scientific Use File ........ 26
7.1 Naming conventions ........ 26
7.2 Linking of competence scores ........ 26
7.2.1 Samples ........ 27
7.2.2 The design of the link study ........ 27
7.2.3 Results ........ 27
7.3 Mathematical competence scores ........ 29

# 1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper the results of these analyses are presented for mathematical competence in starting cohort 4 (ninth grade) in grade 12. First, the main concepts of the mathematics competence test are introduced. Then, the mathematics competence data of starting cohort 4 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamental changes in the presented results.

# 2. Testing Mathematical Competence

The framework and test development for the test of mathematical competence are described in Weinert et al. (2011), Neumann et al. (2012), and Ehmke et al. (2009). In the following, we briefly describe specific aspects of the mathematics test that are necessary for understanding the scaling results presented in this paper.

In the test, students usually face a certain situation followed by only one task related to it; sometimes there are two tasks. Each of the items belongs to one of the following content areas, namely, (a) quantity, (b) space and shape, (c) change and relationships, and (d) data and chance. Furthermore, the framework also describes as a second and independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

The mathematical competence test included three types of response formats: simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR). In MC items, the test taker had to identify the correct answer from several, usually four, response options. In CMC tasks, a number of subtasks with two response options were presented. SCR items required the test taker to write down an answer into an empty box. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt, and Weinert (2012).

The competence test for mathematics that was administered in the present study included 30 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses identified a poor fit for one item (mag12d071_c). Therefore, this item was removed from the final scaling procedure. Furthermore, while one item (mag9r051_sc4g12_c) showed differential item functioning (DIF) between the easy and difficult test, another item (mag9d201_sc4g12_c) showed DIF between the assessment settings (see section 3.1). Consequently, the two respective items were treated as test unique items. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 31 items.

# 3. Data

## 3.1 The Design of the Study

The study followed a three-factorial (quasi-)experimental design. These factors referred to (a) the difficulty of the administered test, (b) the assessment setting (i.e., the context of test administration), and (c) the position of the mathematics test within the test battery.

In order to measure participants' mathematical competence with great accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies, two different versions of the mathematics competence test were developed that differed in their average difficulty (i.e., an easy and a difficult test). Both tests included 21 items that represented the four content areas (see Table 1). Twelve items were identical in both test versions (see Table 1), whereas 9 items were unique to the easy and the difficult test. The different response formats of the items are summarized in Table 2. The CMC item consisted of 3 subtasks.

Table 1

*Number of Items for the Different Content Areas by Difficulty of the Test*

| **Text types** | **Easy test** | **Both tests** | **Difficu lt test** |
|---|---|---|---|
| Quantity | 3 | 3 | 2 |
| Space and Shape | 2 | 3 | 3 |
| Change and Relationships | 2 | 4 | 1 |
| Data and Chance | 2 | 2 | 3 |
| Total number of items | 9 | 12 | 9 |

The panel study aimed at retesting all students that were initially included in the starting cohort 4 for ninth grade (see Duchhardt & Gerdes, 2013). Because some students left their original schools during the course of the longitudinal study, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings:

Students that remained at the same school as in the first assessment were tested at school in a group setting; in contrast, students that left their original school were tracked and, subsequently, individually tested at home (for details regarding the data collection process see the respective field report for wave 7). Thus, the context of test administration differed between the two groups.

Table 2

*Number of Items for the Different Response Formats by Difficulty of the Test*

| **Response format** | **Easy test** | **Difficult test** |
|---|---|---|
| Simple multiple choice items | 20 | 18 |
| Complex multiple choice items | - | 1 |
| Short Constructed Response | 1 | 2 |
| Total number of items | 21 | 21 |

The study assessed different competence domains including, among others, mathematics, computer literacy, and reading competence. The competence tests for these three domains were always presented first within the test battery. However, for students that were individually tested the tests were administered to participants in different sequence (see Table 3). For each participant the mathematics test was either administered as the first or the second test (i.e., after the computer literacy or the reading test). There was no multi-matrix design regarding the order of the items *within* a specific test. All subjects received the test items in the same order.

## 3.2 Sample

A total of 5,733 individuals received the mathematics competence test. Since at least three valid item responses were available for all subjects, the analyses presented in this paper are based on the whole sample of 5,733 individuals. The number of participants within each (quasi-)experimental condition is given in Table 3. While students that remained in their original school only were administered the difficult test, the students that had left their original schools were assigned either to the easy or the difficult test, based on the type of school they were attending in the prior wave (Duchhardt & Gerdes, 2013). Participants that attended a school leading to a high school graduation received the difficult test; the easy test was administered otherwise. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 3

*Number of Participants by the (Quasi-)Experimental Conditions*

| *Assessment setting:* | | **At school** | | **At home** | | **Total** |
|---|---|---|---|---|---|---|
| | *Test position*: | **first position** | **second position** | **first position** | **second position** | |
| *Test difficulty* | **Easy test** | - | - | 1,030 | 263 | 1,293 |
| | **Difficult test** | - | 3,900 | 345 | 195 | 4,440 |
| | **Total** | - | 3,900 | 1,375 | 458 | 5,733 |

# 4. Analyses

## 4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and, finally, e) multiple kinds of missing responses within CMC items that are not determined.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. Because of the multi-stage testing design 20 items were not administered to all participants. For respondents receiving the easy test 10 difficult items were missing by design, whereas 10 easy items were missing by design for respondents receiving the difficult test (see Table 1). As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

## 4.2 Scaling Model

Item and person parameters were estimated using a partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC item was scored as missing. Categories of polytomous variables with less than *N* = 200 responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category (see Appendix A).

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous item was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Mathematical competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012) while the data available in the SUF is described in section 7.

## 4.3 Checking the Quality of the Test

The mathematics competence test was specifically constructed to be implemented in NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

Before aggregating the subtasks of the CMC item to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, point-biserial correlations of the correct responses with the total score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to generate polytomous variables that were included in the final scaling model.

The MC items consisted of one correct response and one or more distractors (i.e., incorrect response options). The quality of the distractors within MC items was examined using the point-biserial correlation between an incorrect response and the total score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to a polytomous variable, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the corrected discrimination as computed in ConQuest) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall, judgment of the fit of an item was based on all fit indicators.

The mathematics competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they are easier for male than for female participants though being equal in ability), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables test position, gender, school types (high school vs. vocational school), the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). Moreover, in light of the quasi-experimental design measurement invariance analyses were also conducted for the test difficulty and administration setting. Differential item functioning (DIF) analyses were estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy of further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The mathematics competence test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The test was constructed to measure a unidimensional mathematical competence score. The assumption of unidimensionality was investigated by specifying a four-dimensional model based on the four different content areas. Every item was assigned to one content area (between-item-multidimensionality). The correlations among the dimensions as well as differences in model fit between the unidimensional model and the respective multidimensional models were used to evaluate the unidimensionality of the test. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) $Q_3$. Because in case of locally independent items, the $Q_3$ statistic tends to be slightly negative, we report the corrected $Q_3$ that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of $Q_3$ falling below .20 indicate essential unidimensionality.

## 4.4 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

# 5. Results

## 5.1 Missing Responses

### 5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by experimental condition (i.e., test difficulty and administration setting). Overall, there were very few invalid responses. Between 96% and 99% of the respondents did not have any invalid response at all; overall less than one percent had more than one invalid response. There was no difference in the amount of invalid responses between the different experimental conditions.

