NEPS SURVEY PAPERS

Insa Schnittjer, Anna-Lena Gerken and Lara Aylin Petersen

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 2 IN GRADE 4

NEPS Survey Paper No. 69
Bamberg, April 2020

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 in Grade 4

Insa Schnittjer, Anna-Lena Gerken, and Lara Aylin Petersen

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

Email address of the lead author:

schnittjer@uni-landau.de

Bibliographic Data:

Schnittjer, I., Gerken, A.-L., & Petersen, L. A. (2020). *NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 2 in Fourth Grade* (NEPS Survey Paper No. 69). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP69:1.0

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 in Grade 4

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 4 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 6,725 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 relatively high omission rates in the short constructed response (SCR) items as well as some recognizable gaps in the scale of item difficulties. 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 as well as the longnitudinal linking parapeters.

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) as well as Fuß, Gnambs, Lockl, and Attig (2019).

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 4 of starting cohort 2 (kindergarten). First, the main concepts of the mathematical test are introduced. Then, 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:

- 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 were three types of response formats. These were simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR).

In MC items the test taker had to find the correct answer from four response options. In CMC tasks a number of subtasks with two response options were presented. SCR tasks required 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. As in the previous studies of this starting cohort, the competence tests for these domains were always presented in the same order. 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 4 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 four content areas (see Appendix C for the assignment of the items to the content areas), whereas Table 2 shows the distribution of the response formats. Two of the nine SCR items included either two or three subtasks; however, both items were scored dichotomously as the other SCR Items. The CMC item included five subtasks. Because of insufficient cell frequencies (see below), some categories had to be collapsed, and, therefore the CMC item was scored dichotomously, as well.

Table 1: Number of Items by Content Areas

Content area	Frequency
Quantity	9
Space and shape	5
Change and relationships	4
Data and chance	6
Total number of items	24

Table 2: Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	14
Complex Multiple-Choice	1
Short Constructed Response	9
Total number of items	24

Due to the transition from kindergarten to elementary school, not all students in this starting cohort could be assessed in the school context. So the starting cohort was divided into two subsamples that exhibited different assessment settings: Students that remained in the school context were tested at school in a group setting; the remaining students were tracked and individually tested at home. Thus, the context of the test administration differed between the two groups.

3.2 Sample

A total of 6,724 students received the mathematics test. For six respondents less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few responses, this cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 6,718 test takers. Of these, 5,272 students were tested in the class context, whereas 1,446 students were tested individually. 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 were 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. Because of time limits, not all persons finished the test within the given time. All missing responses after the last valid response were coded as not reached. As partial credit 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. The polytomous items were 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 Rasch model (Rasch 1960). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

The CMC item consisted of a set of five subtasks that were aggregated to a polytomous variable, 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 mag4d14s_c categories had to be collapsed resulting in a dichotomous score and, therefore, was scored 1 for only correct responses over all subtasks and 0 for one or more incorrect response over the subtasks. Therefore there were no polytomously scored items in the test.

Simple MC items and SCR items were also 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 partial credit 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 and the SCR 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 variables that were included in the final scaling model. Due to unsatisfactory numbers in categories, these categroeis had to be collapsed and therefore the only CMC item had to be scored dichotomously, too.

The MC items consisted of one correct response option and two to four 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).

The SCR items require the test-taker to give mostly one-word answers, such as number. All SCR items were scored dichotomous even if there was more than a one response required.

After aggregating the subtasks to polytomous variables, and scoring them all dichotomously, the fit of the dichotomous MC and SCR items, and the finally dichotomous CMC item to the Rasch model (Rasch, 1960) 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, 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). Moreover, in light of the two different assessment settings, invariance analyses were also conducted for the administration setting. 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 usually scaled assuming Rasch-homogeneity. The Rasch (1960) model 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 two-parametric logistic model (2PL; Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the mathematics 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, TAM in R was used. To guarantee the capability with the multidimensional model, the unidimensional model was estimated in TAM too. The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (9,000 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 4.2.5 (Wu, Adams, & Wilson, 2015). The 2PL model was estimated in mdltm (Matthias von Davier, 2005). The multi-dimensional-model was estimated in TAM version 2.8-21 (Kiefer, Robitzsch, & Wu, 2017) in R version 3.4.2 (R Core Team, 2017).

4 Results

4.1 Missing Responses

4.1.1 Missing responses per person

The number of invalid responses per person was rather small, as can been seen in Figure 1. In fact, 84.2 % of test takers gave no invalid response at all. Only 2.3 % of the respondents had more than one invalid response.

