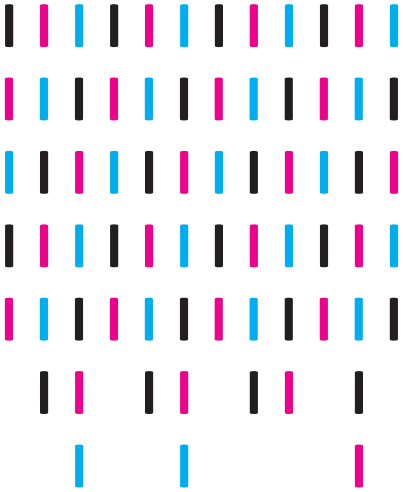

NEPS SURVEY PAPERS

Insa Schnittjer and Anna-Lena Gerken

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 2

NEPS Survey Paper No. 47
Bamberg, November 2018

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 2 for Grade 2

Insa Schnittjer[1,2] and Anna-Lena Gerken[1]

[1]IPN – Leibniz Institute for Science and Mathematics Education at Kiel University
[2]University of Koblenz-Landau

Email address of the lead author:

schnittjer@uni-landau.de

Bibliographic Data:

Schnittjer, I.,& Gerken, A.-L. (2018): *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 2 for Grade 2* (NEPS Survey Paper No. 47). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP47:1.0

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., Schnittjer, 2016).

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 2 for Grade 2

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 2 of starting cohort 2 (kindergarten). The mathematics test contained 24 items with different response formats representing different content areas and cognitive components. The test was administered to 6,168 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 three complexe multiple choice 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

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) and Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models that are based on the 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 2 of starting cohort 2 (kindergarten). First, the main concepts of the mathematical test are introduced. Subsequently, the mathematical competence data of starting cohort 2 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:

- sets, numbers and operations,
- units and measuring,
- 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 (see Apendix C). 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 two types of response formats. These are simple multiple-choice (MC) and complex multiple-choice (CMC). In MC items the test taker has to find the correct answer from four response options. In CMC tasks a number of subtasks with two response options are presented.

3 Data

3.1 The Design of the Study

The study assessed different competence domains including mathematical competence, reading competence, and domain-general cognitive functioning. The competence tests for these domains took place on two testing days. The tests were always presented in the same order, starting with the mathematics test on the first testing day. The other two tests took place on the second testing day. 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. A special challenge of this test was to take into account that the reading competences of this age group are very heterogeneous. Therefore, the items were read out to the children from a test administrator. There were up to 20 children in one test session. As a consequence, it was up to the test administrator to keep the time limits for the whole group in mind.

The mathematics test in grade 2 consisted of 24 items which represented different content-related and process-related components and used different response formats. The characteristics of the 24 items are depicted in the following tables. Table 1 shows the distribution of the five content areas, whereas table 2 shows the distribution of response formats. Two of the three CMC items included four subtasks, the third CMC item included five subtasks.

Table 1: Number of Items by Content Areas

Content area	Frequency
Sets, numbers and operations	5
Units and measuring	4
Space and shape	5
Change and relationships	5
Data and chance	5
Total number of items	24

Table 2: Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	21
Complex Multiple-Choice	3
Total number of items	24

3.2 Sample

A total of 6,168 students took the mathematics test. For one respondent less than three valid item responses were available. Because no reliable ability scores can be estimated based on

such few responses, this case was excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 6,167 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. Omitted items occurred when test takers skipped some items. Due to time limits, not all the test administrators could finish the last item instruction in time or some children did not follow the instructions to the end of the test. All missing responses after the last valid response were coded as not reached, regardless of whether it was due to slow instructions given from the test administrator, or due to individual reasons. 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 non-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. This happened for all three of the CMC items. For the items mag1d09s_sc2g2_c and mag2g12s_c, the two lowest categories were collapsed. For item mag1r19s_sc2g2_c, the four lower categories had to be collapsed, so it was consequently scored dichotomously.

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). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF are 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 three distractors. 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 good, greater than 0.2 acceptable, and below 0.2 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, the number of books at home (as a proxy for socioeconomic status), 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 mathematics test was evaluated by specifying a five-dimensional model based on the five content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Quasi Monte Carlo integration in TAM in R was used. To guarantee the compatibility with the multidimensional model, the unidimensional model was estimated in TAM as well. The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (11,000 nodes). 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 4.5.2 (Adams, Wu, & Wilson, 2015).The GPCM model was estimated in MDLTM (Matthias von Davier, 2005). To check the multidimensionality, the IRT models were estimated in TAM version 2.4-9 (Kiefer, Robitzsch, & Wu, 2016) in R version 3.4.1 (R Core Team, 2016) using the Quasi Monte Carlo integration with 11,000 nodes.

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, 93.4 % of test takers gave no invalid response at all. Less than 1 % of the respondents had more than one invalid response.

