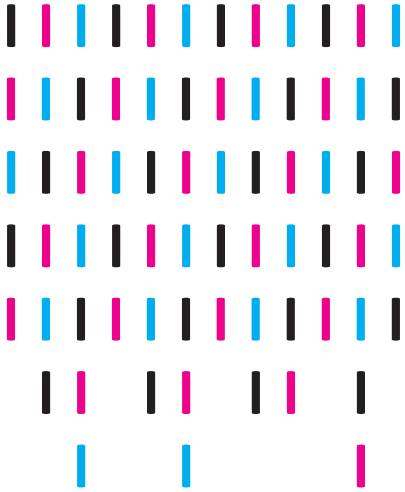

NEPS SURVEY PAPERS

Insa Schnittjer and Anna-Lena Gerken

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 3 IN GRADE 7

NEPS Survey Paper No. 16
Bamberg, January 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

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 3 in Grade 7

Insa Schnittjer and Anna-Lena Gerken

Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the lead author:

schnittjer@ipn.uni-kiel.de

Bibliographic Data:

Schnittjer, I. & Gerken, A.-L. (2017): *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 in Grade* 7 (NEPS Survey Paper No. 16). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP16:1.0

Acknowledgements:

We would like to thank Steffi Pohl and Kerstin Haberkorn for developing and providing standards for the technical reports and Timo Gnambs and Luise Fischer for giving valuable feedback on previous drafts of this manuscript.

The present report has been modeled along previous reports published by the NEPS. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Duchhardt, 2015; Haberkorn, Pohl, Hardt, & Wiegand, 2012; Jordan & Duchhardt, 2013; Koller, Haberkorn, & Rohm, 2014; Pohl, Haberkorn, Hardt, & Wiegand, 2012).

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 in Grade 7

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and designs tests for assessing these different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedure for the mathematical competence test in grade 7 of starting cohort 3 (fifth grade). The mathematics test contained 23 items with different response formats representing different content areas and cognitive components. The test was administered to 6,194 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test´s dimensionality were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability good item fit and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the large number of items targeted toward a lower mathematical ability as well as the relatively high omission rates in some items. Overall, the mathematics test had acceptable psychometric properties that allowed for an estimation of reliable mathematics competence scores. Besides the scaling results, this paper also describes the data available in the Scientific Use File and provides 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
3.3 Missing Responses 6
3.4 Scaling Model 6
3.5 Checking the Quality of the Scale 7
3.6 Software 8
4. Results 8
4.1 Missing Responses 8
4.1.1 Missing responses per person 8
4.1.2 Missing responses per item 10
4.2 Parameter Estimates 12
4.2.1 Item parameters 12
4.2.2 Test targeting and reliability 14
4.3 Quality of the test 16
4.3.1 Distractor analyses 16
4.3.2 Item fit 16
4.3.3 Differential item functioning 16
4.3.4 Rasch-homogeneity 19
4.3.5 Unidimensionality 19
5. Discussion 20
6. Data in the Scientific Use File 21
6.1 Naming conventions 21
6.2 Linking the data of Grade 5 and Grade 7 21
6.3 Mathematical competence scores 22
References 23
Appendix 25

1. Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured 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 competence domains measured in the NEPS is given by Weinert et al. (2011).

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 grade 7 of starting cohort 3 (fifith grade). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of starting cohort 3 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 of this report are based on the data available some time before data release. Due to data protection and data cleaning issues, the data in the Scientific Use File (SUF) may differ slightly from the data set used for analyses in this paper. However, fundamentally different results are not expected.

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. (2013) 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.

The items are not arranged in units. Thus, 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:

- quantity,
- space and shape,
- change and relationships,
- data and chance.

Each item was constructed in such a way as to primarily address a specific content area. The framework also describes as a second and independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

In the mathematics test there are three types of response formats. These are simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR). In MC items the test taker has to find the correct answer from several, usually four, response

options. In CMC tasks a number of subtasks with two response options are presented. SCR items require the test taker to write down an answer into an empty box.

3. Data

3.1 The Design of the Study

The study assessed different competence domains including, among others, reading competence and mathematical competence. The competence tests for these domains were always presented first within the test battery. In order to control for test position effects, the two tests were assigned to test takers in different order. Half of the subjects received a booklet that contained the reading test first followed by the mathematics test, while the other half of the sample received the two tests in the opposite order. The subjects were assigned to one of the two booklets keeping the order from 5th grade test. Subjects that did not take the test in grade 5 were assigned randomly to one of the two booklets. There was no multi-matrix design regarding the order of the items *within* the mathematics test. All students received the same mathematics items in the same order.

The mathematics test in grade 7 consisted of 23 items which represented different content-related and process-related components and used different response formats. The characteristics of the 23 items are depicted in the following tables. Table 1 shows the distribution of the four content areas, whereas Table 2 shows the distribution of response formats. The CMC item originally included four subtasks; but one subtask was excluded from the analysis due to an unsatisfactory item fit.

Table 1: Number of Items by Content Areas

Content area	Frequency
Quantity	5
Space and shape	5
Change and relationships	7
Data and chance	6
Total number of items	23

Table 2: Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	20
Complex Multiple-Choice	1
Short-constructed response	2
Total number of items	23

3.2 Sample

A total of 6,194 students received the mathematics test. For three respondents less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paperare based on a sample of 6,191 test takers. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

3.3 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.