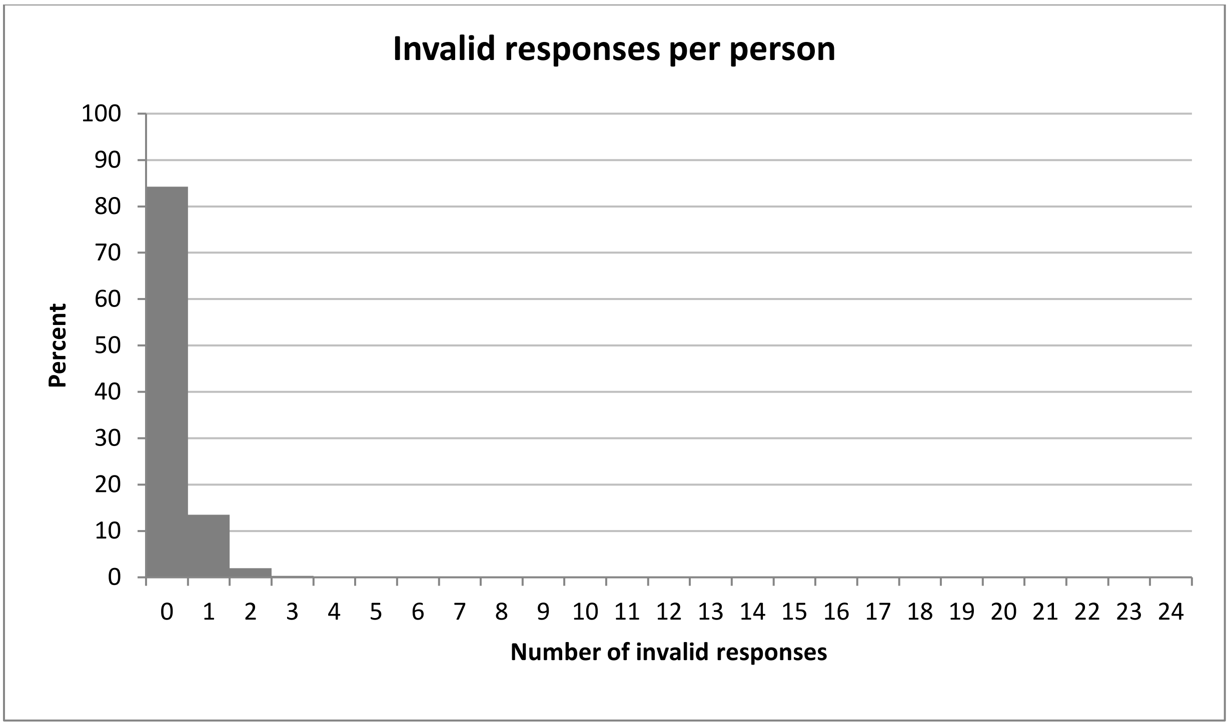

Figure 1: Number of invalid responses

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 45.0 % of the respondents omitted no item at all, whereas 5.1 % of the respondents omitted more than 5 items.

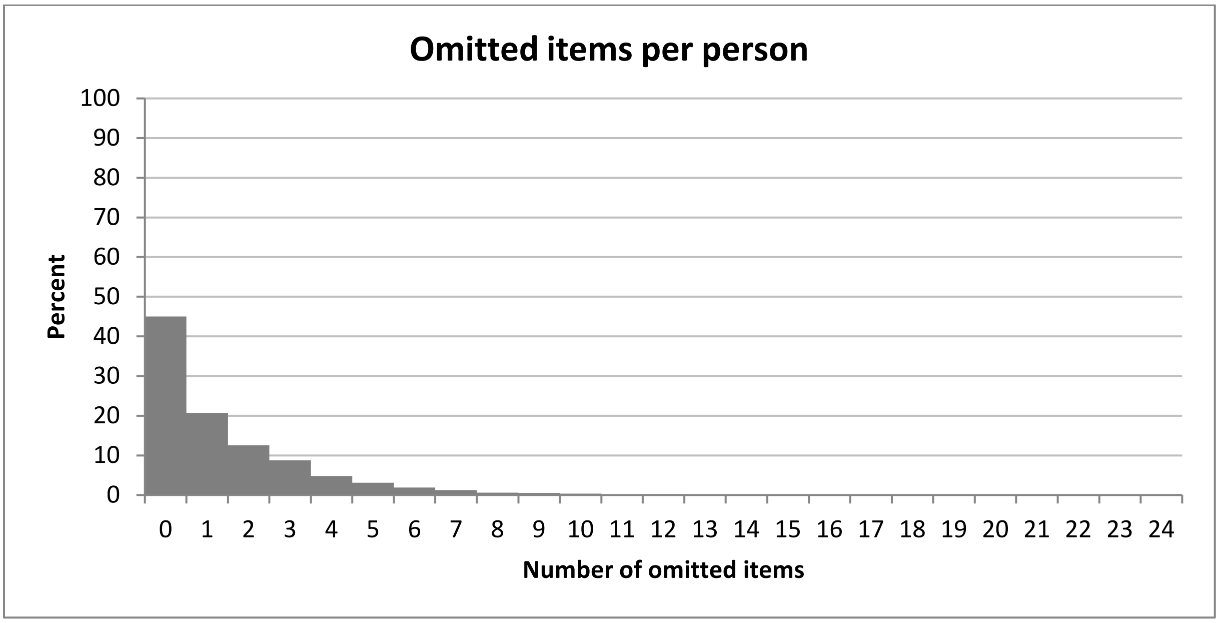

Figure 2: Number of omitted items

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, 63.3 % reached the end of the test, whereas 19.3 % of the test takers had one to five items not reached. 17.4 % of the students had more than five items not reached.