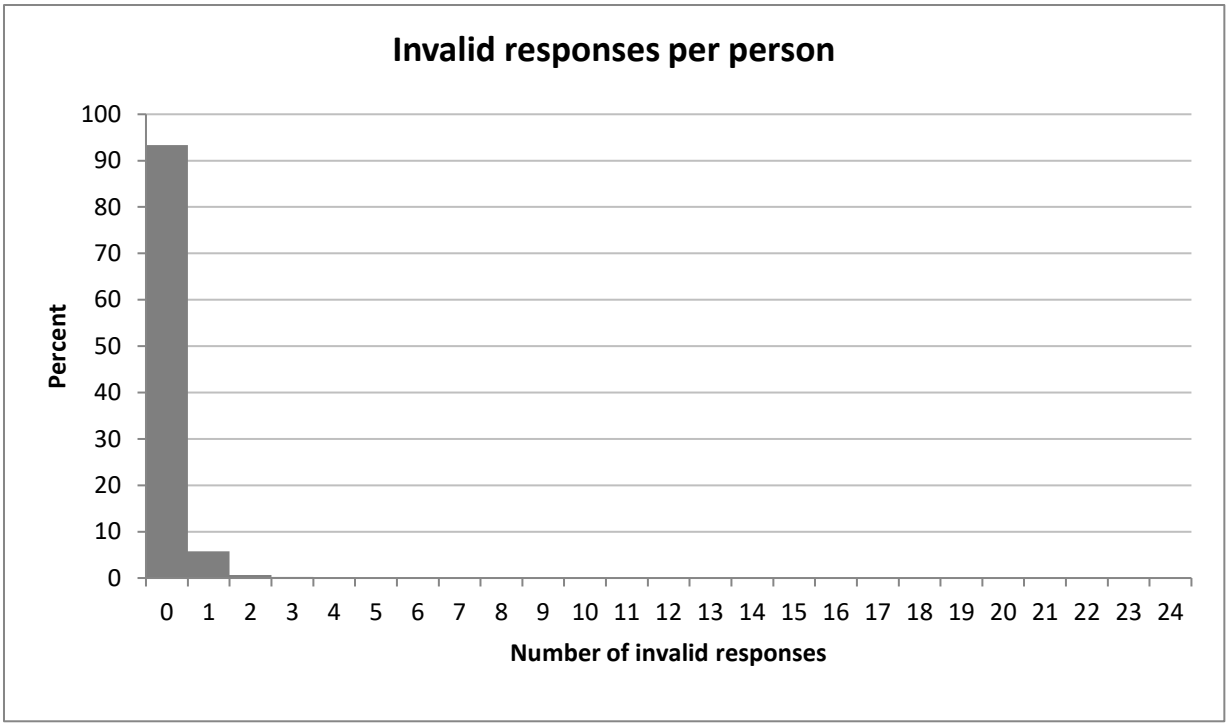

Figure 1: Number of invalid responses

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 65.0 % of the respondents omitted no item at all, whereas 1.2 % of the respondents omitted more than 5 items.

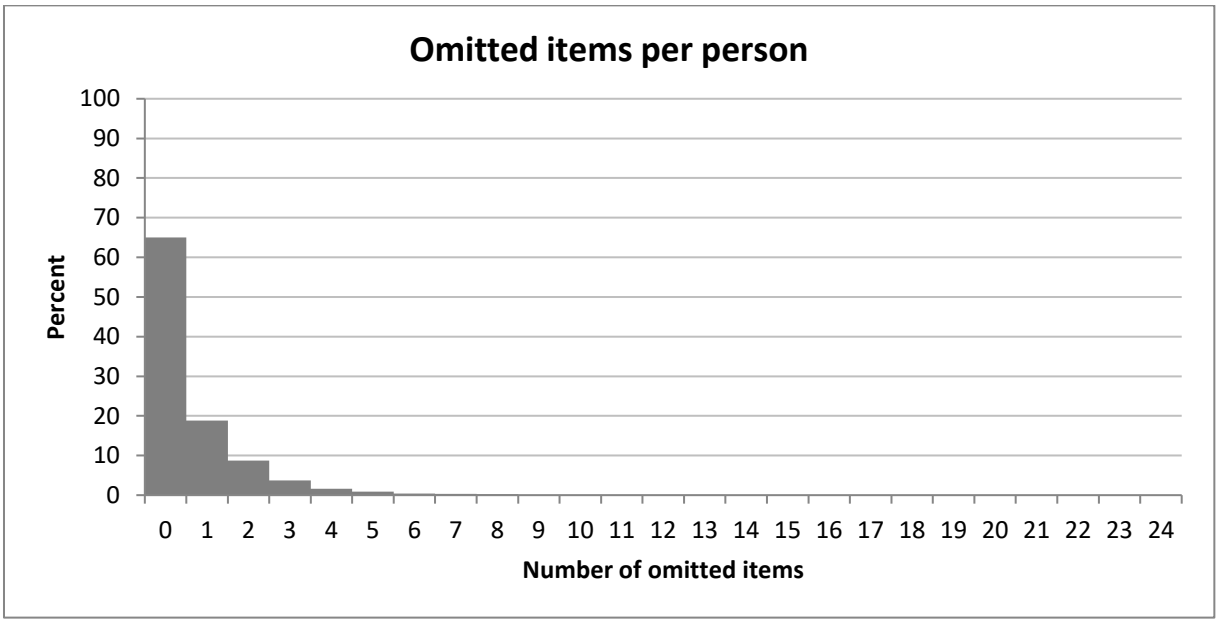

Figure 2: Number of omitted items

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, regardless of whether it was the single test taker that did work on the test till its end or whether the test administrator did not keep a reasonable pace in order to finish it within the time limit. As can be seen, 94.7 %

reached the end of the test, whereas 4.3 % of the test takers did not reach one to five items. Nevertheless, only 1.0 % of the students did not reach more than five items.