In this study, all respondents received the same set of items. As a consequence, there are no items that were not administered to a person. Invalid responses occured, for example, when two response options were selected where only one was required or when simply illegible answers were provided in the SCR format. Omitted items occurred when test takers skipped some items. Due to time limits not all persons finished the test within the given time limit. All missing responses after the last valid response were coded as not reached. 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. 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 the items functioned.

3.4 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. For item mag7r02s_c the two lowest CMC item categories were collapsed.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items 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 competencies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in the form of plausible values (Mislevy, 1991). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 6.

3.5 Checking the Quality of the Scale

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

Before aggregating the subtasks of CMC items 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 error (WMNSQ), the respective *t*-value, point -biserial correlations of the responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC variables that were included in the final scaling model.

The MC items consisted of one correct response option and one or more distractors (i.e., incorrect response options). The quality of the distractors within MC items was evaluated using the point-biserial correlation between selecting an incorrect response and the total correct 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 polytomous variables, 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 > 1.2 (t-value > |8|) were judged as a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total correct score (equal to the discrimination value as computed in ConQuest) greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The mathematical competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, school types (high school vs. non-high school), the number of books at home (as a proxy for socioeconomic status), the position of the test, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) was examined using a

multi-group 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 considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 as small 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 competence data in the NEPS are 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 may 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 dimensionality of the mathmatics test was evaluated by specifying a four-dimensional model based on the four content areas. Every item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Monte Carlo estimation in ConQuest was used (the number of nodes per dimension was chosen in such a way as to obtain stable parameter estimates). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The IRT models were estimated in ConQuest version 2.0 (Wu, Adams, & Wilson, 1997). The 2PL model was estimated in mdltm (Matthias von Davier, 2005).

4. Results

4.1 Missing Responses

4.1.1 Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was very small. In fact, 96.1% of test takers gave no invalid response at all. Less than 4% of the respondents had one or more invalid responses.

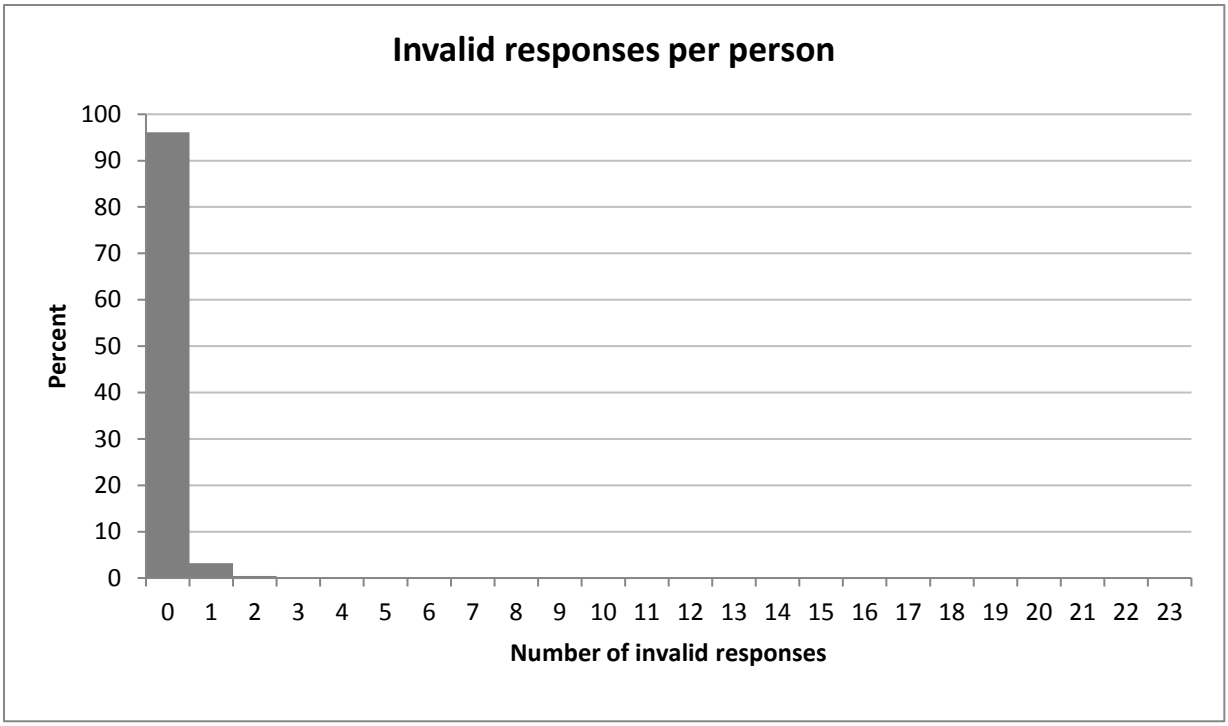

Figure 1: Number of invalid responses

Missing responses may also occur when test takers skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. It shows that 74.5% of the respondents omitted no item at all, whereas 0.9% of the respondents omitted more than 5 items.