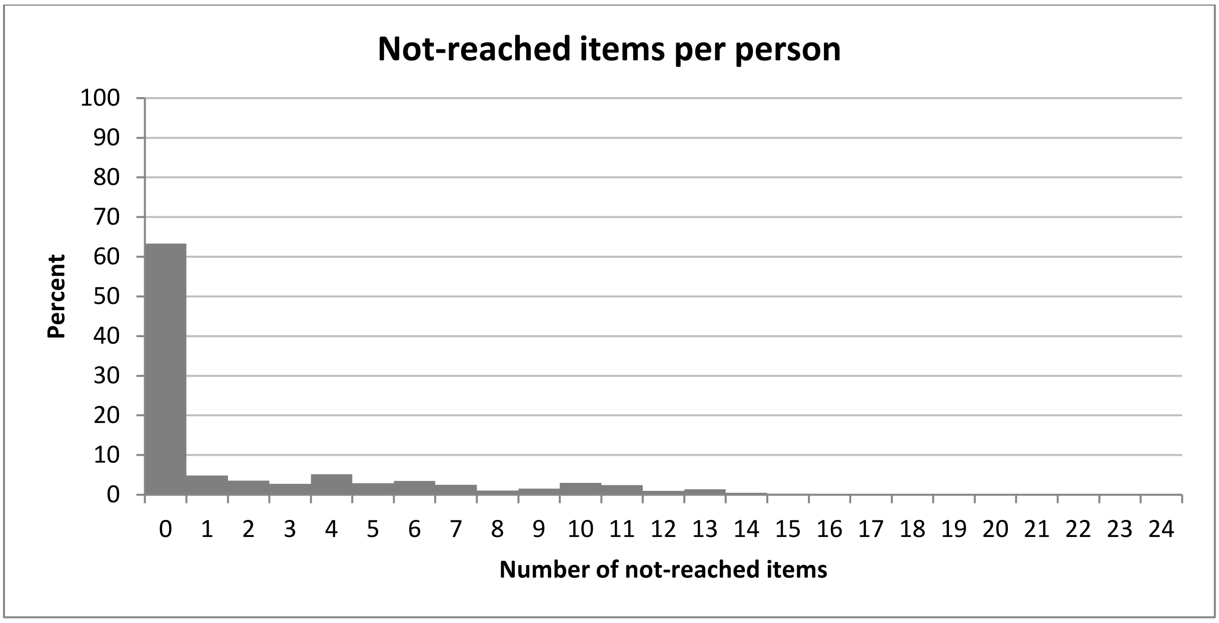

Figure 3: Number of not-reached items

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, 27.9 % of the test takers showed no missing response at all, whereas 28.5 % showed more than five missing responses.

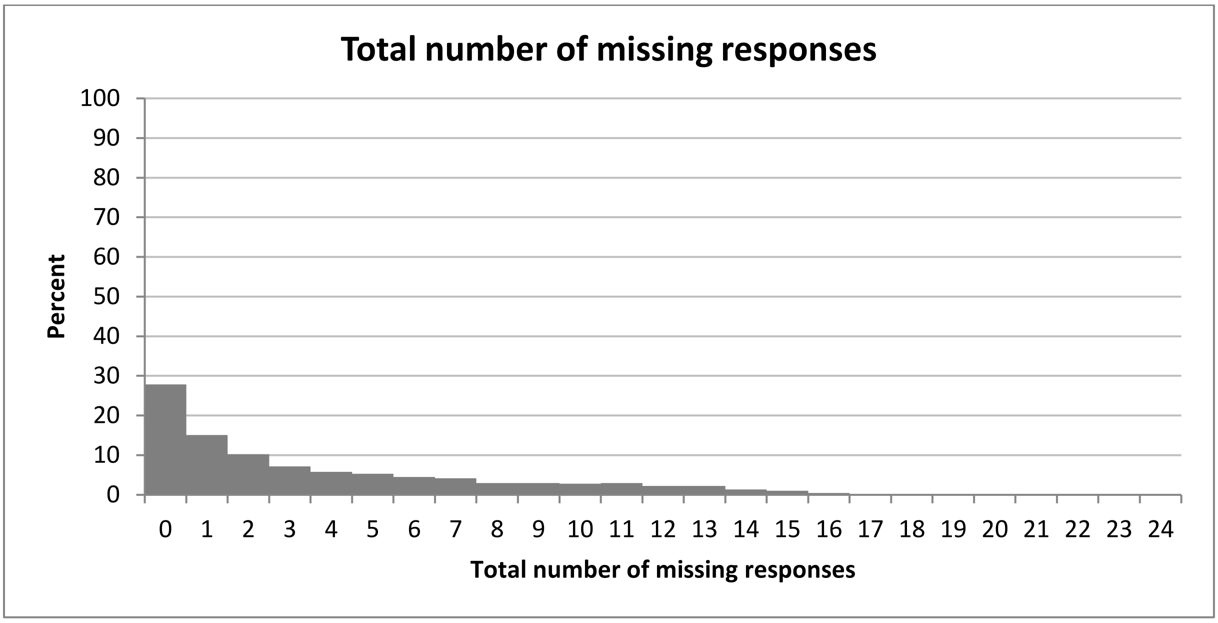

Figure 4: Total number of missing responses

Overall, there was a rather high amount of missing. The Short Constructed Response Format seemed to be hard to handle for the students. As can be seen in Figure 5, there was a higher amount of missings for Short Constructed Response Formats than for the Multiple-Choice or Complex Multiple-Choice Formats (except for the last items, which always show more missing than the rest, due to time limitations).