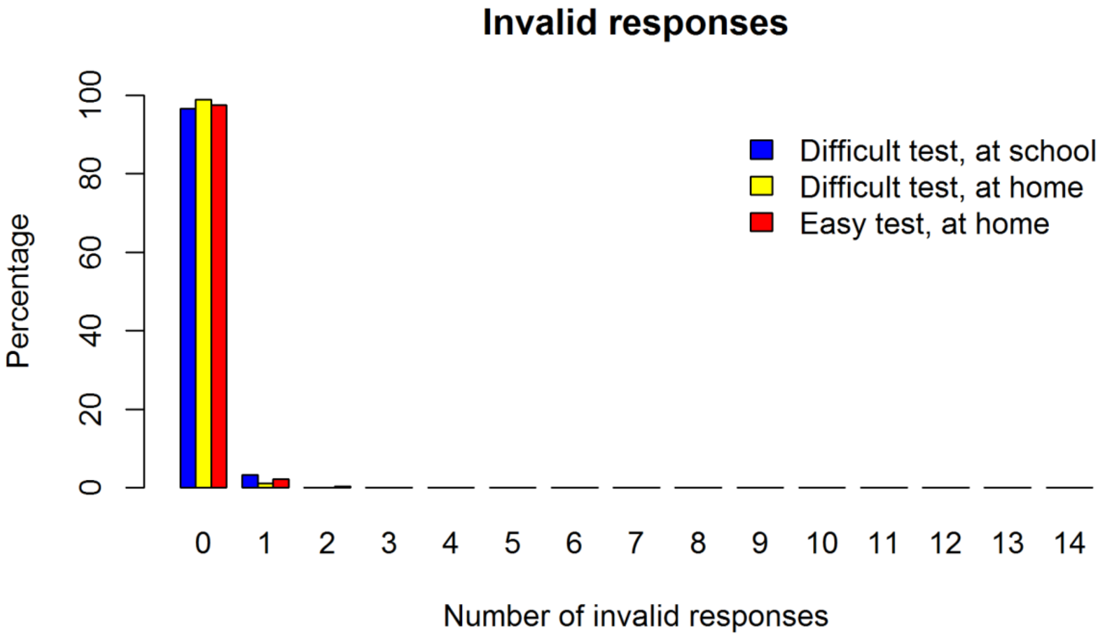

*Figure 1. Number of invalid responses by experimental condition*

Missing responses may also occur when respondents omit items. As illustrated in Figure 2 most respondents, 58% to 67%, did not skip any item and less than six percent omitted more than three items. There was no difference in the amount of omitted items between the different experimental conditions.

Another source of missing responses is items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was rather high especially for respondents that received the difficult test, because many respondents were unable to finish the test within the allocated time limit (Figure 3). Between 57% and 74% of the respondents finished the entire test. Of the participants administered the difficult test about 29% did not reach the last five items, whereas this was the case for about 19% of the participants administered the easy test.

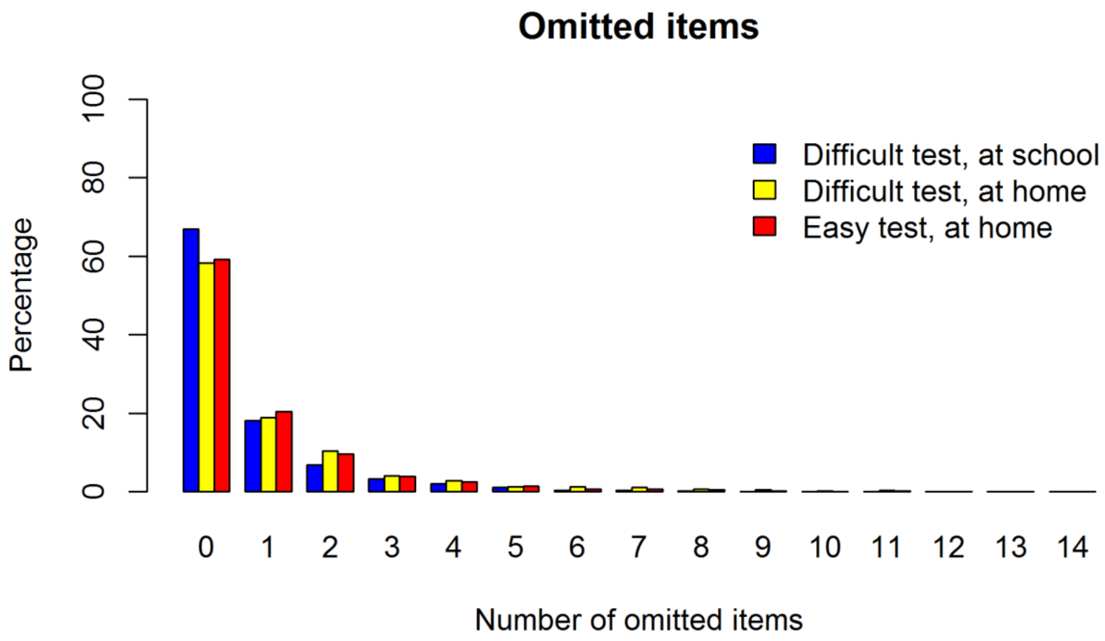

*Figure 2. Number of omitted items by experimental condition*

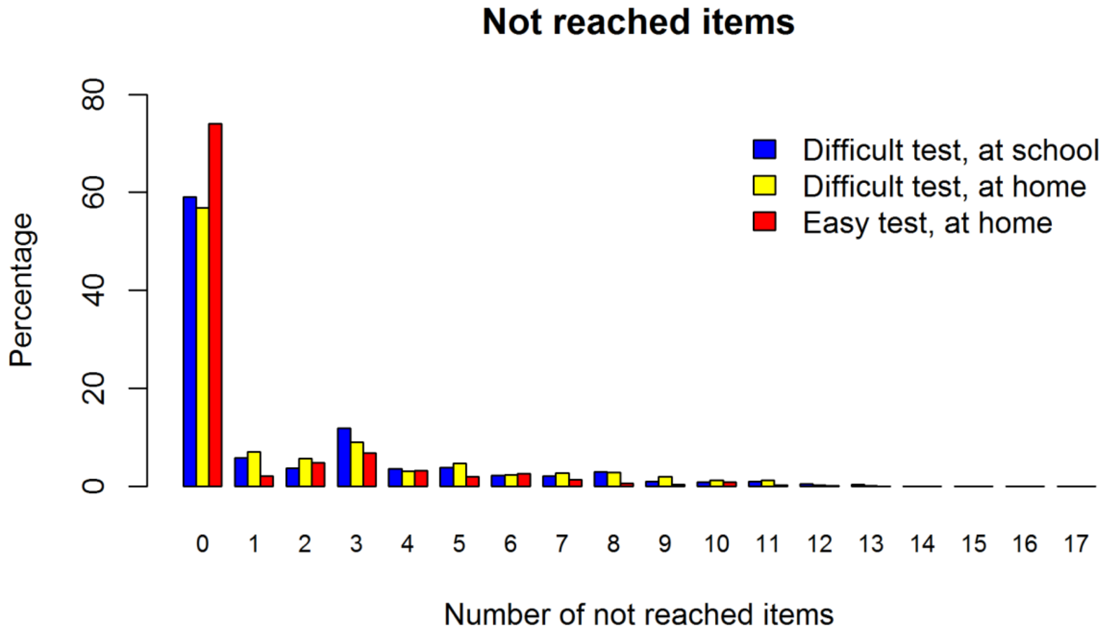

*Figure 3. Number of not-reached items by experimental condition*

The total number of missing responses, aggregated over invalid, omitted and not-reached, per person, is illustrated in Figure 4. On average, the respondents showed between *M* = 2.04 (*SD* = 2.93) and *M* = 2.96 (*SD* = 3.46) missing responses in the different experimental conditions. About 34% to 45% of the respondents had no missing response at all and about 23% to 32% of the participants had four or more missing responses.