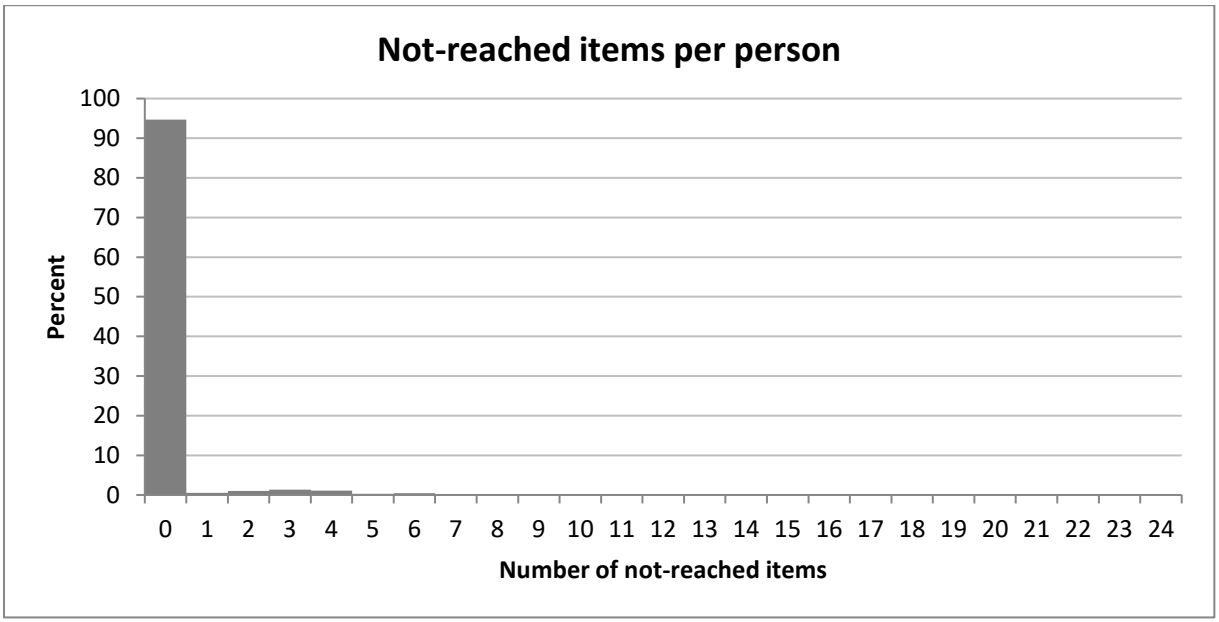

Figure 3: Number of not-reached items

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, 58.1 % of the test takers showed no missing response at all, whereas 2.8 % showed more than five missing responses.

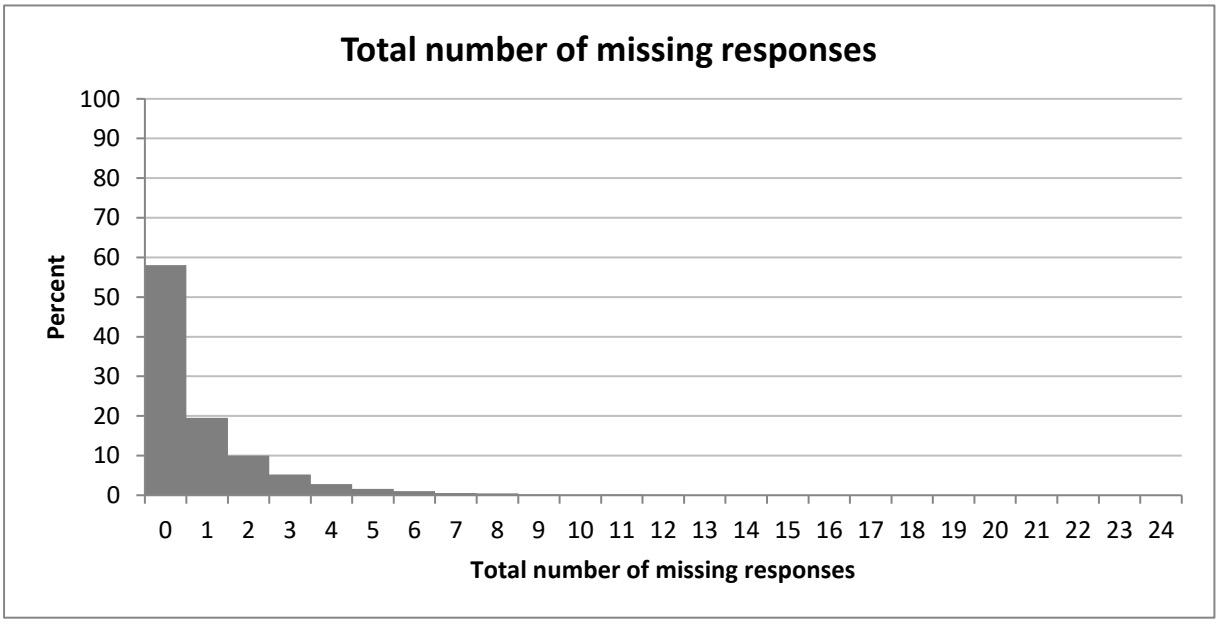

Figure 4: Total number of missing responses

Overall, there was a negligible number of invalid, and a reasonable number 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.

All CMC items displayed higher rates of invalid responses and even higher omission rates. Item 10 (mag1d09s_sc2g2_c) was explained line by line due to it being the first CMC item, which explaines the lower invalid and especially the lower omission rates. Some children seemed to have some difficulties with the CMC format, as 40 children selected only positive answers and 25 chose only negative answers. 79 children merely selected the first item and none else. Overall, 369 children only picked a single answer, the way that an MC task would have required it. Because a CMC task is classified as omitted if a subitem is omitted, the high omission rates of 7,17% (mag1r19s_sc2g2_c) or 10.5% (mag2g12s_c) respectively can be explained.

The number of persons that did not reach an item increased with the position of the item in the test up to 5.32%. The total number of missing responses per item varied between 0.96% (mag2r031_c) and 12.83% (mag1r19s_sc2g2_c).