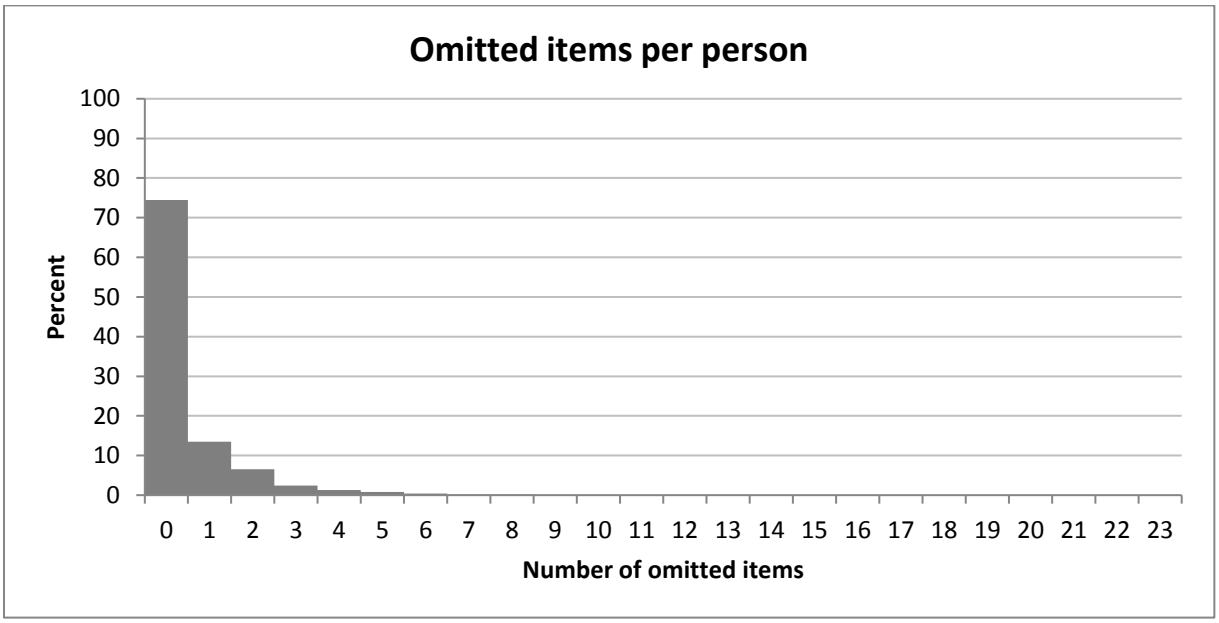

Figure 2: Number of omitted items

All missing responses after the last valid response are defined as not reached. Figure 3 shows the number of items that were not reached by a person. As can be seen, only 92.1% reached the end of the test, whereas 6.4% of the test takers did not reach one to five items. Only 1.5% of the students did not reach more than five items.

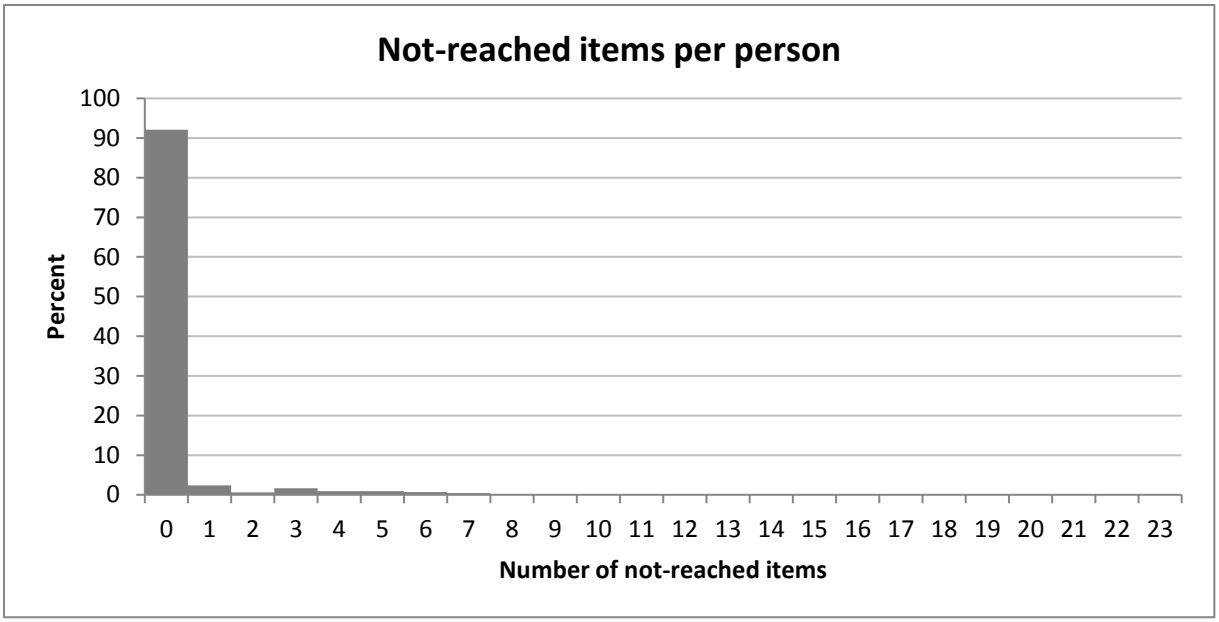

Figure 3: Number of not-reached items

Figure 4 shows the total number of missing responses per person, which is the sum of invalid, omitted, not-reached, and not-determinable missing responses. In total, 67.3% of the test takers showed no missing response at all, whereas 3.0% showed more than five missing responses.