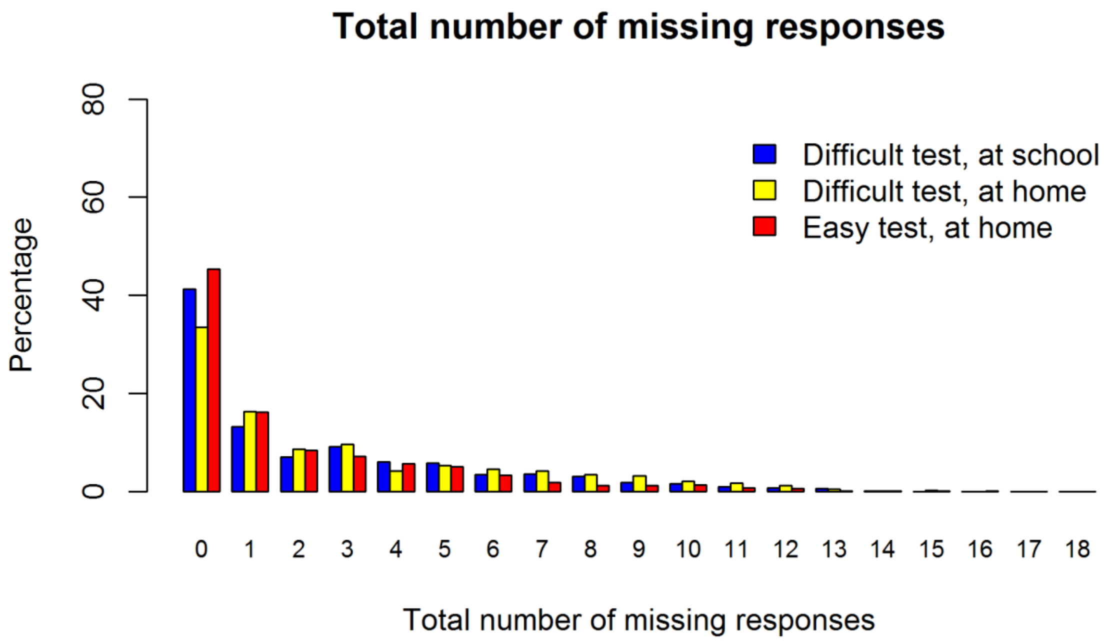

*Figure 4. Total number of missing responses by experimental condition*

In sum, the amount of invalid missing responses is small, whereas a reasonable part of missing responses occurs due to omitted items. The number of not-reached items is, however, rather large and has the greatest impact on the total number of missing responses.

### 5.1.2 Missing responses per item

Tables 4 and 5 provide information on the occurrence of different kinds of missing responses per item for the easy and difficult test version. Overall, in both tests the omission rates were rather low, varying across items between 0.00% and 10.00%. There was only one item with an omission rate exceeding 10% (mag9r061_sc4g12_c, when the item was administered in the home condition). For the difficult test omission rates correlated with the item difficulties at about .16 in the school context and about .03 at home; for the easy test the respective correlation was larger with .38 in the home setting. Generally, the percentage of invalid responses per item (columns 6 and 10 in Tables 4 and 5) was rather low with the maximum rate being 1.92%. With an item's progressing position in the test, the amount of persons that did not reach the item (columns 4 and 8 in Tables 4 and 5) rose up to a considerable amount of 26% to 43% for the different experimental conditions. Particularly, in the difficult condition the last items of the test were not reached by many respondents (see Figure 5).

Table 4

*Percentage of Missing Values for the Difficult Test by Assessment Setting*

| | | *At school* | | | | *At home* | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Item** | **Position** | ***N*** | **NR** | **OM** | **NV** | ***N*** | **NR** | **OM** | **NV** |
| maa3q071_sc4g12_c | 1 | 3836 | 0.00 | 1.64 | 0.00 | 519 | 0.00 | 3.89 | 0.00 |
| mag12v101_c | 2 | 3781 | 0.00 | 3.03 | 0.03 | 503 | 0.00 | 6.85 | 0.00 |
| mag12q121_c | 3 | 3845 | 0.00 | 1.33 | 0.08 | 522 | 0.00 | 3.15 | 0.19 |
| mag12v122_c | 4 | 3715 | 0.00 | 4.69 | 0.05 | 498 | 0.00 | 7.78 | 0.00 |
| mag12r011_c | 5 | 3807 | 0.00 | 2.31 | 0.08 | 521 | 0.00 | 3.52 | 0.00 |
| mag12v061_c | 6 | 3779 | 0.03 | 3.00 | 0.08 | 514 | 0.00 | 4.81 | 0.00 |
| mag12r091_c | 7 | 3643 | 0.08 | 6.49 | 0.03 | 501 | 0.00 | 7.22 | 0.00 |
| mag9r051_sc4g12_c | 8 | 3850 | 0.15 | 0.97 | 0.15 | 529 | 0.00 | 2.04 | 0.00 |
| mag12q081_c | 9 | 3666 | 0.56 | 5.44 | 0.00 | 505 | 0.19 | 6.30 | 0.00 |
| mag12d021_c | 10 | 3818 | 1.10 | 0.92 | 0.08 | 533 | 0.56 | 0.74 | 0.00 |
| mag12q051_c | 11 | 3677 | 2.18 | 3.51 | 0.03 | 502 | 1.85 | 5.19 | 0.00 |
| mag9d201_sc4g12_c | 12 | 3740 | 3.15 | 0.87 | 0.08 | 508 | 3.15 | 2.78 | 0.00 |
| mag9v121_sc4g12_c | 13 | 3708 | 4.15 | 0.74 | 0.03 | 502 | 5.19 | 1.85 | 0.00 |
| mas1q02s_sc4g12_c | 14 | 3334 | 7.21 | 7.08 | 0.13 | 445 | 8.15 | 9.44 | 0.00 |
| mas1d081_sc4g12_c | 15 | 3361 | 9.41 | 3.69 | 0.72 | 446 | 10.93 | 6.30 | 0.19 |
| maa3d112_sc4g12_c | 16 | 3166 | 11.69 | 7.08 | 0.05 | 414 | 13.33 | 10.00 | 0.00 |
| mag9r061_sc4g12_c | 17 | 2856 | 15.56 | 9.28 | 1.92 | 357 | 18.15 | 15.37 | 0.37 |
| maa3r011_sc4g12_c | 18 | 3090 | 19.26 | 1.51 | 0.00 | 412 | 21.30 | 2.41 | 0.00 |
| mag12r041_c | 20 | 2592 | 31.18 | 2.13 | 0.23 | 364 | 30.37 | 1.85 | 0.37 |
| mag12v131_c | 21 | 2475 | 34.97 | 1.54 | 0.03 | 338 | 36.11 | 1.30 | 0.00 |
| mag12d031_c | 22 | 2303 | 40.90 | 0.00 | 0.05 | 307 | 43.15 | 0.00 | 0.00 |

*Note*. Position = Item position within test, $N$ = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.
The item on position 12 was treated as a unique item in each testlet (see section 2).
Item 19 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

Table 5

*Percentage of Missing Values for the Easy Test*

| | | *At home* | | | |
|---|---|---|---|---|---|
| **Item** | **Position** | ***N*** | **NR** | **OM** | **NV** |
| maa3q071_sc4g12_c | 1 | 1230 | 0.00 | 4.80 | 0.08 |
| mag12v101_c | 2 | 1184 | 0.00 | 8.43 | 0.00 |
| mag12q121_c | 3 | 1257 | 0.00 | 2.78 | 0.00 |
| mag12v122_c | 4 | 1214 | 0.00 | 5.80 | 0.31 |
| maa3d131_sc4g12_c | 5 | 1245 | 0.08 | 3.56 | 0.08 |
| maa3d132_sc4g12_c | 6 | 1214 | 0.15 | 5.96 | 0.00 |
| mag12r091_c | 7 | 1184 | 0.15 | 8.28 | 0.00 |
| mag9r051_sc4g12_c | 8 | 1246 | 0.23 | 3.02 | 0.39 |
| mag9v011_sc4g12_c | 9 | 1255 | 0.31 | 2.47 | 0.15 |
| mag12d021_c | 10 | 1273 | 0.54 | 0.85 | 0.15 |
| mag12q051_c | 11 | 1211 | 0.85 | 5.41 | 0.08 |
| mag9d201_sc4g12_c | 12 | 1230 | 1.78 | 2.94 | 0.15 |
| mag9v121_sc4g12_c | 13 | 1237 | 2.17 | 2.09 | 0.08 |
| maa3r121_sc4g12_c | 14 | 1248 | 2.78 | 0.70 | 0.00 |
| mag12q111_c | 15 | 1225 | 4.18 | 1.08 | 0.00 |
| mag9r061_sc4g12_c | 16 | 893 | 6.81 | 23.2 | 0.93 |
| maa3q101_sc4g12_c | 17 | 1137 | 8.82 | 3.09 | 0.15 |
| mag9q101_sc4g12_c | 18 | 1093 | 12.06 | 3.33 | 0.08 |
| mag12r041_c | 20 | 1016 | 18.95 | 2.17 | 0.31 |
| mag12v131_c | 21 | 963 | 23.82 | 1.70 | 0.00 |
| mag12v132_c | 22 | 956 | 25.99 | 0.00 | 0.08 |

*Note*. Position = Item position within test, $N$ = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.