Table 3: Percentage of Missing values

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items
mag1v051_sc2g2_c	1	6,027	0.18	2.09	0.00
mag2v071_c	2	5,898	0.02	4.35	0.00
mag2r031_c	3	6,108	0.05	0.91	0.00
mag2d061_c	4	5,929	0.41	3.45	0.00
mag1d131_sc2g2_c	5	5,808	0.49	5.33	0.00
mag2r131_c	6	5,731	0.19	6.86	0.02
mag2v121_c	7	6,055	0.19	1.61	0.02
mag2z061_c	8	5,951	0.28	3.21	0.02
mag2r111_c	9	5,998	0.16	2.56	0.02
mag1d09s_sc2g2_c	10	5,912	0.60	3.29	0.03
mag1z121_sc2g2_c	11	6,087	0.37	0.89	0.03
mag2g12s_c	12	5,454	0.83	10.49	0.05
mag1d081_sc2g2_c	13	6,082	0.52	0.81	0.05
mag2g021_c	14	6,033	0.23	1.75	0.19
mag2r151_c	15	6,098	0.44	0.49	0.19
mag1v021_sc2g2_c	16	5,977	0.31	2.58	0.19
mag1z071_sc2g2_c	17	6,023	0.06	1.93	0.34
mag2d101_c	18	6,049	0.21	1.20	0.50
mag1g031_sc2g2_c	19	6,023	0.23	1.12	0.99
mag2v041_c	20	6,017	0.02	1.17	1.25
mag2z011_c	21	5,833	0.06	3.00	2.35

mag1r19s_sc2g2_c	22	5,376	1.65	7.17	3.71
mag2g091_c	23	5,755	0.39	1.54	4.75
mag2z051_c	24	5,834	0.08	0.00	5.32

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, the relative frequency of the responses was evaluated 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 15.12% and 99.42% across all items. On average, the rate of correct responses was 67.45% (*SD* = 19.79%). From a descriptive point of view, the items covered a relatively wide range of difficulties.

Table 4a: Item Parameters

Item	Posi-tion	Percentage correct	Diffi-culty	*SE*	WMNS Q	*t*	r_{it}	Discr.
mag1v051_sc2g2_c	1	73.64	-1.250	0.033	0.95	-2.9	0.49	1.49
mag2v071_c	2	65.34	-0.779	0.032	1.03	2.1	0.45	1.14
mag2r031_c	3	84.22	-2.007	0.032	1.07	3.2	0.30	0.84
mag2d061_c	4	61.78	-0.584	0.032	0.96	-3.2	0.51	1.44
mag1d131_sc2g2_c	5	55.65	-0.267	0.041	0.94	-5.6	0.54	1.52
mag2r131_c	6	44.62	0.253	0.032	1.06	5.2	0.41	0.89
mag2v121_c	7	62.39	-0.633	0.034	0.98	-1.8	0.49	1.27
mag2z061_c	8	15.12	2.054	0.037	1.06	2.4	0.27	0.74
mag2r111_c	9	51.03	-0.056	0.043	1.12	11.0	0.35	0.66
mag1d09s_sc2g2_c	10	n.a.	-0.549	0.040	1.10	6.4	0.38	0.38
mag1z121_sc2g2_c	11	20.98	1.609	0.032	0.95	-2.7	0.44	1.30
mag2g12s_c	12	n.a.	-1.483	0.033	0.95	-3.2	0.52	0.69
mag1d081_sc2g2_c	13	83.39	-1.934	0.032	0.91	-4.0	0.48	1.99
mag2g021_c	14	39.57	0.512	0.032	1.03	3.0	0.42	0.97
mag2r151_c	15	64.66	-0.741	0.037	1.04	2.8	0.43	0.99
mag1v021_sc2g2_c	16	54.02	-0.199	0.036	1.00	0.1	0.48	1.17
mag1z071_sc2g2_c	17	58.92	-0.449	0.033	0.99	-1.0	0.49	1.28
mag2d101_c	18	79.14	-1.614	0.032	1.00	0.1	0.41	1.22
mag1g031_sc2g2_c	19	75.74	-1.383	0.035	0.88	-7.5	0.56	2.21
mag2v041_c	20	64.10	-0.718	0.033	0.92	-6.3	0.55	1.70
mag2z011_c	21	55.55	-0.288	0.039	1.05	4.2	0.43	0.99

mag1r19s_sc2g2_c	22	67.76	-0.829	0.033	1.12	7.9	0.33	0.54
mag2g091_c	23	61.93	-0.603	0.032	0.96	-3.2	0.52	1.49
mag2z051_c	24	81.98	-1.828	0.032	0.95	-2.4	0.45	1.71

Note. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (GPCM).
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).

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variables) 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 -2.007 (mag2r031_c) and 2.054 (mag2z061_c) with a mean of -0.574. Overall, the item difficulties were distributed acceptably well across the proficiency scale, with a tendency of being too easy. However, there were only four items with a difficulty above zero, two of them with a very high difficulty above 1.5. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were very small (*SE*(ß) ≤ 0.05).

Table 4b: Step Parameters of Polytomous Items

Item	Position in the test	step 1 (*SE*)	step 2 (*SE*)	step 3
mag1d09s_sc2g2_c	10	-0.506 (0.037)	-0.229 (0.038)	0.735
mag2g12s_c	12	-1.121 (0.045)	1.060 (0.044)	0.061

Note. mag1r19s_sc2g2_c was scored dichotomously and therefore cannot be found in table 4b but in table 4a.

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 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 theright side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.108, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = 0.798, WLE reliability = 0.787) was good. Although the items covered a wide range of the ability distribution (*Min*= -2.007, *Max*= 2.054), the items were overall too easy ($Mean_{Difficulty}$= -0,574). As a consequence, person abilities in medium- and low-ability regions were measured relatively precisely, whereas higher ability estimates had larger standard errors.