4.1.2 Missing responses per item

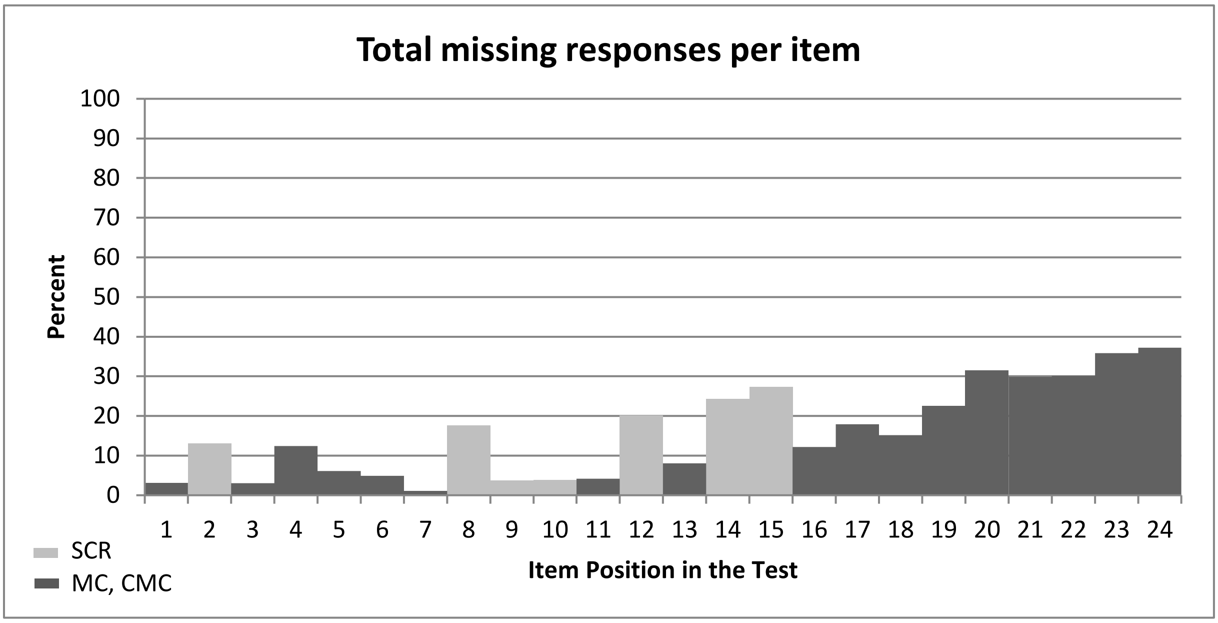

Figure 5: Total missing responses per item

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

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 12.02 % (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
mag5d041_sc2g4_c	1	6,510	0.18	2.92	0.00
mag4q101_c	2	5,837	1.10	12.01	0.00
mag4r021_c	3	6,513	0.03	3.02	0.00
mag5v271_sc2g4_c	4	5,882	0.03	12.40	0.01
mag4q011_c	5	6,309	0.24	5.84	0.01
mag4r071_c	6	6,387	1.19	3.69	0.04
mag4d131_c	7	6,645	0.03	0.98	0.07
mag5q231_sc2g4_c	8	5,532	0.98	16.45	0.22
mag5q301_sc2g4_c	9	6,467	0.30	3.14	0.30
mag4v121_c	10	6,463	0.18	3.10	0.52
mag5d051_sc2g4_c	11	6,437	0.06	3.14	0.98
mag4q060_c	12	5,368	6.04	11.55	2.40
mag4d031_c	13	6,177	0.16	4.53	3.36
mag5q140_sc2g4_c	14	5,086	2.71	13.56	5.82
mag4v111_c	15	4,882	1.00	17.52	8.81
mag4r041_c	16	5,898	0.04	1.77	10.39
mag4r042_c	17	5,514	0.07	6.43	11.42
mag4q051_c	18	5,700	0.07	1.18	13.90
mag4q091_c	19	5,202	0.43	4.76	17.37
mag4q092_c	20	4,600	0.58	10.64	20.30
mag4d14s_c	21	4,714	0.06	4.02	25.50
mag5v071_sc2g4_c	22	4,694	0.21	1.65	28.27
mag5r191_sc2g4_c	23	4,307	2.34	1.73	31.82
mag4d081_c	24	4,220	0.51	0.00	36.68

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 the CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 9.87 % and 92.07 % across all items. On average, the rate of correct responses was 53.34 % (*SD* = 27.80 %). From a descriptive point of view, the items covered a relatively wide range of difficulties.

The estimated item difficulties are depicted in Table 4. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. Due to insufficient frequencies in categories, categories of the CMC item had to be collapsed, and therefore this item was scored dichotomously. The estimated item difficulties varied between -2.839 (mag4d131_c) and 2.630 (mag4q060_c) with a mean of -0.037. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were very small (*SE*(ß) ≤ 0.05).

Table 4: Item Parameters

Item	Position	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.
mag5d041_sc2g4_c	1	53.99	-0.268	0.030	1.02	1.5	0.33	0.80
mag4q101_c	2	11.00	2.291	0.044	0.99	-0.3	0.28	0.97
mag4r021_c	3	38.42	0.507	0.030	1.07	6.4	0.23	0.56
mag5v271_sc2g4_c	4	23.07	1.179	0.034	1.04	2.5	0.25	0.63
mag4q011_c	5	14.89	1.955	0.039	1.04	1.6	0.19	0.59
mag4r071_c	6	28.71	0.998	0.032	1.05	3.5	0.25	0.62
mag4d131_c	7	91.07	-2.839	0.049	0.94	-1.8	0.29	1.37
mag5q231_sc2g4_c	8	22.55	1.171	0.035	0.97	-2.1	0.35	0.99
mag5q301_sc2g4_c	9	28.62	1.051	0.032	0.93	-5.3	0.40	1.27
mag4v121_c	10	49.79	-0.068	0.030	1.03	2.9	0.31	0.72
mag5d051_sc2g4_c	11	75.90	-1.582	0.036	0.92	-4.6	0.40	1.39
mag4q060_c	12	7.89	2.630	0.050	1.04	1.1	0.16	0.62
mag4d031_c	13	35.80	0.547	0.031	1.07	6.2	0.25	0.56
mag5q140_sc2g4_c	14	38.94	-0.026	0.033	0.89	-10.7	0.50	1.56
mag4v111_c	15	9.57	2.135	0.047	0.99	-0.4	0.29	1.00
mag4r041_c	16	79.44	-2.624	0.049	0.95	-1.5	0.28	1.16
mag4r042_c	17	53.96	-0.783	0.034	1.07	4.9	0.27	0.63
mag4q051_c	18	62.43	-1.226	0.035	0.93	-4.1	0.41	1.21
mag4q091_c	19	52.75	-0.881	0.035	1.00	0.0	0.34	0.84
mag4q092_c	20	34.67	0.001	0.035	0.99	-0.7	0.37	0.88
mag4d14s_c	21	n.a.	-1.723	0.036	0.96	-2.7	0.40	1.03
mag5v071_sc2g4_c	22	56.52	-1.723	0.042	1.02	0.7	0.30	0.84
mag5r191_sc2g4_c	23	23.03	0.686	0.037	1.03	2.0	0.31	0.71
mag4d081_c	24	48.85	-1.510	0.042	0.99	-0.3	0.33	0.92

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 (2PL).