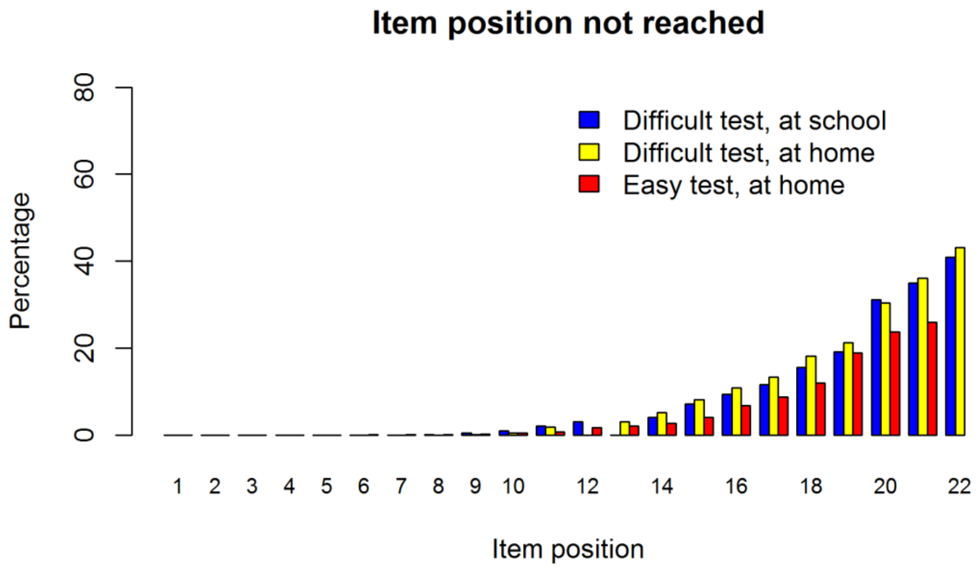

*Figure 5. Item position not reached by experimental conditions*

## 5.2 Parameter Estimates

### 5.2.1 Item parameters

The second column in Table 6 presents the percentage of correct responses in relation to all valid responses for each item. Because there is a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The percentage of correct responses within dichotomous items varied between 16% and 78% with an average of 49% (*SD* = 15%) correct responses.

Table 6

*Item Parameters*

| | Item | Percentage correct | Item difficulty | *SE* | WMNSQ | *t* | $r_{\text{it}}$ | Discr. | $Q_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1. | maa3q071_sc4g12_c | 60.56 | -0,510 | 0,030 | 1,01 | 0,9 | 0.34 | 0.96 | 0.03 |
| 2. | mag12v101_c | 57.94 | -0,375 | 0,030 | 0,97 | -2,5 | 0.39 | 1.18 | 0.03 |
| 3. | mag12q121_c | 34.92 | 0,747 | 0,031 | 1,04 | 3,2 | 0.29 | 0.80 | 0.03 |
| 4. | mag12v122_c | 48.39 | 0,070 | 0,030 | 1,08 | 7,7 | 0.26 | 0.65 | 0.04 |
| 5. | maa3d131_sc4g12_c | 40.48 | -0,099 | 0,062 | 1,00 | 0,1 | 0.31 | 1.02 | 0.03 |
| 6. | maa3d132_sc4g12_c | 16.14 | 1,328 | 0,082 | 0,94 | -1,2 | 0.34 | 1.34 | 0.03 |
| 7. | mag12r011_c | 46.95 | 0,299 | 0,033 | 0,96 | -3,5 | 0.4 | 1.19 | 0.03 |
| 8. | mag12v061_c | 35.24 | 0,894 | 0,035 | 1,01 | 0,4 | 0.33 | 0.96 | 0.03 |
| 9. | mag12r091_c | 39.34 | 0,512 | 0,031 | 1,08 | 6,7 | 0.25 | 0.64 | 0.04 |

Table 6 (continued)

| | Item | Percentage correct | Item difficulty | *SE* | WMNSQ | *t* | $r_{\mathrm{it}}$ | Discr. | $Q_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 10. | mag9r051_sc4g12_c[a] | 70.66 | -0,891 | 0,036 | 0,96 | -2,6 | 0.38 | 1.30 | 0.03 |
| | | 36.84 | 0,064 | 0,063 | 0,99 | -0,3 | 0.32 | 1.02 | 0.03 |
| 11. | mag9v011_sc4g12_c | 59.20 | -0,984 | 0,062 | 0,98 | -0,8 | 0.32 | 1.17 | 0.02 |
| 12. | mag12q081_c | 23.76 | 1,545 | 0,039 | 0,92 | -4,1 | 0.41 | 1.52 | 0.03 |
| 13. | mag12d021_c | 56.81 | -0,342 | 0,030 | 1,01 | 1,2 | 0.33 | 0.91 | 0.03 |
| 14. | mag12q051_c | 27.61 | 1,143 | 0,033 | 1,03 | 2,2 | 0.28 | 0.80 | 0.03 |
| 15. | mag9d201_sc4g12_c[b] | 78.48 | -1,295 | 0,043 | 0,94 | -2,5 | 0.38 | 1.53 | 0.04 |
| | | 43.38 | -0,217 | 0,052 | 0,98 | -0,9 | 0.35 | 1.16 | 0.03 |
| 16. | mag9v121_sc4g12_c | 47.84 | 0,090 | 0,030 | 0,93 | -6,7 | 0.45 | 1.41 | 0.02 |
| 17. | maa3r121_sc4g12_c | 66.75 | -1,377 | 0,064 | 1,03 | 1,0 | 0.26 | 0.92 | 0.03 |
| 18. | mag12q111_c | 43.35 | -0,270 | 0,062 | 1,01 | 0,6 | 0.30 | 0.95 | 0.02 |
| 19. | mas1q02s_sc4g12_c | n.a. | -1,247 | 0,036 | 0,99 | -0,4 | 0.38 | 1.19 | 0.03 |
| 20. | mas1d081_sc4g12_c | 71.29 | -0,934 | 0,039 | 1,02 | 1,0 | 0.31 | 0.91 | 0.03 |
| 21. | maa3d112_sc4g12_c | 30.98 | 1,109 | 0,039 | 1,04 | 2,5 | 0.27 | 0.74 | 0.03 |
| 22. | mag9r061_sc4g12_c | 51.73 | -0,033 | 0,034 | 0,93 | -5,4 | 0.44 | 1.34 | 0.03 |
| 23. | maa3q101_sc4g12_c | 33.69 | 0,188 | 0,067 | 1,09 | 3,2 | 0.18 | 0.52 | 0.04 |
| 24. | mag9q101_sc4g12_c | 65.97 | -1,354 | 0,068 | 0,94 | -2,2 | 0.39 | 1.68 | 0.04 |
| 25. | maa3r011_sc4g12_c | 59.28 | -0,331 | 0,038 | 0,90 | -7,1 | 0.48 | 1.65 | 0.04 |
| 27. | mag12r041_c | 58.59 | -0,514 | 0,035 | 1,05 | 3,9 | 0.30 | 0.76 | 0.03 |
| 28. | mag12v131_c | 50.90 | -0,151 | 0,036 | 1,10 | 7,7 | 0.25 | 0.60 | 0.03 |
| 29. | mag12v132_c | 64.54 | -1,337 | 0,072 | 0,97 | -1,0 | 0.32 | 1.22 | 0.03 |
| 30. | mag12d031_c | 58.62 | -0,340 | 0,044 | 0,95 | -2,8 | 0.42 | 1.26 | 0.03 |

*Note*. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, $r_{\mathrm{it}}$ = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, $Q_3$ =Average absolute residual correlation for item (Yen, 1983).