Scale in logits	Person ability	Item difficulty
	X	
	X	
	X	
	XX	
2	XX	8
	X	
	XXX	
	XXX	11
	XXXX	
	XXXXXX	
	XXXXX	
1	XXXXX	
	XXXXXXX	
	XXXXXXX	
	XXXXXXXXX	
	XXXXXXXXXX	14
	XXXXXXXXXX	
	XXXXXXXXX	6
	XXXXXXXXXX	
0	XXXXXXXXXX	9
	XXXXXXXXX	5 16
	XXXXXXXX	21
	XXXXXXXX	10 17
	XXXXXXXXX	4 7 23
	XXXXXXXXX	2 15 20 22
	XXXXXXXX	
-1	XXXXXXX	
	XXXXX	1
	XXXXX	19
	XXXX	12
	XXXX	18
	XXX	
	XX	13 24
-2	X	3
	X	
	X	

Figure 5: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 32.5 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, the point-biserial correlations between selecting each incorrect response (distractor) in MC items and the students' total correct scores was evaluated. 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 -0.41 to 0.12 (*Mean* = -0.17). Only one distractor

reached a positive correlation. 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 0.19 to 0.49 with a mean of 0.37 indicating that more proficient students were also more likely to identify 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	0.37	-0.17
Minimum	0.19	-0.41
Maximum	0.49	0.12

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 0.88 (mag1g031_sc2g2_c) and the highest being 1.12 (mag2r111_c and mag1r19s_sc2g2_c). The two items with the largest WMNSQ (mag2r111_c and mag1r19s_sc2g2_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 0.27 (mag2z061_c) and 0.56 (mag1g031_sc2g2_c) with an average correlation of 0.45. Thus, the value 0.27 was caused by the most difficult item of the test, its correlation was still acceptable and taking into account the lack of other difficult items, it even depicted a good correlation.

4.3.3 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). DIF was examined for the variables gender, the number of books at home, 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, β(female) – β(male). A positive value indicates a higher difficulty for females, a negative value a lower difficulty for females compared to males.

Overall, 3,159 (51.2 %) of the test takers were female and 3,007 (48.8 %) were male, one student did not give a valid response. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.350 logits, Cohen's d = 0.337). There was only one item for which the difference in item difficulties between female and male groups exceeded |0.6| logits (mag2r031_c). This item showed the lowest difficulty in the test. However, this item showed good fits in the other categories and belongs to the category of space and shape items with focus on space, so therefore there was no reason to exclude the item from the analyses. There were five items for which the gender DIF exceeded |0.4| logits (mag2g021_c, mag2z011_c, mag1r19s_sc2g2_c, mag2g091_c, mag2z051_c). However, these differences were considered not to be severe.

Table 6: Differential Item Functioning

Item	Position	Gender	Migration status	Books		
		female vs. male	without vs. with	missing vs. <100 books	missing vs. >100 books	<100 books vs. >100 books
mag1v051_sc2g2_c	1	-0.310	-0.116	-0.006	0.112	0.110
mag2v071_c	2	0.108	0.336	0.096	-0.196	-0.300
mag2r031_c	3	-0.630	-0.196	0.046	0.07	0.014
mag2d061_c	4	-0.326	0.148	0.104	0.084	-0.028
mag1d131_sc2g2_c	5	-0.344	-0.026	0.028	0.066	0.030
mag2r131_c	6	0.032	0.130	-0.002	-0.192	-0.198
mag2v121_c	7	0.088	-0.144	0.038	0.288	0.242
mag2z061_c	8	-0.366	-0.954	-0.166	0.064	0.222
mag2r111_c	9	-0.394	0.050	-0.142	-0.152	-0.018
mag1d09s_sc2g2_c	10	-0.204	0.124	0.018	0.032	0.016
mag1z121_sc2g2_c	11	-0.274	0.234	0.008	0.358	0.342
mag2g12s_c	12	0.204	0.064	-0.01	-0.034	-0.028
mag1d081_sc2g2_c	13	-0.012	-0.312	0.254	0.740	0.478
mag2g021_c	14	0.428	-0.022	-0.032	-0.230	-0.206
mag2r151_c	15	0.068	-0.198	0.08	0.230	0.142
mag1v021_sc2g2_c	16	0.028	0.192	-0.066	-0.036	0.022
mag1z071_sc2g2_c	17	0.088	0.204	-0.098	-0.270	-0.180
mag2d101_c	18	0.156	0.070	0.114	0.006	-0.116
mag1g031_sc2g2_c	19	0.014	-0.134	-0.004	0.218	0.214
mag2v041_c	20	0.220	-0.018	-0.006	0.058	0.058
mag2z011_c	21	0.448	0.126	-0.164	-0.376	-0.218
mag1r19s_sc2g2_c	22	-0.412	-0.370	-0.066	-0.130	-0.072
mag2g091_c	23	0.548	0.140	-0.036	-0.086	-0.058
mag2z051_c	24	0.520	0.036	-0.004	-0.078	-0.084
Main effect (model with DIF)		**0.356**	**-0.370**	**0.212**	**0.760**	**0.554**
Main effect (model without DIF)		**0.350**	**-0.358**	**0.212**	**0.766**	**0.556**

There were 5,661 (91.8 %) participants without migration background, 506 (8.2 %) participants with migration background. On average, participants without migration background performed considerably better in the mathematics test than those with migration background (main effect = -0.358 logits, Cohen's *d* = 0.336). Comparing the two groups, DIF

exceeding 0.6 logits occurred in item mag2z061_c with -0.954 logits, indicating that this item was a lot easier for participants without migration background than for people with migration background. Since this is also the most difficult item with an estimate above 2.1 logits, it seems reasonable that due to the small number of participants with migration background, this item showed a large difference in difficulty between theses two groups. The item showed good to very good fit indicators in all other categories. Therefore, there is no reason to exclude this item from the analyses. There were no items with a considerable DIF of 0.4 to 0.6 logits considering migration background.