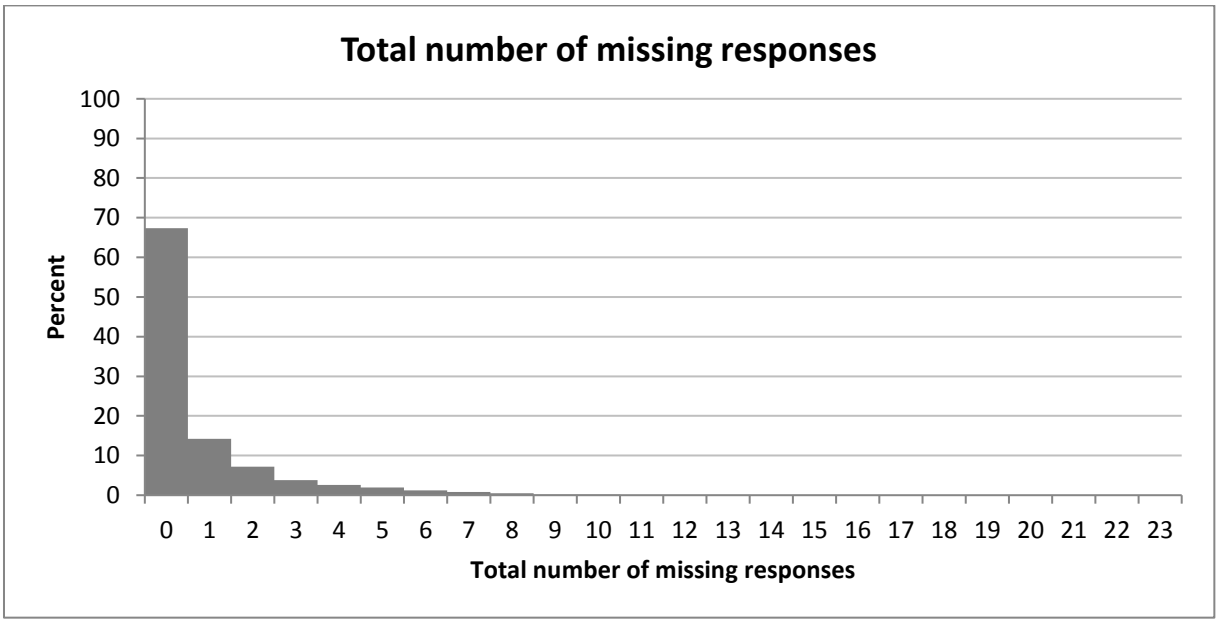

Figure 4: Total number of missing responses

Overall, there was a negligible amount of invalid, and a reasonable amount of not-reached or omitted items.

4.1.2 Missing responses per item

Table 3 shows the number of valid responses for each item, as well as the percentage of missing responses.

Overall, the number of invalid responses per item was very small. The omission rates were acceptable. Two items were omitted by more than 5% of the participants. The highest omission rate (6.99%) occurred for the most difficult item mag9v091_sc3g7_c.

The number of persons that did not reach an item increased with the position of the item in the test up to 7.88%.

The total number of missing responses per item varied between 0.66% (item mag5d051_sc3g7_c) and 12.02% (item mag9v091_sc3g7_c).

Table 3: Percentage of Missing values

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted missings	Percentage of not-reached items
mag9q071_sc3g7_c	1	6,079	0.39	1.42	0.00
mag7v071_c	2	6,043	0.06	2.33	0.00
mag7r081_c	3	6,083	0.60	1.15	0.00
mag7q051_c	4	6,139	0.18	0.66	0.00
mag5q301_sc3g7_c	5	6,124	0.11	0.97	0.00
mag9d151_sc3g7_c	6	6,105	0.44	0.95	0.00
mag5d051_sc3g7_c	7	6,150	0.05	0.61	0.00
mag5d052_sc3g7_c	8	6,129	0.40	0.60	0.00
mag9v011_sc3g7_c	9	6,055	0.08	2.12	0.00
mag9v012_sc3g7_c	10	5,924	0.05	4.26	0.00
mag7q041_c	11	6,065	0.00	2.02	0.02
mag7d042_c	12	6,146	0.08	0.61	0.03
mag7r091_c	13	6,039	0.08	2.29	0.08
mag9q181_sc3g7_c	14	6,120	0.06	0.95	0.13
mag7d011_c	15	6,136	0.05	0.61	0.23
mag7v012_c	16	6,014	0.06	2.47	0.32
mag7v031_c	17	5,992	0.13	2.31	0.78
mag5r251_sc3g7_c	18	5,827	0.05	4.36	1.47
mag7d061_c	19	5,741	0.03	4.81	2.42
mag5v321_sc3g7_c	20	5,600	0.61	5.62	3.31
mag9v091_sc3g7_c	21	5,447	0.08	6.99	4.94
mag5r191_sc3g7_c	22	5,715	1.11	1.07	5.51
mag7r02s_c	23	5,654	0.19	0.44	7.88

4.2 Parameter Estimates

4.2.1 Item parameters

In order to get a first descriptive measure of the item difficulties and check for possible estimation problems, we evaluated the relative frequency of the responses given before performing any IRT analyses. Using each subtask of a CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 27.43% and 93.35% across all items. On average, the rate of correct responses was 60.87% (*SD* = 17.54%). From a descriptive point of view, the items covered a relatively wide range of difficulties.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are depicted in Table 4a. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The step parameters of the polytomous item are depicted in Table 4b. The estimated item difficulties varied between -3.132 (item mag5d051_sc3g7_c) and 1.183 (item mag9v091_sc3g7_c) with a mean of -0.559. Overall, the item difficulties were distributed well across the proficiency scale, although with a tendency to being too easy. However, there were no very difficult items. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were very small (*SE*(ß) ≤ 0.05).