Item 26 was excluded from the analyses due to an unsatisfactory item fit (see section 2). Percent correct scores are not informative for polytomous CMC item scores. These are denoted by n.a. For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

[a] Item 10 was scaled separately in the difficult (first row) and easy test (second row) due to differential item functioning (DIF). [b] Item 15 was scaled separately for the group (first row) and the individual setting (second row) due to DIF.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are given in Table 6. The step parameters for the polytomous variable are depicted in Table 7. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for the polytomous variable) ranged from -0.88 (item maa3r121_sc4g12_c) to 2.04 (item mag12q081_c) with an average difficulty of 0.38. Due to the large sample size the standard errors (*SE*) of the estimated item difficulties (column 4 in Table 6) were rather small (all *SE*s ≤ 0.07).

Table 7

*Step Parameters (with Standard Errors) for the Polytomous Item*

| Item | Step 1 | Step 2 | Step 3 |
|---|---|---|---|
| Mas1q02s _sc4g12_c | -0,213<br>(0.034) | 0,033<br>(0.037) | *0,18* |

### 5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 6, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.989, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .766) was good. Although the items covered a wide range of the ability distribution, there were no items to cover the lower and upper peripheral ability areas. As a consequence, person ability in medium ability regions will be measured relative precisely, whereas lower and higher ability estimates will have larger standard errors of measurement.

| | | |
|---|---|---|
| 3 | | |
| | X | |
| | X | |
| | X | |
| 2 | X | |
| | XX | |
| | XX | |
| | XXX | |
| | XXX | 12 |
| | XXXX | 6 |
| | XXXXX | |
| | XXXX | 14 21 |
| 1 | XXXXX | |
| | XXXXXX | 8 |
| | XXXXXXX | 3 |
| | XXXXXXX | |
| | XXXXXXXX | 9 |
| | XXXXXXXXX | 7 |
| | XXXXXXXXXX | 23 |
| | XXXXXXXXX | 4 10b 16 |
| 0 | XXXXXXXXX | 5 22 |
| | XXXXXXXX | 15b 28 |
| | XXXXXXXXXX | 2 13 18 25 30 |
| | XXXXXXXX | 1 27 |
| | XXXXXXXXX | |
| | XXXXXXXX | |
| | XXXXXXX | 10a |
| -1 | XXXXX | 11 20 |
| | XXXXXX | |
| | XXXXX | 15a 19 |
| | XXXX | 17 24 29 |
| | XXX | |
| | XXX | |
| | XX | |
| | X | |
| -2 | X | |
| | X | |
| -3 | | |

*Figure 6.* Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph, with each ‘X’ representing 32.4 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 6).

## 5.3 Quality of the test

### 5.3.1 Fit of the subtasks of the complex multiple choice item

Before the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of the CMC item separately, there were 34 items. The probability of a correct response range from 16% to 90% across all items (*Mdn* = 54%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.90 to 1.10, the respective *t*-value from -7.2 to 7.9, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to a polytomous variable seems to be justified.

### 5.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC items and the polytomous CMC item. Altogether, item fit can be considered to be very good (see Table 6). Values of the WMNSQ ranged from 0.9 (item maa3r011_sc4g12_c) to 1.10 (mag12v131_c). Only three items exhibited a *t*-value of the WMNSQ greater than 6 and none exceeded a value of 8. Thus, there is no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .18 (item maa3q101_sc4g12_c) to .48 (item maa3r011_sc4g12_c) and had a mean of .34. All item characteristic curves showed a good fit of the items.

### 5.3.3 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total score. The point-biserial correlations for the distractors ranged from -.45 to .07 with a mean of -.14. These results indicate that the distractors worked well.

### 5.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status), migration background, school type, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, the effect of the two experimental factors was also studied. Thus, we compared the two assessment settings (at school or at home) and for the common items that were administered to all participants we examined measurement invariance for the easy and difficult test. The differences between the estimated item difficulties in the various groups are summarized in Table 8. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 9).

Table 8

*Differential Item Functioning*

| Item | Gender | Books | Migration | School | Position | Setting | Booklet |
|---|---|---|---|---|---|---|---|
| | male vs.<br>female | < 100 vs.<br>≥ 100 | without<br>vs. with | no sec.<br>vs. sec. | first vs.<br>second | school<br>vs. home | easy vs.<br>difficult |
| maa3q071_sc4g12_c | 0.262<br>(0.275) | 0.032<br>(0.034) | 0.040<br>(0.041) | -0.036<br>(-0.041) | 0.566<br>(0.756) | -0.106<br>(-0.122) | 0.070<br>(0.089) |
| mag12v101_c | -0.206<br>(-0.216) | 0.040<br>(0.042) | 0.032<br>(0.033) | 0.034<br>(0.039) | -0.116<br>(-0.155) | -0.080<br>(-0.092) | 0.130<br>(0.165) |
| mag12q121_c | 0.450*<br>(0.472) | 0.120<br>(0.126) | 0.004<br>(0.004) | -0.036<br>(-0.041) | -0.134<br>(-0.179) | -0.092<br>(-0.105) | 0.178<br>(0.226) |
| mag12v122_c | 0.304<br>(0.319) | -0.170<br>(-0.178) | 0.106<br>(0.108) | -0.398*<br>(-0.456) | -0.058<br>(-0.077) | 0.264<br>(0.303) | -0.176<br>(-0.223) |
| maa3d131_sc4g12_c | 0.176<br>(0.184) | 0.152<br>(0.159) | -0.036<br>(-0.037) | 0.430<br>(0.493) | -0.178<br>(-0.238) | | |
| maa3d132_sc4g12_c | -0.434<br>(-0.455) | -0.088<br>(-0.092) | 0.080<br>(0.081) | 0.108<br>(0.124) | 0.138<br>(0.184) | | |
| mag12r011_c | -0.178<br>(-0.187) | -0.010<br>(-0.01) | -0.072<br>(-0.073) | -0.022<br>(-0.025) | 0.160<br>(0.214) | -0.236<br>(-0.271) | |
| mag12v061_c | -0.300<br>(-0.314) | -0.244<br>(-0.256) | 0.124<br>(0.126) | -0.398*<br>(-0.456) | 0.008<br>(0.011) | 0.554*<br>(0.635) | |
| mag12r091_c | 0.238<br>(0.249) | -0.208<br>(-0.218) | 0.146<br>(0.149) | -0.410*<br>(-0.470) | -0.176<br>(-0.235) | 0.376*<br>(0.431) | -0.344<br>(-0.437) |
| mag9r051_sc4g12_c[a] | 0.082<br>(0.086) | 0.142<br>(0.149) | -0.114<br>(-0.116) | -0.040<br>(-0.046) | -0.174<br>(-0.232) | -0.004<br>(-0.005) | |
| | 0.076<br>(0.080) | 0.500<br>(0.524) | -0.116<br>(-0.118) | 0.242<br>(0.277) | 0.070<br>(0.093) | | |
| mag9v011_sc4g12_c | -0.266<br>(-0.279) | 0.034<br>(0.036) | -0.066<br>(-0.067) | -0.138<br>(-0.158) | -0.172<br>(-0.230) | | |
| mag12q081_c | -0.120<br>(-0.126) | 0.042<br>(0.044) | 0.032<br>(0.033) | 0.282<br>(0.323) | 0.162<br>(0.216) | -0.034<br>(-0.039) | |
| mag12d021_c | -0.226<br>(-0.237) | 0.000<br>(0.000) | 0.038<br>(0.039) | -0.026<br>(-0.030) | -0.068<br>(-0.091) | -0.012<br>(-0.014) | 0.082<br>(0.104) |
| mag12q051_c | 0.178<br>(0.187) | -0.182<br>(-0.191) | 0.184<br>(0.187) | -0.162<br>(-0.186) | 0.038<br>(0.051) | 0.100<br>(0.115) | -0.038<br>(-0.048) |
| mag9d201_sc4g12_c[b] | -0.498*<br>(-0.522) | 0.132<br>(0.138) | -0.060<br>(-0.061) | 0.404<br>(0.463) | | | |