The number of books at home was used as a proxy for socioeconomic status. There were 1,839 (29.8 %) test takers with 0 to 100 books at home, 3,420 (55.5 %) test takers with more than 100 books at home, and 908 (14.7 %) test takers without any information. Group differences and DIF were investigated by using all three groups. Participants with 100 or fewer books at home performed on average 0.556 logits (Cohen's *d* = -0.553) worse in mathematics than participants with more than 100 books. Comparing the two groups, DIF exceeding |0.4| logits only occurred in item mag1d081_sc2g2_c (0.478 logits).

Furthermore, participants with 100 or fewer books at home performed on average 0.212 logits better than participants without a valid answer (Cohen´s *d* = -0.214), whereas participants with more than 100 books at home performed even 0.766 logits (Cohen´s *d* = -0.760) better that participants with a missing on the book variable. Only item mag 1d081_sc2g2_c (0.740 logits) exceeded |0.4| logits.

In Table 7, we compared the models that only included main effects to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for all three DIF variables, with an exception of the two subgroups of participants with missing and fewer than 100 books at home. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents an overparameterization of models. Using BIC, the more parsimonious models including only the main effects of migration status and number of books at home were preferred over the more complex DIF models. However, for the variable gender, BIC favored the models estimating DIF.

Note that the analyses including the number of books at home 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	172,079.89	30	172,139.89	172,341.69
		DIF	171,504.40	54	171,612.40	171,975.65
Migration status		main effect	172,205.98	30	172,265.98	172,467.79
		DIF	172,134.31	54	172,242.31	172,605.57
Books	<100	main effect	145,323.00	30	145,383.00	145,580.03
	vs. 100	DIF	145,176.88	54	145,284.88	145,639.53
	missing	main effect	119,005.41	30	119,065.41	119,256.60
	vs. 100	DIF	118,858.39	54	118,966.39	119,310.52
	_____	main effect	79,056.00	30	79,116.00	79,293.55

	missing vs. 100	DIF	79,035.54	54	79,143.54	79,463.13

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 (GPCM) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4a), ranging from 0.38 (item mag1d09s_sc2g2_c) to 2.21 (item mag1g031_sc2g2_c). The average discrimination parameter fell at 1.19. Model fit indices suggested a slightly better model fit of the GPCM model (AIC = 173,018.66, BIC = 173,503.00, number of parameters = 72) as compared to the PCM (AIC = 175,380.40, BIC = 175,703.30, number of parameters = 48). Despite the empirical preference for the GPCM model, the PCM 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 (PCM) 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 version 4.5.2, so that we had to use a substitude program 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, where the use of a substitute program was also necessary (see 3.6).

4.3.5 Unidimensionality

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

To estimate this multidimensional model, the Quasi Monte Carlo integration implemented in TAM in R version 3.4.1 (R Core Team, 2016) was used. (Due to convergence problems even with 25 nodes per dimension, model parameters could not be estimated in ConQuest using the Gauss-Hermite quadrature method. This might be caused by the fact that there is more than three dimensions as well as the high correlations between them.)The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained, which occurred at 11,000 nodes.

The variances and correlations of the five dimensions are shown in Table 8. Four out of the five dimensions exhibited a substantial variance. In dimension two (space and shape), the highest estimation falls at 0.253, which means that there ar no difficult items in this dimension. This might explain the rather small variance of 0.586 in dimension two. The correlations between the five dimensions were – as expected – very high, varying between 0.790 and 0.962, and thus, indiciated an essentially unidimensional test (cf. Carstensen, 2013), even though, according to model fit indices, the five-dimensional model fitted the data slightly better (AIC =171,513.2, BIC = 171,802.46, number of parameters = 43) than the unidimensional model (AIC =172,305.6, BIC = 172,500.68, number of parameters = 29). These results indicate that the five content areas measure a common construct, although it is not completely unidimensional.

Model fit between the unidimensional and the five-dimensional model is compared in Table 9.

Table 8: Results of Five-Dimensional Scaling

	Cange and Relationship	Space and shape	Data and chance	Sets, numbers and operations	Units and measurement
Change and relationships (5 items)	1.487				
Space and shape (5 items)	0.916	0.586			
Data and chance (5 items)	0.815	0.888	01.848		
Sets, numbers and operations (5 items)	0.962	0.903	0.790	1.244	
Units and measuring (4 items)	0.950	0.881	0.756	0.953	1.881

Note. Variances of the dimensions are depicted in the diagonal; correlations are given in the off-diagonal.

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

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	172,247.6	29	172,305.6	172,500.68
Five-dimensional	171,427.2	43	171,513.2	171,802.46

Note. Contrary to the calculations for the PCM and GPCM, results in this table were achieved by using TAM in R (see 3.6).

5 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 2 and at describing how the mathematics competence score had been estimated.

The number of different kinds of missing responses was evaluated and the number of all kinds of missing responses was 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. Moreover, discrimination values of the items (either estimated in a GPCM model or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. Only one item showed a considerable DIF for one of the subgroups, but with regard to the small subgroup this was negligible (see 4.3.3). Therefore, the analyses indicated that the test was fair for the examined subgroups.

The test had a good reliability (EAP/PV-reliability = .798, WLE reliability = .787) and distinguished well between test takers, as indicated by the test's variance (1.108). The item

distribution along the ability scale was acceptable, although the test had a tendency to be too easy for the sample.

Fitting a five-dimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model-fit than the unidimensional partial credit model. However, very high correlations of .790 and higher between the five dimensions indicated that the unidimensional model described the data reasonably well.