Table 4a: Item Parameters

Item	Posi-tion	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.
mag9q071_sc3g7_c	1	55.80	-0.284	0.029	1.00	0.2	0.49	0.98
mag7v071_c	2	39.67	0.533	0.029	1.16	12.6	0.32	0.49
mag7r081_c	3	45.18	0.247	0.029	1.07	5.9	0.42	0.75
mag7q051_c	4	43.13	0.348	0.029	0.97	-2.8	0.52	1.10
mag5q301_sc3g7_c	5	53.33	-0.153	0.029	0.92	-7.0	0.57	1.33
mag9d151_sc3g7_c	6	74.63	-1.328	0.032	0.95	-2.9	0.50	1.22
mag5d051_sc3g7_c	7	93.35	-3.132	0.054	0.94	-1.5	0.35	1.54
mag5d052_sc3g7_c	8	81.06	-1.780	0.035	0.95	-2.5	0.46	1.23
mag9v011_sc3g7_c	9	60.78	-0.543	0.029	0.94	-5.1	0.55	1.25
mag9v012_sc3g7_c	10	45.49	0.238	0.029	0.99	-0.9	0.50	1.05
mag7q041_c	11	62.37	-0.634	0.030	0.97	-2.8	0.52	1.12
mag7d042_c	12	82.05	-1.860	0.036	1.01	0.6	0.39	0.94
mag7r091_c	13	51.85	-0.090	0.029	0.98	-1.9	0.52	1.10
mag9q181_sc3g7_c	14	81.54	-1.823	0.036	0.97	-1.7	0.44	1.16
mag7d011_c	15	74.45	-1.321	0.032	1.06	3.8	0.39	0.78
mag7v012_c	16	54.32	-0.213	0.029	0.99	-0.5	0.50	1.02
mag7v031_c	17	57.21	-0.360	0.029	1.07	5.4	0.44	0.80
mag5r251_sc3g7_c	18	59.93	-0.490	0.03	1.00	0.2	0.49	0.99
mag7d061_c	19	36.49	0.684	0.031	1.12	9.3	0.35	0.57
mag5v321_sc3g7_c	20	45.38	0.259	0.03	0.99	-0.9	0.50	1.03
mag9v091_sc3g7_c	21	27.43	1.183	0.034	0.94	-4.0	0.49	1.25
mag5r191_sc3g7_c	22	71.34	-1.142	0.032	0.93	-4.4	0.53	1.30
mag7r02s_c	23	n.a.	-1.191	0.043	1.02	1.4	0.33	0.41

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, rit = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).

Percent correct scores are not informative for polytomous CMC and MA 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).

Table 4b: Step Parameters of Polytomous Item

Item	Position in the test	step 1 (*SE*)	step 2
mag7r02s_c	23	-1.088 (0.027)	1.088

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person´s abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 5, item difficulties of the reading 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 1.156, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = 0.761, WLE reliability = 0.721) was good. Although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person abilities in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors.

Scale in logits	Person ability	Item difficulty
3		
	X	
	X	
	X	
	XX	
2	XX	
	XX	
	XX	
	XXXX	
	XXXX	
	XXXXX	
	XXXXXXX	21
1	XXXXXX	
	XXXXXX	
	XXXXXXX	19
	XXXXXXXX	2
	XXXXXXXXX	4
	XXXXXXXXX	3 10 20
	XXXXXXXX	
0	XXXXXXXXX	13
	XXXXXXXX	1 5 16
	XXXXXXXXX	17
	XXXXXXXXX	9 18
	XXXXXXXXX	11
	XXXXXXX	
-1	XXXXXXX	
	XXXX	22
	XXXXX	6 15 23
	XXXXX	
	XXXX	
	XX	
	XX	8 12 14
-2	XX	
	X	
	X	
	X	
	X	
-3		
		7

Figure 5: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 36.4 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4a).

4.3 Quality of the test

4.3.1 Distractor analyses

To investigate how well the distractors performed in the test, we evaluated – for the MC items – the point-biserial correlations between selecting each incorrect response (distractor) and the students' total correct scores. This distractor analysis was performed on the basis of preliminary analyses treating all subtasks of the CMC item as single items. The point-biserial correlations for the distractors ranged from -0.42 to -0.01 with a mean of -0.21. These results indicate that the distractors worked well. In contrast, the point-biserial correlations between selecting the correct response and student's total correct scores ranged from 0.32 to 0.54 with a mean of 0.45 indicating that more proficient students were also more likely to identified the correct response option.