| Item | Gender | Books | Migration | School | Position | Setting | Booklet |
|---|---|---|---|---|---|---|---|
| | -0.084 | -0.096 | -0.088 | 0.596* | 0.062 | | -0.118 |
| | (-0.088) | (-0.101) | (-0.090) | (0.683) | (0.083) | | (-0.150) |
| mag9v121_sc4g12_c | -0.056 | 0.210 | -0.140 | 0.184 | 0.140 | -0.254 | 0.288 |
| | (-0.059) | (0.220) | (-0.143) | (0.211) | (0.187) | (-0.291) | (0.365) |
| maa3r121_sc4g12_c | -0.742* | -0.050 | 0.008 | -0.370 | -0.404 | | |
| | (-0.778) | (-0.052) | (0.008) | (-0.424) | (-0.539) | | |
| mag12q111_c | 0.248 | -0.282 | 0.042 | -0.182 | 0.032 | | |
| | (0.260) | (-0.295) | (0.043) | (-0.208) | (0.043) | | |
| mas1q02s_sc4g12_c | -0.072 | 0.164 | -0.012 | -0.012 | -0.098 | 0.034 | |
| | (-0.075) | (0.172) | (-0.012) | (-0.014) | (-0.131) | (0.039) | |
| mas1d081_sc4g12_c | 0.260 | -0.064 | 0.000 | 0.194 | 0.242 | -0.282 | |
| | (0.273) | (-0.067) | (0.000) | (0.222) | (0.323) | (-0.323) | |
| maa3d112_sc4g12_c | 0.400 | -0.120 | 0.148 | -0.144 | 0.752 | 0.146 | |
| | (0.419) | (-0126) | (0.151) | (-0.165) | (1.004) | (0.167) | |
| mag9r061_sc4g12_c | 0.196 | 0.092 | -0.014 | 0.302 | -0.022 | -0.276 | 0.396* |
| | (0.205) | (0.096) | (-0.014) | (0.346) | (-0.029) | (-0.316) | (0.503) |
| maa3q101_sc4g12_c | 0.252 | -0.298 | 0.106 | -0.044 | 0.074 | | |
| | (0.264) | (-0.312) | (0.108) | (-0.050) | (0.099) | | |
| mag9q101_sc4g12_c | -0.114 | -0.044 | -0.136 | -0.268 | 0.140 | | |
| | (-0.120) | (-0.046) | (-0.139) | (-0.307) | (0.187) | | |
| maa3r011_sc4g12_c | -0.194 | 0.042 | -0.018 | 0.060 | -0.396 | -0.172 | |
| | (-0.203) | (0.044) | (-0.018) | (0.069) | (-0.529) | (-0.197) | |
| mag12r041_c | 0.288 | -0.022 | -0.048 | -0.138 | -0.370 | 0.122 | -0.088 |
| | (0.302) | (-0.023) | (-0.049) | (-0.158) | (-0.494) | (0.140) | (-0.112) |
| mag12v131_c | 0.292 | -0.270 | 0.198 | -0.472* | -0.082 | 0.308 | -0.380 |
| | (0.306) | (-0.283) | (0.202) | (-0.541) | (-0.109) | (0.353) | (-0.482) |
| mag12v132_c | -0.214 | 0.240 | -0.260 | 0.114 | -0.026 | | |
| | (-0.224) | (0.251) | (-0.265) | (0.131) | (-0.035) | | |
| mag12d031_c | 0.002 | 0.206 | -0.112 | 0.350 | -0.114 | -0.350 | |
| | (0.002) | (0.216) | (-0.114) | (0.401) | (-0.152) | (-0.401) | |

| Item | Gender | Books | Migration | School | Position | Setting | Booklet |
|---|---|---|---|---|---|---|---|
| main effect | 0.596 | -0.682 | 0.424 | -1.144 | -0.074 | 1.076 | -1.040 |
| (model with DIF) | (0.625) | (-0.715) | (0.432) | (-1.311) | (-0.099) | (1.233) | (-1.320) |
| main effect | 0.548 | -0.674 | 0.410 | -1.084 | -0.060 | 1.060 | -1.046 |
| (model without DIF) | (0.576) | (-0.708) | (0.418) | (-1.249) | (-0.080) | (1.219) | (-1.330) |

*Note*. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium").

[a] DIF for Item mag9r051_sc4g12_c was calculated separately in the difficult (first row) and easy (second row) test. [b] DIF for Item mag9d201_sc4g12_c was calculated separately in the group setting (first row) and individual setting (second row).
* Absolute standardized difference is significantly, $p < .05$, greater than 0.25 (see Fischer et al., 2016).

Gender: The sample included 2,690 (47%) males and 3,028 (53%) females. Fifteen respondents that did not indicate their gender were excluded from the analysis. On average, male participants had a higher estimated mathematics ability than females (main effect = 0.596 logits, Cohen's *d* = 0.625). Only one item (item maa3r121_sc4g12_c) showed DIF greater than 0.6 logits. An overall test for DIF (see Table 9) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF; see table 8). Model comparisons using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF. The deviation was small in both cases. Thus, overall, there was no pronounced DIF with regard to gender.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,632 (29%) test takers with 0 to 100 books at home and 3,819 (67%) test takers with more than 100 books at home. 282 (5%) test takers had no valid response and were excluded from the analysis. There was considerable average difference between the two groups. Participants with 100 or less books at home performed on average -0.682 logits (Cohen's *d* = -0.715) lower in mathematics than participants with more than 100 books. However, there was no considerable DIF on the item level. Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 9).

Migration background: There were 4,226 participants (74%) with no migration background, 1,382 subjects (24%) with a migration background and 125 individuals (2%) that did not indicate their migration background. In comparison to subjects with migration background, participants without migration background had, on average, a slightly higher mathematics ability (main effect = 0.424 logits, Cohen's *d* = 0.432). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.6 logits. Moreover, the overall test for DIF using the AIC and BIC also favored the main effects model that did not include item-level DIF.

School type: Overall, 3,638 subjects (64%) who took the mathematics test attended secondary school (German: "Gymnasium") whereas 2,095 (37%) were enrolled in other school types. Subjects in secondary schools showed a higher mathematics ability on average (1.144 logits; Cohen's *d* = -1.311) than subjects in other school types. There was no noteworthy item DIF; no item exhibited DIF greater than 0.6 logits. However, the overall model test using AIC indicated a slightly better fit for the more complex DIF model, because several items showed

DIF effects between 0.4 and 0.6 Logits, whereas the test using BIC favored the main effects model; thus, these differences were not considered severe.

Table 9

*Comparisons of Models with and without DIF*

| **DIF variable** | **Model** | ***N*** | **Deviance** | **Number of parameters** | **AIC** | **BIC** |
|---|---|---|---|---|---|---|
| Gender | main effect | 5,718 | 132,412.404 | 35 | 132,482.404 | 132,715.202 |
| | DIF | 5,718 | 132,080.367 | 65 | 132,210.367 | 132,642.706 |
| Books | main effect | 5,451 | 126,058.588 | 35 | 126,128.588 | 126,359.712 |
| | DIF | 5,451 | 125,961.913 | 65 | 126,091.913 | 126,521.144 |
| Migration | main effect | 5,608 | 130,037.147 | 35 | 130,107.147 | 130,339.265 |
| | DIF | 5,608 | 130,002.084 | 65 | 130,132.084 | 130,563.160 |
| School | main effect | 5,733 | 131,827.666 | 35 | 131,897.666 | 132,130.555 |
| | DIF | 5,733 | 131,619.259 | 65 | 131,749.259 | 132,181.769 |
| Position | main effect | 1,649 | 37,217.470 | 34 | 37,285.470 | 37,469.340 |
| | DIF | 1,649 | 37,165.040 | 63 | 37,291.040 | 37,631.740 |
| Setting | main effect | 5,733 | 11,3879.072 | 24 | 113,927.072 | 114,086.768 |
| | DIF | 5,733 | 11,3732.891 | 43 | 113,818.891 | 114,105.013 |
| Booklet | main effect | 5,733 | 73,001.9473 | 14 | 73,029.947 | 73,123.103 |
| | DIF | 5,733 | 72,905.9139 | 25 | 72,955.914 | 73,122.264 |

Position: The mathematics competence test was administered in two different positions (see section 3.1 for the design of the study). A subsample of 1,264 (77%) persons received the mathematics test first and 385 (23%) respondents took the mathematics test after having completed either the computer literacy or the reading test at home. Participants that were administered the test in school were excluded from the analysis, because they had no variation in test position. Additionally, in order to prevent confounding test position and proficiency, home-tested participants that were enrolled in secondary school were excluded from the analyses, because of their unequal distribution between the test positions. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show a minor average effect of item position[1]. Subjects who received the mathematics test first performed on average 0.074 logits (Cohen's *d* = -0.099) worse than subjects who received the mathematics test second. One item

[1] Note that this main effect does not indicate a threat to measurement invariance. Instead, it may be an indication of fatigue effects that are similar for all items.