Summerizing the results, 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 24 items, 22 of which were scored as dichotomous variables (21 MC items and one CMC item for which categories had been collapsed) with 0 indicating an incorrect response and 1 indicating a correct response. Two items were scored as polytomous variables (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 variables were scored as 0.5 for each category. Items that were already administered in grade 1 kept their original names ('mag1v051…', 'mag1d131…', 'mag1d09s…', 'mag1z121…', 'mag1d081…', 'mag1v021…', 'mag1z071…', 'mag1g031…' and 'mag1r19s…'). However, for reasons of identification, a suffix was added in front of the '…_c' (scored item) to specify the current test administration ('sc2g2' referring to Starting Cohort 2, Grade 2). See Fuß et al. (2016) for details on the naming conventions in the NEPS.

6.2 Linking the data of Grade 1 and Grade 2

In starting cohort 2, the mathematics competence tests administered in kindergarten, grade 1, and grade 2 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.Therefore, 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 a 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). The process of linking the mathematics competence in kindergarten and grade 1 is described in Schnittjer and Fischer (2018).

In order to link the tests of mathematics competence conducted in grade 1 and grade 2, nine items which already were administered in grade 1 were, again, administered in grade 2 (e.g., mag1v051_sc2g2_c). An empirical study that evaluated different linking 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. Seven of

the nine common items that were administered in grade 1 and grade 2 were found to be measurement invariant across the two measurement points. As such, they served as link items. Therefore, the anchor items design as described in Fischer et al. (2016) was used. For more information on the selection of link items and the method for linking the tests of mathematical competence see Fischer et al. (2016).

6.2.1 Samples

In starting cohort 2, a subsample of 5,813 students participated at both measurement occasions, in grade 1 and also in grade 2. Consequently, these respondants were used to link the two tests across the three measurement points (see Fischer et al., 2016).

6.2.2 Results

To examine whether the two tests administered in the longitudinal 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. For the two-dimensional model, the common items load on the first dimension and the unique items (i.e., the items included in only one test) load on the second dimension. Because in both grades the information criteria favored the one-dimensional model, AIC = 147,619 and BIC = 147,786 for grade 1, and AIC = 162,435 and BIC = 162,657 for grade 2, over the two-dimensional model, AIC =180,305 and BIC =180,497 for Grade 1, and AIC = 195,121 and BIC = 195,327 for grade 2, the unidimensional scales can be assumed for the mathematics tests in grades 1 and 2.

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 2 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 10.

The analyses of differential item functioning identified two items with significant DIF (mag1d09s_c / mag1d09s_sc2g2_c and mag1g031_c / mag1g031_sc2g2_c). Therefore, those items were excluded as anchor item. The mathematics competence tests administered in the two grades were linked using the “mean/mean” method using the seven measurement invariant anchor items (see Fischer et al., 2016).

Table 10: Differential Item Functioning Analysis for the common items in the tests for mathematical competence in Grade 1 and Grade 2.

Grade 1	Grade 2	Δσ	$SE_{\Delta\sigma}$	*F*
mag1v051_c	mag1v051_sc2g2_c	0.058	0.04	1.70
mag1d131_c	mag1d131_sc2g2_c	-0.446	0.04	110.45
mag1d09s_c	mag1d09s_sc2g2_c	-0.672	0.04	259.30
mag1z121_c	mag1z121_sc2g2_c	-0.075	0.06	1.77
mag1d081_c	mag1d081_sc2g2_c	0.479	0.05	97.93
mag1v021_c	mag1v021_sc2g2_c	-0.055	0.04	1.62
mag1z071_c	mag1z071_sc2g2_c	0.321	0.04	55.41
mag1g031_c	mag1g031_sc2g2_c	0.539	0.05	141.36
mag1r19s_c	mag1r19s_sc2g2_c	-0.150	0.05	10.32

Note. Δσ = Difference in item difficulty parameters between Grades 1 and 2 (positive values indicate easier items in Grade 1); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis; F_{crit} = Critical value for the minimum effects hypothesis test for an α of .05; the degrees of freedom (df_1, df_2) are based on the number of measurement points (df_1 = *k*-1) and the number of test takers taking both tests (df_2 = n-1). The critical *F* (1; 5,812) = 125.18. A non-significant test indicates measurement invariance.

In the longitudinal subsample, the mean item difficulty parameters for the seven anchor items (see Table 10) were 0.386 in Grade 1 and -0.488 in Grade 2. Mean/mean linking (Loyd & Hoover, 1980) resulted in a correction term of $c_{1\text{-}2}$ = 0.386 – (-0.488) = 0.874. The correction term for linking kindergarten to 1st grade was $c_{kindergarten\text{-}1}$ = 1.352 (Schnittjer & Fischer, 2018). The sum of the correction terms $c_{kindergarten\text{-}1}$+ $c_{1\text{-}2}$= 2.226 was added to each item difficulty parameter derived in 2nd grade and, thus, resulted in the linked item parameters (see Table 11).

Table 11: Original and linked item difficulty parameters for the mathematics test in 2nd Grade.

item	Anchor item	Original Item difficulties	Linked Item difficulties
mag1v051_sc2g2_c	yes	-1.250	0.976
mag2v071_c	no	-0.779	1.447
mag2r031_c	no	-2.007	0.219
mag2d061_c	no	-0.584	1.642
mag1d131_sc2g2_c	yes	-0.267	1.959
mag2r131_c	no	0.253	2.479
mag2v121_c	no	-0.633	1.593
mag2q061_c	no	2.054	4.280
mag2r111_c	no	-0.056	2.170
mag1d09s_sc2g2_c	no	-0.549	1.677
mag1z121_sc2g2_c	yes	1.609	3.835
mag2g12s_c	no	-1.483	0.743
mag1d081_sc2g2_c	yes	-1.934	0.292
mag2g021_c	no	0.512	2.738
mag2r151_c	no	-0.741	1.485
mag1v021_sc2g2_c	yes	-0.199	2.027
mag1z071_sc2g2_c	yes	-0.449	1.777
mag2d101_c	no	-1.614	0.612
mag1g031_sc2g2_c	no	-1.383	0.843
mag2v041_c	no	-0.718	1.508
mag2q011_c	no	-0.288	1.938
mag1r19s_sc2g2_c	yes	-0.829	1.397
mag2g091_c	no	-0.603	1.623
mag2q051_c	no	-1.828	0.398