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).

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 6, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The respective thresholds ranged from -2.840 (item mag4d131_c) to 2.632 (item mag4q06s_c). Therefore, a rather borad range was spanned. The variance was estimated to be 0.990, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = 0.749, WLE reliability = 0.727) was good. Although the items covered a wide range of the ability distribution, the range of item difficulties showed some larger gaps on the upper and the lower ends of the scale. As a consequence, person abilities in medium regions and mostly in lower regions were measured relatively precisely, whereas very high and very low ability estimates had larger standard errors.

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

Figure 6: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each ‘X’ represents 36.7 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4).

4.3 Quality of the test

Since the items of the mathematical competence test refer to many different stimuli, the assumption of local item independence is plausible.

4.3.1 Fit of the subtasks of complex multiple-choice items

Before the responses to the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks had been checked by analyzing the subtasks together with the simple multiple-choice items via a simple Rasch model. There were 27 variables altogether.

The rates of correct responses given to the subtasks of the CMC item varied from 83.64% to 89.21%. The subtasks ranged between good and very good item fit – WMNSQ ranging between 0.86 and 1.01 and the respective t-values between -4.5 and 0.6. The good model fit of the subtasks was considered to justify their aggregation to a polytomous variable for item (mag4d14s_c). As already described in section 3.1, this only CMC item (mag4d14s_c) was scored dichotomously due to insufficient frequencies in categories.

4.3.2 Distractor analyses

To investigate how well the distractors performed in the test, for the MC items the point-biserial correlations between selecting each incorrect response (distractor) 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 -0.60 to 0.02 with a mean of -0.23. Although one distractor showed a correlation slightly above 0, 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 0.25 to 0.60 with a mean of 0.41 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	0.41	-0.23
Minimum	0.25	-0.60
Maximum	0.60	0.02

4.3.3 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the Rasch model, for all variables were scored dichotomously. 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 0.89 (mag5q14s_sc2g4_c) and the highest being 1.07 (mag4r021_c, mag4d031_c and mag4r042_c). 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 0.16 (mag4q06s_c) and 0.50 (mag5q14s_sc2g4_c) with an average correlation of 0.31. Thus, the value 0.16 was caused by the most difficult item of test, its correlation is still within acceptance and especially since it was an SCR item that asked for three short answers. Also taking into account the fact of large gaps on the ends of the scale, its depicted correlation is acceptable.

4.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home, migration background, and testing mode (group testing vs. individual testing) (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 boys and girls, *ß*(male) – *ß*(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females.

Overall, 3,438 (51.2 %) of the test takers were female and 3,27924 (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 = 0.108 logits, Cohen's *d* = 0.109). All items of the test were fair regarding gender.

There were 4,205 (62.6 %) participants without migration background, 1,471 (21.9 %) participants with migration background, and 1,042 (15.5 %) students that gave no valid answer and therefore were grouped to missings. On average, participants without migration background performed considerably better in the mathematics test than those with migration background (main effect = 0.342 logits, Cohen's *d* = 0.354). Whereas those with migration backgroud performed slightly better than those without a valid respond (main effect = 0.096, Cohen's *d* = 0.096). Therefore, the group of children without migration background also performed considerably better than the group of children that gave no valid respond (main effect = 0.442, Cohen's *d* = 0.453).

Comparing the three groups, DIF exceeding 0.4 logits occurred in item mag4d131_c with 0.422 logits, indicating that this item was a little easier for participants without migration background than for people with migration background. This item was not only the most easiest on the test, but also showed no DIF for the other two subgroups of migration backgroud and therefore the DIF was considered very small. DIF also slightly exceeded 0.4 logits for item mag5q140_sc2g4_c, indicating that this item also was slightly easier for children that gave no valid resond than for those with migration background. Since this item was insignificant for the other two subgroup of migration background and additionally the difference was very small, this DIF was considered not severe. There were no items with a considerable DIF of 0.4 to 0.6 logits or even above 0.6 logits considering migration background.

Table 6: Differential Item Functioning

Item	Position	Gender	Migration status			Books			Mode
		female vs.	withou t vs. with	without vs. miss.	with vs. miss.	<100 vs. >100	<100 vs. miss.	>100 vs. miss.	group vs. indivi-