Table 5: Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.45	-0.21
Minimum	0.32	-0.42
Maximum	0.54	-0.01

4.3.2 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 and polytomous CMC items. Altogether, item fit can be considered to be very good (see Table 4a). Values of the WMNSQ were close to 1 with the lowest value being 0.92 (item mag5q301_sc3g7_c) and the highest being 1.16 (item mag7v071_c). Thus, only the latter showed a small misfit with a WMNSQ above 1.15 (cf. Pohl & Carstensen, 2009). This item had a satisfactory discrimination (r_{it} = .32) even though it had one of the three highest difficulties in this rather easy test. So, overall, the item had an acceptable fit and it was not necessary to exclude the item from the test. The two items with the largest WMNSQ (mag7d061_c and mag7v071_c) showed acceptable, slightly flat item characteristic curves (ICC). Therefore, all ICC showed a good or very good fit of the items. Overall, there was no indication of severe item over- or underfit. The correlations of the item scores with the total scores varied between 0.32 (item mag7v071_c) and 0.57 (item mag5q301_sc3g7_c) with an average correlation of 0.46.

4.3.3 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, school types, the number of books at home, the position of the test, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 6 shows the difference between the estimated difficulties of the items in different subgroups. Female versus male, for example, indicates the difference in difficulty between girls and boys, *ß*(female) – *ß*(male). A positive value indicates a higher difficulty for females, a negative value a lower difficulty for females compared to males.

Table 6: Differential Item Functioning

Item	Gender	Position	Migration status	Books	School type
	female vs. male	Math/Reading test vs. Reading/Math Test	without vs. with	<100 books vs. >100 books	non-high school vs. high school
mag9q071_sc3g7_c	-0.226	0.180	0.226	-0.024	0.068
mag7v071_c	0.212	0.038	0.300	-0.334	-0.554
mag7r081_c	0.084	0.030	-0.064	-0.198	-0.242
mag7q051_c	-0.152	0.086	-0.126	0.366	0.286
mag5q301_sc3g7_c	0.390	0.182	-0.158	0.510	0.578
mag9d151_sc3g7_c	-0.062	-0.090	-0.186	0.118	0.224
mag5d051_sc3g7_c	-0.004	-0.300	-0.120	-0.094	0.080
mag5d052_sc3g7_c	0.132	-0.128	-0.378	0.058	0.158
mag9v011_sc3g7_c	-0.238	-0.116	-0.238	0.042	0.180
mag9v012_sc3g7_c	-0.160	-0.012	0.036	-0.032	-0.030
mag7q041_c	-0.126	-0.184	-0.108	0.064	0.196
mag7d042_c	0.198	-0.206	0.228	-0.300	-0.220
mag7r091_c	-0.136	-0.008	0.130	0.102	0.218
mag9q181_sc3g7_c	0.142	-0.078	-0.014	-0.048	-0.004
mag7d011_c	0.208	-0.012	-0.014	-0.106	-0.276
mag7v012_c	-0.080	0.028	0.034	-0.132	-0.076
mag7v031_c	0.142	0.116	0.206	-0.228	-0.248
mag5r251_sc3g7_c	0.200	-0.068	0.006	0.004	-0.060
mag7d061_c	-0.048	0.140	0.386	-0.308	-0.440
mag5v321_sc3g7_c	-0.180	-0.092	-0.202	0.064	-0.080
mag9v091_sc3g7_c	-0.520	-0.022	0.044	0.088	0.078
mag5r191_sc3g7_c	0.400	0.102	-0.128	0.296	0.336
mag7r02s_c	-0.144	0.076	-0.124	-0.044	-0.164
Main effect (model with DIF)	**0.340**	**-0.108**	**-0.560**	**0.806**	**1.242**
Main effect (Model without DIF)	**0.338**	**-0.110**	**-0.554**	**0.806**	**1.248**

Overall, 2,993 (48.3%) of the test takers were female and 3,197 (51.6%) were male, one student did not give a valid response. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.340 logits, Cohen's *d* = 0.309). There was no item with a large DIF with regard to gender. The only item for which the difference in item difficulties between the two groups exceeded 0.4 logits was item mag9v091_sc3g7_c (0.520 logits). However, this differences was not considered severe.

The test takers received either the mathematics or the reading test first. Whether there was a test position effect was analyzed through another DIF analysis. There were 3,071 (49.6%) subjects who solved the mathematics test first, whereas 3,120 (50.4%) students received the mathematics tests after the reading test. There were no considerable average differences between the two groups (main effect = 0.108 logits, Cohen's *d* = 0.097). There was also no considerable DIF comparing participants with the different test position.

There were 4,199 (67.8%) participants without migration background, 1,314 (21.2%) participants with migration background, and 678 (11.0%) participants without a valid response. Only the first two groups were used for investigating DIF of migration. On average, participants without migration background performed considerably better in the mathematics test than those with migration background (main effect = 0.56 logits, Cohen's *d* = 0.519). There was no considerable DIF comparing the two groups. The largest difference in item difficulties between the two groups was 0.386 logits (mag7d061_c).

The number of books at home was used as a proxy for socioeconomic status. There were 2,336 (37.7%) test takers with 0 to 100 books at home, 3,607 (58.3%) test takers with more than 100 books at home, and 248 (4.0 %) test takers without a valid response. Group differences and DIF were investigated by using the first two groups. Participants with 100 or less books at home performed on average 0.806 logits (Cohen's *d* = 0.779) worse in mathematics than participants with more than 100 books. Comparing the two groups, DIF exceeding 0.4 logits occurred in item mag5q301_sc3g7_c (0.510 logits).