(maa3d112_sc4g12_c) exhibited DIF greater than 0.6 Logits. However, the overall test for DIF using the AIC and BIC favored the more parsimonious main effect model.

Setting: The mathematics competence test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 3,900 (68%) persons received the mathematics test in small groups at school, whereas 1,833 (32%) participants finished the test individually at their private homes. Subjects who finished the mathematics test at school were on average 1.076 logits (Cohen's *d* = 1.233) better than those working at their private homes. However, this difference must not be interpreted as a causal effect of the administration setting because respondents were not randomly assigned to the different settings. Rather, it is likely that self-selection processes occurred, for example, because less proficient students were more likely to leave school and, consequently, were tested at home. More importantly, there was no noteworthy DIF due to the administration setting; all differences in item difficulties were smaller than 0.6 logits. Again, the overall model test using AIC (see Table 9) indicated a slightly better fit for the more complex DIF model, whereas the model test using BIC favored the more parsimonious main effect model. The largest difference in difficulty between the two design groups was 0.554 logits (item mag12v061_c) and was not considered severe.

Booklet: To estimate the participants' proficiency with great accuracy the participants received different tests that either included a larger number of easy or a larger number of difficult items (see section 3.1 for the design of the study). Only a subset of 12 items that were included in both tests was administered to all participants. For these common items we examined potential DIF across the two test versions (easy versus difficult). A subsample of 1,293 (23%) persons received the easy test and 4,440 (77%) persons received the difficult test. As expected, subjects who were administered the easy test scored on average 1.04 logits (Cohen's *d* = -1.320) lower than subjects who received the difficult test. There was no DIF for the common items with regard to the test version. The largest difference in difficulties between the two groups was 0.396 logits (item mag9r061_sc4g12_c).

### 5.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (2PL) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 6), ranging from 0.52 (item maa3q101_sc4g12_c) to 1.68 (item mag9q101_sc4g12_c). The average discrimination parameter fell at 1.07. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 132,404.64, BIC = 132,958.50) as compared to the 1PL model (AIC = 133,153.61, BIC = 133,379.85). Despite the empirical preference for the 2PL model, the 1PL model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

### 5.3.6 Unidimensionality

The unidimensionality of the test was investigated by specifying a multidimensional model and comparing it to a unidimensional model. The four different content areas constituted the multidimensional model. Estimation of the model was carried out in ConQuest using Gauss-Hermite quadrature method.

Table 10

*Results of Four-Dimensional Scaling*

| | **Dim 1** | **Dim 2** | **Dim 3** | **Dim 4** |
|---|---|---|---|---|
| **Quantity**<br>(Dim 1)<br>(8 items) | (1.014) | | | |
| **Space and Shape**<br>(Dim 2)<br>(8 items) | 0.929 | (1.175) | | |
| **Change and Relationships**<br>(Dim 3)<br>(7 items) | 0.957 | 0.953 | (0.947) | |
| **Data and Chance**<br>(Dim 4)<br>(9 items) | 0.927 | 0.912 | 0.922 | (1.174) |

*Note*. Variances of the dimensions are given in the diagonal and correlations are given in the off-diagonal.

The estimated variances and correlations of the four-dimensional model based on the four content areas given in Table 10. The correlations between the dimensions varied between *r* = .912 and *r* = .957. The smallest correlation was found between Dimension 2 (“Space and Shape”) and Dimension 4 (“Data and Chance”). Dimension 1 (“Quantity”) and Dimension 3 (“Change and Relationships”) showed the strongest correlation. All correlations but the correlations between Dimensions 1 and 3 and Dimensions 2 and 3 deviated from a perfect correlation (i.e., they were considerably lower than *r* = .95, see Carstensen, 2013). Nonetheless, the four-dimensional model (AIC = 133153.96, BIC = 133440.08, number of parameters = 43) fitted the data slightly worse than the unidimensional model (AIC = 133,153.61, BIC = 133,379.85, number of parameters = 34). As each item corresponded to one of the four content areas, local item dependence (LID) and the content areas were confounded. As a consequence, the deviation of the correlations from a perfect correlation shown in Table 10, may result from multidimensionality as well as from local item dependence. Given the testing design in the main studies, it is not possible to disentangle the two sources. However, for the unidimensional model the average absolute residual correlations as indicated by the corrected $Q_3$ statistic (see Table 6) were quite low (*M* = -.02, *SD* = .03) — the largest individual residual correlation was .11 — and thus indicated an essentially unidimensional test. Because the mathematics test is constructed to measure a single dimension, a unidimensional mathematics competence score was estimated.

## 6. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the mathematics test in starting cohort 4 for grade 12 and at describing how the mathematics competence score was estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, items with short constructed responses, subtasks of the CMC item, as well as the aggregated polytomous CMC item, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests dimensionality as well as local item dependence.

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the amount of not-reached items was higher in the difficult test than in the easy test, indicating that the difficult test was more time challenging. Other types of missing responses were reasonably small.

The test had a high reliability and distinguished well between test takers. However, the test is mainly targeted at medium-performing students and did less accurately measure mathematics competence of high- and low-performing students. As a consequence, ability estimates will be precise for medium-performing students but less precise for high- and low-performing students.

Summarizing these results, the test has good psychometric properties that facilitate the estimation of a unidimensional mathematical competence score.

## 7. Data in the Scientific Use File

### 7.1 Naming conventions

The data in the scientific use file contain 30 items, of which 29 items were scored as dichotomous variables (27 MC and 2 SCR items) with 0 indicating an incorrect response and 1 indicating a correct response. One item was scored as a polytomous variable (CMC item). MC and SCR items are marked with a ‘0_c’ at the end of the variable name, whereas the variable names of CMC items end in ‘s_c’. For further details on the naming conventions of the variables see Fuß and colleagues (2016). In the IRT scaling model, the polytomous CMC variable was scored as 0.5 for each category. One item was excluded from the estimation of the competence scores due to an unsatisfactory item fit (see section 2).

### 7.2 Linking of competence scores

In starting cohort 4, the mathematics competence tests administered in grades 9 (see Duchhardt & Gerdes, 2013) and 12 include different items that were constructed in such a way as to allow for an accurate measurement of mathematical competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Although the tests in grades 9 and 12 share six common items, linking was not based on an anchor-items design (see Fischer et al., 2016) because two items showed DIF between the two booklets (mag9r051_sc4g12_c) and the administration settings (mag9d201_sc4g12_c). Moreover, two items (mag9v011_sc4g12_c, mag9v121_sc4g12_c) violated the assumption of measurement invariance. Following an anchor-group design, an independent link sample including students from grade 11 (all attending a vocational school) that were not part of starting cohort 4 were administered all items from the grade 9 and the easy test version of the grade 12 mathematics competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 4 across the two grades.

### 7.2.1 Samples

In starting cohort 4, a subsample of 5,540 students participated at both measurement occasions, in grade 9 and also in grade 12. From these, 1,229 participants received the easy test in Grade 12. Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016.). Moreover, an independent link sample of *N* = 619 students (348 women) from grade 11 received both tests within a single measurement occasion.