		male							dual
mag5d041_sc2g4_c	1	0.214	-0.158	0.042	0.198	0.138	0.100	-0.058	-0.158
mag4q101_c	2	-0.120	0.110	-0.004	-0.116	-0.114	0.154	0.248	-0.192
mag4r021_c	3	-0.202	0.184	0.188	0.002	-0.280	0.164	0.422	0.016
mag5v271_sc2g4_c	4	-0.242	0.192	0.252	0.060	-0.170	0.150	0.298	0.210
mag4q011_c	5	-0.108	0.218	0.254	0.036	-0.100	0.016	0.096	-0.128
mag4r071_c	6	0.208	-0.130	-0.064	0.064	-0.092	0.118	0.188	0.386
mag4d131_c	7	0.314	-0.422	-0.138	0.284	0.542	-0.118	-0.680	-0.458
mag5q231_sc2g4_c	8	-0.040	-0.066	-0.136	-0.072	0.038	0.046	-0.012	0.006
mag5q301_sc2g4_c	9	0.130	-0.066	-0.078	-0.014	0.154	-0.032	-0.206	0.052
mag4v121_c	10	-0.070	-0.110	0.026	0.138	-0.080	0.064	0.124	-0.026
mag5d051_sc2g4_c	11	-0.252	-0.010	-0.046	-0.038	0.058	-0.264	-0.342	-0.164
mag4q060_c	12	-0.124	-0.096	0.068	0.164	0.094	0.050	-0.064	0.316
mag4d031_c	13	-0.014	0.280	0.330	0.046	-0.162	0.142	0.282	-0.044
mag5q140_sc2g4_c	14	0.090	0.078	-0.340	-0.420	0.204	-0.134	-0.360	-0.092
mag4v111_c	15	0.212	0.090	-0.020	-0.114	0.144	0.194	0.028	0.204
mag4r041_c	16	0.366	-0.208	-0.336	-0.128	0.326	-0.236	-0.582	0.024
mag4r042_c	17	0.076	0.146	0.114	-0.034	-0.100	0.166	0.244	0.090
mag4q051_c	18	-0.324	-0.120	-0.184	-0.066	0.234	-0.034	-0.288	0.064
mag4q091_c	19	0.268	0.120	-0.206	-0.328	-0.160	-0.158	-0.018	-0.080
mag4q092_c	20	-0.220	0.008	0.042	0.032	-0.164	-0.168	-0.024	0.080
mag4d14s_c	21	-0.016	-0.388	-0.218	0.170	0.270	-0.092	-0.382	-0.170
mag5v071_sc2g4_c	22	-0.082	0.076	0.064	-0.012	-0.190	-0.274	-0.104	0.150
mag5r191_sc2g4_c	23	0.014	0.068	0.078	0.008	-0.084	0.120	0.184	0.010
mag4d081_c	24	0.218	-0.188	-0.064	0.122	0.248	-0.044	-0.314	-0.206
Main effect (model with DIF)		**-0.108**	**-0.336**	**-0.436**	**-0.098**	**0.614**	**-0.128**	**-0.722**	**0.00**
Main effect (model without DIF)		**-0.108**	**-0.342**	**-0.442**	**-0.096**	**0.620**	**-0.128**	**-0.750**	**-0.002**

The number of books at home was used as a proxy for socioeconomic status. There were 2,092 (31.1 %) test takers with 0 to 100 books at home, 3,678 (54.7 %) test takers with more than 100 books at home, and 948 (14.1 %) test takers without any information. Participants with 100 or less books at home performed on average 0.620 logits (Cohen's *d* = -0.666) worse on this mathematics test than participants with more than 100 books. Still, they performed on average 0.128 logits better than test takers that gave no information on this matter. Persons with more than 100 books at home performed even 0.750 logits better than persons without any information about their number of books at home. Comparing the three groups, DIF exceeding |0.4| logits only occurred in three items. In item mag4r021_c a difference of 0.422 logits for the difference between persons with more than 100 books and persons with missings occured, in mag4r041_c the same two subgroups showed a differenze of -0.582 logits. Finally item mag4d131_c showed a difference of 0.542 logits for the group of test takres with more that 100 books at home versus the group of test takers with less than 100 books at home. This item also was the only item exceeding a group differnce of 0.6 logits. This difference of 0.680 logits appeared comparing the group of test takers with more than 100 books to the group that gave no information on this matter. However, since all three items showed good item fit in the other categories, and taking into account that the item exceeding 0.6 logits, being the most easiest item on the mathematics test, there was no evidence to be found for excluding an item.

As described in chapters 3.1 and 3.2, there were 5,272 test takers that were tested within class context and 1,446 test takers that were tested individually at home. Comparing the two testing groups, there was barely any difference in test difficulty for those two goups (0.002 logits). Only the most easiest item (mag4d131_c) of the test was slightly easier for test takers that were tested individually, indicated by exceeding |0.4| logits slightly, by achieving a difference of 0.458 logits.

Overall testfairness could be conformed for all tested subgorups. In Table 7, we compared the models that only included main effects to models that additionally estimated DIF effects. Overall, Akaike's (1974) information criterion (AIC) favored the models estimating DIF for all four DIF variables. However, only two comparisons of subgroups favored models estimating only the main effect. This appeared for the comparison of the group with migration status to the group that gave no information on the migration status, as well as the comparison of the subgroups of test takers with less than 100 books at home versus test takers that gave no information on the amount of 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, amount of books at home and of mode were preferred over the more complex DIF models. However, for the variable gender, more complex models estimating DIF were preferred.

Table 7: Comparison of Models with and without DIF

DIF variable		Model	Deviance	Number of parameters	AIC	BIC
Gender		main effect	172,079.89	26	172,139.89	172,341.69
		DIF	171,504.40	50	171,612.40	171,975.65
Migration status	without vs. with	main effect	117,769.25	26	117,821.25	117,994.00
		DIF	117,668.41	50	117,768.41	118,100.61
	without vs. missing	main effect	108,784.90	26	108,836.90	109,007.60
		DIF	108,706.44	50	108,806.44	109,134.71
	with vs. missing	main effect	51,949.25	26	52,001.25	52,152.81
		DIF	51,904.05	50	52,004.05	52,295.51
Books	<100 vs. >100	main effect	119,090.95	26	119,142.95	119,316.12
		DIF	118,949.97	50	119,049.97	119,382.99
	<100 vs. missing	main effect	62,850.26	26	62,902.26	63,058.77
		DIF	62,808.83	50	62,908.83	63,209.81
	>100 vs. missing	main effect	95,717.16	26	95,769.16	95,936.58
		DIF	95,544.53	50	95,644.53	95,966.50
Mode		Main	139,484.31	26	139,536.31	139,713.44
		DIF	139,390.23	50	139,490.23	139,830.85

Note. The analyses including the amount of books at home contain fewer cases and, thus, the information criteria cannot be compared across analyses with different DIF variables

4.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. In order to test this assumption, a two-parametric logistic model (2PL) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4a), ranging from 0.56 (item mag4r021_c) to 1.56 (item mag5q140_sc2g4_c). The average discrimination parameter fell at 0.91. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 138,597.60, BIC = 139,019.98, number of parameters = 62) as compared to the 1PL model (AIC = 139,535.78, BIC = 139,794.65, number of parameters = 38). 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 1PL model 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 4.2.5 so that we had to MDLTM (see 3.6, Davier, 2005). As a consequence, the results for AIC and BIC using the 1PL model might differ slightly from the later results (see 4.3.5) comparing multi-dimensionality to unidimensionality of the test, estimated in R (see 3.6).