3,282 (53.0 %) of the participants were highschool students, whereas 2,909 (47.0 %) attended different types of school. On average, highschool students showed a considerably higher mathematical competence than the other students (main effect = 1.242, Cohen's d = 1.358). There was no item with a considerable DIF. Differences in item difficulties exceeding 0.4 logits were observed in three items (mag7v071_c, mag5q301_sc3g7_c, and mag7d061_c), the maximum being 0.578 logits.

In Table 7, we compared the models that included only main effects to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for all five DIF variables. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents an overparametrization of models. Using BIC, the more parsimonious models including only the main effects of migration status and test position were preferred over the more complex DIF models. However, BIC preferred the models including both main effect and DIF effects of gender, the number of books and school type, respectively, to the models including only the respective main effect. (Note that the analyses including migration contain fewer cases and, thus, the information criteria cannot be compared across analyses with different DIF variables.)

Table 7: Comparison of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	main effect	156,894.83	26	156,946.83	157,121.82
	DIF	156,627.77	49	156,725.77	157,055.57
Migration status	main effect	139,487.58	26	139,539.58	139,711.57
	DIF	139,358.26	49	139,456.26	139,780.39
Position	main effect	157,050.98	26	157,102.98	157,277.98
	DIF	156,977.15	49	157,075.15	157,404.96
Books	main effect	149,912.01	26	149,964.01	150,137.95
	DIF	149,672.69	49	149,770.69	150,098.50
School type	main effect	155,203.71	26	155,255.71	155,430.72
	DIF	154,796.18	49	154,894.18	155,223.99

4.3.4 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 different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4a), ranging from 0.41 (item mag7r02s_c) to 1.54 (items mag5d051_sc3g7_c). The average discrimination parameter fell at 1.02. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 156,766.353, BIC = 157,183.66, number of parameters = 39) as compared to the 1PL model (AIC = 116,321.26, BIC = 116,568.55, number of parameters = 37). Despite the empirical preference for the 2PL model, the 1PL model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model (1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

Note that these calculations could not be made by conquest 2.0 so that we had to use a substitude Programm called MDLTM (see 3.6, Davier, 2005). In consequence, the results for AIC and BIC using the 1PL model might differ from the later results (see 4.3.5) comparing multi-dimensionality to unidimensionality of the test.

4.3.5 Unidimensionality

The unidimensionality of the test was investigated by specifying a four-dimensional model based on the four different content areas. Each item was assigned to one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Monte Carlo estimation implemented in ConQuest was used. The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained. The variances and correlations of the four dimensions are shown in Table 8. Three of the four dimensions exhibit a substantial variance. In the fourth dimension (dimension 3) six out of seven items show difficulties ranging from -0.543 to +0.533 so the difficulties are very homogenous in this dimension, which could explain the small variance. The correlations between the four dimensions were – as expected – very high, varying between 0.930 and 0.953, and, thus, indiciated an

essentially unidimensional test (cf. Carstensen, 2013). Moreover, according to model fit indices, the unidimensional model fitted the data slightly better (see Table 9)) than the four-dimensional model . These results indicate that the three cognitive requirements measure a common construct.

Table 8: Results of Four-Dimensional Scaling

	Dim 1	Dim 2	Dim 3	Dim 4
Quantity (5 items)	1.737			
Space and shape (5 items)	0.950	1.150		
Change and relationships (7 items)	0.953	0.940	0.488	
Data and chance (6 items)	0.930	0.930	0.940	1.244

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

Table 9: Comparison of the Unidimensional and the Four-Dimensional Model.

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	157063.129	25	157113.129	157281.400
Four-dimensional	158907.585	34	158975.585	159204.430

Note. Contrary to the calculations for the 1PL and 2PL models, results in this table were achieved by using Conquest 2.0.

5. Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 3 and at describing how the mathematics competence score had been estimated.

The amount of different kinds of missing responses was evaluated and all kinds of missing responses were rather low. Furthermore item as well as test quality were examined. As indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, ICC – the items exhibited a good item fit. Also, discrimination values of the items (either estimated in a 2PL model or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test is fair for the examined subgroups.

The test had a good reliability and distinguished well between test takers, as indicated by the test's variance. The item distribution along the ability scale was acceptable, although the test had a tendency to being slightly too easy for the sample.

Fitting a four-dimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly worse model than the unidimensional

partial credit model. Moreover, high correlations between the four dimensions indicate that the unidimensional model described the data well.

In summary, the test had good psychometric properties that facilitated the estimation of a unidimensional mathematics competence score.

6. Data in the Scientific Use File

6.1 Naming conventions

The data in the Scientific Use File contain 23 items, of which 22 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. One item was scored as a polytomous variable (CMC items). MC items are marked with a '_c' at the end of the variable name, whereas the variable names of CMC items end in 's_c'. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Items that were already administered in grade 5 kept their original names ('mag5q301…', 'mag5d051…', 'mag5d052…', 'mag5r251…', 'mag5v321…' and 'mag5r191…'). However, for reasons of identification a suffix was added in front of the '…_c' (scored item) to specify the current test administration ('sc3g7' referring to Starting Cohort 3, grade 7).