### 7.2.2 The design of the link study

The test administered in grade 9 included 22 items (see Duchhardt & Gerdes, 2013), whereas the easy test version administered in grade 12 included 21 items (see above). Moreover, the mathematics test was administered at different positions in the test battery. A random sample of 318 students received the mathematics test before working on a reading test, whereas the remaining 301 students received the reading test before the mathematics test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the reading items in the same order.

### 7.2.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The Bayesian information criterion slightly favored the two-dimensional model, BIC = 20,811.31, over the one-dimensional model, BIC = 20,840.76. Also, Akaike's information criterion suggested a better fit for the two-dimensional model, AIC = 20,643.05, as compared to the one-dimensional model, AIC = 20,681.35. However, an examination of the residual correlations for the one-dimensional model using the corrected $Q_3$ statistic (Yen, 1984) indicated a largely unidimensional scale—the average absolute residual correlation was $M = .00$ ($SD = .05$, $Max = .15$). This indicates that the mathematics competence tests administered in grades 9 and 12 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 4 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 11.

Table 11

*Differential Item Functioning Analyses between the Starting Cohort and the Link Sample.*

| | Grade 9 | | | | | Grade 12 | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Nr.** | **Item** | **Δσ** | ***SE*$_{\Delta\sigma}$** | ***F*** | **Nr.** | **Item** | **Δσ** | ***SE*$_{\Delta\sigma}$** | ***F*** |
| 1. | mag9q071_c | 0.18 | 0.11 | 2.71 | 1. | maa3q071_sc4g12_c | -0.49 | 0.11 | 19.63 |
| 2. | mag9v131_c | -0.66 | 0.11 | 34.68 | 2. | mag12v101_c | -0.01 | 0.11 | 0.01 |
| 3. | mag9v13s_c | 0.10 | 0.11 | 0.78 | 3. | mag12q121_c | -0.10 | 0.13 | 0.61 |
| 4. | mag9r261_c | -0.21 | 0.19 | 1.15 | 4. | mag12v122_c | 0.10 | 0.11 | 0.72 |
| 5. | mag9r111_c | -0.13 | 0.12 | 1.20 | 5. | maa3d131_sc4g12_c | -0.43 | 0.11 | 14.80 |
| 6. | mag9d171_c | -0.04 | 0.11 | 0.11 | 6. | maa3d132_sc4g12_c | -0.18 | 0.13 | 1.72 |
| 7. | mag9d151_c | 0.04 | 0.12 | 0.12 | 9. | mag12r091_c | 0.14 | 0.12 | 1.29 |
| 10. | mag9v012_c | -0.24 | 0.15 | 2.73 | 13. | mag12d021_c | 0.07 | 0.11 | 0.40 |
| 11. | mag9q161_c | 0.29 | 0.12 | 6.04 | 14. | mag12q051_c | 0.25 | 0.15 | 2.75 |
| 13. | mag9r191_c | 0.28 | 0.11 | 6.22 | 17. | maa3r121_sc4g12_c | 0.11 | 0.12 | 0.93 |
| 15. | mag9q181_c | -0.18 | 0.16 | 1.17 | 18. | mag12q111_c | 0.01 | 0.11 | 0.01 |
| 16. | mag9r25s_c | 0.00 | 0.12 | 0.00 | 23. | maa3q101_sc4g12_c | 0.15 | 0.12 | 1.59 |
| 18. | mag9q081_c | -0.33 | 0.11 | 8.61 | 27. | mag12r041_c | 0.04 | 0.13 | 0.10 |
| 20. | mag9q021_c | -0.13 | 0.12 | 1.21 | 28. | mag12v131_c | 0.13 | 0.13 | 0.95 |
| 21. | mag9v091_c | 0.17 | 0.12 | 2.08 | 29. | mag12v132_c | 0.20 | 0.14 | 2.08 |
| 22. | mag9q211_c | 0.85 | 0.12 | 46.90 | | | | | |

*Note*. Common items (i.e., items repeatedly administered to the starting cohort in grades 9 as well as 12) were excluded from the analysis. Δσ = Difference in item difficulty parameters between the longitudinal subsample in grades 9 or 12 and the link sample (positive values indicate easier items in the link sample); *SE*$_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is $F_{0154}$ (1, 1,846) = 49.42. A non-significant test indicates measurement invariance.

Analyses of differential item functioning between the link sample and starting cohort 4 identified neither for grade 9 (difference in logits: *Min* = 0.00, *Max* = 0.85) nor for grade 12 (difference in logits: *Min* = 0.01, *Max* = |-0.49|) items with pronounced DIF. Therefore, the mathematical competence tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 0.496$. This correction term was subsequently added to each difficulty parameter derived in grade 12 (see Table 6) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) resulting in $0.142$ and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

## 7.3 Mathematical competence scores

In the SUF manifest mathematics competence scores are provided in the form of two different WLEs, "mag12_sc1" and "mag12_sc1u", including their respective standard error, "mag12_sc2" and "mag12_sc2u". For "mag12_sc1u", person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for differences in the test position. In grade 9, the mathematics test was always presented second within the test battery, whereas in grade 12 the mathematics test was either presented as the first or the second test within the test battery (see page 5). To correct for differences in the test position, we added the main effect related to the test position (see Table 8) to the WLE scores of respondents that received the mathematics test before working on another test. As a result, the WLE scores provided in "mag12_sc1u" can be used for longitudinal comparisons between grades 9 and 12. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in "mag12_sc1" are not linked to the underlying reference scale of grade 9. However, they are corrected for the position of the mathematics test within the booklet. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the mathematics test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

Plausible values that allow for an investigation of latent relationships of competence scores with other variables will be provided in future data releases. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012).

## References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.). *Research Outcomes of the PISA Research Conference 2009* (pp. 199-214). New York, NY: Springer.

Duchhardt, C. & Gerdes, A. (2013): *NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 4 in ninth grade* (NEPS Working Paper No. 22). Bamberg: University of Bamberg, National Educational Panel Study.

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A., & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne – Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze & M. Grüßing (eds.). Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht (313-327). Münster: Waxmann.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2012). Modeling and assessing of mathematical competence over the lifespan. Manuscript submitted for publication.

Organisation for Economic Cooperation and Development (OECD). (2014). PISA 2012 Technical Report. Paris, France: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. doi:10.1111/jedm.12028

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. doi:10.1007/s11618-011-0182-7

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

# Appendix

## Appendix A: ConQuest-Syntax for estimating linked WLEs in starting cohort 4

```
Title SC4 G12 MATH: Partial Credit Model;

/* load data */
datafile [FILENAME].sav ! filetype=spss,
         responses = maa3q071_sc4g12_c mag12v101_c mag12q121_c
                     mag12v122_c maa3d131_sc4g12_c maa3d132_sc4g12_c
                     mag12r011_c mag12v061_c mag12r091_c
                     mag9r051_sc4g12_d_c mag9r051_sc4g12_e_c
                     mag9v011_sc4g12_c mag12q081_c mag12d021_c
                     mag12q051_c
                     mag9d201_sc4g12_g_c mag9d201_sc4g12_i_c
                     mag9v121_sc4g12_c maa3r121_sc4g12_c mag12q111_c
                     mas1q02s_sc4g12_c mas1d081_sc4g12_c maa3d112_sc4g12_c
                     mag9r061_sc4g12_c maa3q101_sc4g12_c mag9q101_sc4g12_c
                     maa3r011_sc4g12_c mag12r041_c mag12v131_c
                     mag12v132_c mag12d031_c,
         pid=ID_t >> daten.dat;


/* collapse response categories with less than 200 responses */
recode (0,1,2,3,4)     (0,0,1,2,3)     ! item (21);  /* mas2q02s_c */

/* scoring */
codes 0,1,2,3,4;
score (0,1)         (0,1)              ! items (1-20, 22-31);
score (0,1,2,3)     (0,0.5,1,1.5)      ! item (21);


/* load linked item parameters */
import anchor_parameters << anchor_parameters.txt;

/* model specification */
set constraint = none;
model item + item*step;

/* estimate model */
estimate ! method=gauss, nodes=15, iterations=1000, convergence=0.0001,
stderr=empirical;

/* save results to file */
show ! estimate=latent    >> show.txt;
show cases ! estimate=wle >> wle.txt;
```