4.3.6 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 Quasi Monte Carlo estimation implemented in R in the package “TAM” was used. The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained, which occurred at 9,000 nodes.

The variances and correlations of the four dimensions are shown in Table 8. Three of the four dimensions exhibitt a substantial variance. In dimension 3 (change and relationship) the item difficulties ranged on the upper and lower end of the scale, which means there was a lack of items with an average difficulty on this dimension. Furthermore there were two very difficult items with only one conterplay on the lower end of the scale. However, point-biserial correlations of these items were rather low, due to their extends on the estimation scale. This might explain the rather small variance of 0.676 in dimension three. The correlations between the four dimensions were – as expected – very high, varying between 0.856 and 0.949, and, thus, indiciated an essentially unidimensional test (cf. Carstensen, 2013). Even though, according to model fit indices, the four-dimensional model fitted the data slightly better (AIC = 139,298.20, BIC = 139,529.83, number of parameters = 34) than the unidimensional model (AIC =139,534.30, BIC = 139,704.61, number of parameters = 25). However, these results indicate that the four 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 Four-Dimensional Scaling

	Quantity	**Space and shape**	**Cange and Relationship**	**Data and chance**
Quantity (5 items)	1.376			
Space and shape (5 items)	0.949	1.195		
Change and relationships (7 items)	0.932	0.917	0.676	
Data and chance (6 items)	0.924	0.898	0.856	0.933

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	139,484.3	25	139,534.30	139,704.61
Four-dimensional	139,230.2	34	139,298.20	139,529.83

Note. Contrary to the calculations for the 1PL and 2PL models, 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 amount of different kinds of missing responses was evaluated and most kinds of missing responses were rather low. However, looking more closely to the response formats, the number of missings was quite large for items using the Short Construced Response format (SCR). Therefore, the SCR format seemed be hard to handle with for the students on 4th grade. 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 2PL model or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. Ignoring the subgroup of test takers that gave no information on the DIF variables, only the most easiest item of the test showed a considerable, but not sincere DIF (see 4.3.4). Therefore, the analyses indicated that the test was 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 range of item difficulties showed some larger gaps on the upper and the lower ends of the scale. As a consequence, person abilities in medium regions and mostly in lower regions were measured relatively precisely, whereas very high and very low ability estimates had larger standard errors.

Fitting a four-dimensional Rasch model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model than the unidimensional model. Nevertheless, 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 24 items, that were all scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. The only CMC item was also scored dichotomously, due to collapsed categories. MC items are marked with a '_c' at the end of the variable name, whereas the variable name of the CMC item end in 's_c'. Items that were already administered in other grades kept their original names ('mag5d041…', 'mag5v271…', 'mag5q231…', 'mag5q301…', 'mag5d051…', 'mag5q140…', 'mag5v071…', and 'mag5r191…'). However, for reasons of identification a suffix was added in front of the '…_c' (scored item) to specify the current test administration ('sc2g4' referring to Starting Cohort 2, Grade 4).

6.2 Linking the data of Grade 4

In starting cohort 2, the mathematics competence tests administered in kindergarten, grade 1, grade 2, and grade 4 for the large part include different items that were constructed in such a way that allows 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 a reference scale that all subsequent measurement waves will refer to. The process of linking the mathematics competence in kindergarten and grade 1 will be described in the technical report of grade 1 (Starting Cohort 2, First Grade; Schnittjer & Fischer, 2018). Furthermore, the process of linking the mathematics competence in grade 1 and grade 2 was described in the technical report of grade 2 (Starting Cohort 2, Second Grade; Schnittjer & Gerken, 2018).

For the domain of mathematical competence, linking typically is achieved using overlapping items (also known as common items). In this case, we are following an anchor-group design, because of the enormous competence growth from grade 2 to grade 4. An independent link sample, including students from grade 4, that were not part of starting cohort 2 were administered all items from the grade 2 and the grade 4 mathematics competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 2 across the two grades. 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 NEPS tests. For more information on the selection of link samples 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 5,359 students participated at both measurement occasions, in grade 2 and grade 4. Consequently, these respontants were used to link the two tests across both grades (see Fischer et al., 2016). However, due to a change in methods as well as an enormous competence growth from grade two to grade 4, there were no common items in the two tests. Therefore, for linking those two measurements, an independant link sample of 299 4th grade students was selected. These 4th grade students were administered to both competence tests, the 2nd grade test as well as the 4th grade test of starting cohort 2.

6.2.2 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items of both tests to a two-dimensional model, loading the items of the 2nd grade test on one dimension and the items of the 4th grade test on the other dimension. As shown in Table 10, AIC as well as BIC clearly favoured the one-dimensional model. Furthermore, the corrected Q_3 statistics (Yen, 1984) underlines unidimensionality ($M(Q_3) = 0$, $SD(Q_3) = 0.07$). Therefore, a unidimensional scale can be assumed for the mathematics competence tests in grade 2 and grade 4.

Table 10: Comparison of the Unidimensional and the two-Dimensional Model.

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	10,045.994	53	10,151.994	10,348.118
Two-dimensional	11,077.614	55	11,187.614	11,391.138

Note. The results in this table were achieved by using ConQuest 4.2.5.