6.2 Linking the data of Grade 5 and Grade 7

In starting cohort 3, the mathematics competence tests administered in grades 5 (see Duchardt & Gerdes., 2013) and 7 for the large part 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 competencies as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competencies across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). The process of linking combines adjacent measurement points on the same scale. As such, the first wave of each competence scale within a cohort is used as reference scale that all subsequent measurement waves will refer to. For the domain of mathematical competence, linking is achieved using overlapping items (also known as common items). In order to link the tests of mathematical competence conducted in grade 5 and grade 7, six items which already were administered in grade 5 were, again, administered in grade 7 (e.g., mag5q301_sc3g7_c).

An empirical study that evaluated different link methods with regard to the appropriateness of linking NEPS data (Fischer et al., 2016) showed that the method of mean/mean linking (see Kolen & Brennan, 2004) is appropriate for the present test. All of the six common items that were administered in grade 5 and grade 7 were found to be measurement invariant across the two measurement points. As such, they served as link items. For more information on the selection of link items and the method for linking the tests of mathematical competence in starting cohort 3 (grade 5 and grade 7) see Fischer et al. (2016).

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs, mag7_sc1 and mag7_sc1u, including their respective standard errors, mag7_sc2 and mag7_sc2u. Both WLE scores are linked to the underlying reference scale of grade 5. Mag7_sc1u is uncorrected for the position of the math test within the booklet and can be used, if the focus of research lies on longitudinal issues, such as competence development. Therefore, resulting differences in WLE scores can be interpreted as competence development across measurement points. Consequently, mag7_sc1 that corrected for the position of the math test within the booklet can be used if the research interest is based on cross-sectional issues. The ConQuest Syntax for estimating the WLE scores from the items are provided in Appendix A, the fixed item parameters are provided in Appendix B. Students that did not take part in the test or those that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE scores for mathematical competence.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-722.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: MIT Press.

Davier, M. von, (2005). A general diagnostic model applied to language testing data (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

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.

Duchhardt, C. (2015). *NEPS Technical Report for Mathematics—Scaling results for the additional study Baden Wuerttemberg* (NEPS Working Paper No. 59). Bamberg: Leibniz Institute for Educational Trajectories, 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* (pp. 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.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2012). Incorporating different response formats in the IRT-scaling model for competence data. Manuscript submitted for publication.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *Technical report of reading – Scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Jordan, A.-K., & Duchhardt, C. (2013). *NEPS Technical Report for Mathematics—Scaling results of Starting Cohort 6–Adults* (NEPS Working Paper No. 32). Bamberg: University of Bamberg, National Educational Panel Study.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (pp. 201-205). New York: Springer.

Koller, I., Haberkorn, K., & Rohm, T. (2014). *NEPS Technical Report for Mathematics—Scaling results of Starting Cohort 6 for adults in main study 2012* (NEPS Working Paper No. 48). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177-196.

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online*, 5(2), 80-102.

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 the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, 5(2), 189-216.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *Technical report of reading – Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Haberkorn, K., Carstensen, C.H. (2015). *Measuring competencies across the lifespan – Challenges of linking test scores.* In M. Stemmler, A. von Eye, &W. Wiedermann (EDS), *Dependent data in social science research* (pp.281.308). Berlin, Germany: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded edition, 1980, Chicago: University of Chicago Press).

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & von Maurice, J. (Eds.). *Education as a lifelong process: The German National Educational Panel Study (NEPS).* (pp. 67-86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ACER Conquest: Generalised item response modelling software*. Melbourne: ACER Press.

Appendix

Appendix A: ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort III

```
Title Starting Cohort III, MATHEMATICS: Partial Credit;

data filename.dat;
format pid 1-07 responses 15-37;
labels << labels.nam;

codes 0,1,2;

score (0,1) (0,1)          !item(1-22);
score (0,1,2) (0,0.5,1)    !item(23);

model item + item*step
set constraint=cases;

estimate! method=quadrature, nodes=15, converge=.0001;
show !estimates=latent >> filename.shw;
itanal >> filename.ita;
show cases !estimates=wle >> filename.wle;
```

Appendix B: Fixed Item Parameters

```
1       0.44143     /* item mag9q071_sc3g7_c */
2       1.25859     /* item mag7v071_c */
3       0.97236     /* item mag7r081_c */
4       1.07410     /* item mag7q051_c */
5       0.57271     /* item mag5q301_sc3g7_c */
6      -0.60250     /* item mag9d151_sc3g7_c */
7      -2.40669     /* item mag5d051_sc3g7_c */
8      -1.05405     /* item mag5d052_sc3g7_c */
9       0.18273     /* item mag9v011_sc3g7_c */
10      0.96409     /* item mag9v012_sc3g7_c */
11      0.09185     /* item mag7q041_c */
12     -1.13458     /* item mag7d042_c */
13      0.63589     /* item mag7r091_c */
14     -1.09693     /* item mag9q181_sc3g7_c */
15     -0.59524     /* item mag7d011_c */
16      0.51259     /* item mag7v012_c */
17      0.36619     /* item mag7v031_c */
18      0.23529     /* item mag5r251_sc3g7_c */
19      1.40956     /* item mag7d061_c */
20      0.98465     /* item mag5v321_sc3g7_c */
21      1.90919     /* item mag9v091_sc3g7_c */
22     -0.41600     /* item mag5r191_sc3g7_c */
23     -0.46519     /* item mag7r02s_c */
24     -1.088       /* step parameter mag7r02s_c */
25      0.05442     /* correcting for test position -  first position*/
26     -0.05442     /* correcting for test position -  second position*/
```