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (section 4.2). Linked item difficulty parameters were derived by adding $c_{kindergarten\text{-}2}$ to the original item parameters.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs, “mag2_sc1” and “mag2_sc1u”, including their respective standard errors, “mag2_sc2” and “mag2_sc2u”. For “mag2_sc1u”, person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for differences in the test position, that varied in grade 1. There the mathematics test was either presented on first position of the test battery, or presented on second position. Whereas the

2nd grade mathematics test was always presented on first positon. To correct for differences in test position, half of the main effect related to the test position was added to the WLE scores of respondants that received the mathematics test in 1st grade on second position. As a result the WLE scores provided in “mag2_sc1u” can be used for longitudinal comparisons between kindergarten, grade 1, and grade 2.

The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “mag2_sc1” were not linked to the underlying reference scale of kindergarten. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A, the fixed item parameters for estimating the uncorrected WLE scores 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.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. 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.

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.

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.

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.

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.

Kiefer T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 1.995-0).

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

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., 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.

R Core Team (2016). R: A language and environment for statistical computing (Version 3.2.4) [Software]. Retrieved from https://www.R-project.org/.

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

Schnittjer, I., & Fischer, L. (2018). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1* (NEPS Working Paper No. 47). Bamberg: Leibniz Instutite for Educational Trajectories, National Educational Panel Study.

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

Van den Ham, A.-K. (2016). *Ein Validitätsargument für den Mathematiktest der National Educational Panel Study für die neunte Klassenstufe.* Unpublished doctoral dissertation, Leuphana University Lüneburg, Lüneburg.

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 II

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

data filename.dat;
format pid 4-10 responses 12-35; /* insert number of columns with data*/
labels << labels.nam;

codes 0,1,2,3,4,5;

recode (0,1,2,3,4)   (0,0,1,2,3)   !item (10); /* collapsing the lowest categories * /
recode (0,1,2,3,4)   (0,0,1,2,3)   !item (12); /* collapsing the lowest categories * /
recode (0,1,2,3,4,5) (0,0,0,0,0,1) !item (22); /* collapsing the lowest categories * /

score (0,1)          (0,1)         !item (1-9,11,13-24);
score (0,1,2,3)      (0,0.5,1,1.5) !item (10,12);

set constraint=cases;

model item + item*step;
estimate;

show !estimates=latent >> filename.shw;
itanal >> filename.ita;
show cases !estimates=wle >> filename.wle;
```

Appendix B: Fixed Item Parameters

```
1       0.976  /*mag1v051_sc2g2_c*/
2       1.447  /*mag2v071_c*/
3       0.219  /*mag2r031_c*/
4       1.642  /*mag2d061_c*/
5       1.959  /*mag1d131_sc2g2_c*/
6       2.479  /*mag2r131_c*/
7       1.593  /*mag2v121_c*/
8       4.280  /*mag2q061_c*/
9       2.170  /*mag2r111_c*/
10      1.677  /*mag1d09s_sc2g2_c*/
11      3.835  /*mag1z121_sc2g2_c*/
12      0.743  /*mag2g12s_c*/
13      0.292  /*mag1d081_sc2g2_c*/
14      2.738  /*mag2g021_c*/
15      1.485  /*mag2r151_c*/
16      2.027  /*mag1v021_sc2g2_c*/
17      1.777  /*mag1z071_sc2g2_c*/
18      0.612  /*mag2d101_c*/
19      0.843  /*mag1g031_sc2g2_c*/
20      1.508  /*mag2v041_c*/
21      1.938  /*mag2q011_c*/
22      1.397  /*mag1r19s_sc2g2_c*/
23      1.623  /*mag2g091_c*/
24      0.398  /*mag2q051_c*/
25       0.0035        /* correcting for test position - first position in grade 1*/
```

Appendix C: Content Areas of Items in the Mathematics Test for Grade 2

Position	Item	Content area
1	mag1v051_sc2g2_c	Change and relationships
2	mag2v071_c	Change and relationships
3	mag2r031_c	Space and shape
4	mag2d061_c	Data and chance
5	mag1d131_sc2g2_c	Data and chance
6	mag2r131_c	Space and shape
7	mag2v121_c	Change and relationships
8	mag2z061_c	Sets, numbers, and operations
9	mag2r111_c	Space and shape
10	mag1d09s_sc2g2_c	Data and chance
11	mag1z121_sc2g2_c	Sets, numbers, and operations
12	mag2g12s_c	Units and measuring
13	mag1d081_sc2g2_c	Data and chance
14	mag2g021_c	Units and measuring
15	mag2r151_c	Space and shape
16	mag1v021_sc2g2_c	Change and relationships
17	mag1z071_sc2g2_c	Sets, numbers, and operations
18	mag2d101_c	Data and chance
19	mag1g031_sc2g2_c	Units and measuring
20	mag2v041_c	Change and relationships
21	mag2z011_c	Sets, numbers, and operations
22	mag1r19s_sc2g2_c	Space and shape
23	mag2g091_c	Units and measuring
24	mag2z051_c	Sets, numbers, and operations

Note. Up to now, the internal validity of the individual dimensions of mathematical competence as dependent measures has not yet been confirmed (van den Ham, 2016).