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 the the longitudinal subsample from starting cohort 2. 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 11.

Analyses of differential item functioning between the link sample and starting cohort 2 did not identify items with significant (α = .05) DIF for grade 2 (difference in logits: *Min* = -1.334, *Max* = 1.558), nor for grade 4 (difference in logits: *Min* = -0.460, *Max* = 0.456). Since the differences in logits for grade 2 were very large, we only used items for calculating the linking correction term *c* that had no difference in logits greater than |.6| (see Table 11; items marked with " * " were not used for the calculation). The differences in logits for grade 4 were less than |.6| and therefore all items of grade 4 were used for calculating the correction term *c*. The mathematical competence tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term for grade 2 and 4 was calculated as *c* = 2.394. Added to the correction term for kindergarten and first grade (see Schnittjer & Gerken, 2018), a total correction term of 4.620 was derived. This correction term was subsequently added to each difficulty parameter estimated in grade 4 (see Table 4) to derive the linked item parameters. The link error, reflecting the uncertainty in the linking process, was calculated according to equation 4 in Fischer et al. (2016) as 0.074 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

Table 11: Differential Item Functioning Analyses between the Starting Cohort and the Link Sample.

	Grade 2				Grade 4			
	Item	**$\Delta\sigma$**	**$SE_{\Delta\sigma}$**	**F**	**Item**	**$\Delta\sigma$**	**$SE_{\Delta\sigma}$**	**F**
1	mag1v051_sc2g2_c*	0.655	0.269	5.9	mag5d041_sc2g4_c	-0.098	0.139	0.5
2	mag2v071_c	-0.283	0.317	0.8	mag4q101_c	0.175	0.203	0.7
3	mag2r031_c	-0.524	0.595	0.8	mag4r021_c	-0.046	0.134	0.1
4	mag2d061_c	0.080	0.262	0.1	mag5v271_sc2g4_c	0.025	0.157	0.0
5	mag1d131_sc2g2_c	-0.329	0.274	1.4	mag4q011_c	0.114	0.175	0.4
6	mag2r131_c	-0.103	0.213	0.2	mag4r071_c	-0.251	0.141	3.2
7	mag2v121_c*	0.760	0.215	12.5	mag4d131_c	0.054	0.279	0.0
8	mag2q061_c*	1.034	0.152	46.0	mag5q231_sc2g4_c	-0.032	0.155	0.0
9	mag2r111_c*	1.558	0.161	94.1	mag5q301_sc2g4_c	-0.460	0.136	11.4
10	mag1d09s_sc2g2_c	0.320	0.322	1.0	mag4v121_c	-0.003	0.136	0.0
11	mag1z121_sc2g2_c*	-1.227	0.200	37.7	mag5d051_sc2g4_c	-0.294	0.195	2.3
12	mag2g12s_c	-0.182	0.675	0.1	mag4q06s_c	-0.107	0.218	0.2
13	mag1d081_sc2g2_c	-0.019	0.468	0.0	mag4d031_c	-0.036	0.138	0.1
14	mag2g021_c*	-1.334	0.289	21.3	mag5q14s_sc2g4_c	-0.190	0.151	1.6
15	mag2r151_c*	0.759	0.223	11.6	mag4v111_c	0.144	0.221	0.4
16	mag1v021_sc2g2_c	0.449	0.207	4.7	mag4r041_c	-0.439	0.320	1.9
17	mag1z071_sc2g2_c	0.233	0.235	1.0	mag4r042_c	0.050	0.164	0.1
18	mag2d101_c	-0.199	0.431	0.2	mag4q051_c	0.129	0.171	0.6
19	mag1g031_sc2g2_c	0.022	0.359	0.0	mag4q091_c	0.456	0.163	7.9
20	mag2v041_c	-0.168	0.297	0.3	mag4q092_c	-0.062	0.172	0.1
21	mag2q011_c	-0.442	0.281	2.5	mag4d14s_c	0.052	0.186	0.1
22	mag1r19s_sc2g2_c*	0.903	0.218	17.1	mag5v071_sc2g4_c	0.330	0.217	2.3
23	mag2g091_c*	-0.889	0.379	5.5	mag5r191_sc2g4_c	0.050	0.189	0.1
24	mag2q051_c*	-1.076	0.721	2.2	mag4d081_c	0.432	0.218	3.9

Note. $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 2 / grade 4 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154}(1,5658) = 122.32. A non-significant test indicates measurement invariance.
* These items had differnces in logits gerater than |.6| and were excluded from caluculating the link constant und link error.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs, mag4_sc1 and mag4_sc1u, including their respective standard errors, mag4_sc2 and mag4_sc2u. There are four measurement points for mathematical competence in starting cohort 2 and therefore the uncorrected WLEs (mag4_sc1u) are linked to the first measurement point in the last year of kindergarten. As a result the WLE scores provided in mag4_sc1u can be used for longitudinal comparisons between the measurement points. In contrast, the WLE scores in mag4_sc1 are not linked to the underlying reference scale of grade 2 and therefore should be used 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. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

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. (2019). *Competence data in NEPS: Overview ofmeasures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg, Germany: 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.

Kiefer, T., Robitzsch, A., & Wu, M. (2017). *TAM: Test Analysis Modules*. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 2.8-21).

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

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

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.

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

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

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

Schnittjer, I., & Fischer, L. (2018): NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1 (NEPS Survey Paper No. 46). Bamberg: Leibniz Institute 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- Grade 4

```
Title Starting Cohort II, MATHEMATICS: Rasch Model;


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 (0,1,2,3,4,5)   (0,0,0,0,0,1)   !item (21); /* collapsing the lowest categories * /


score (0,1)            (0,1)           !item (1-24);


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       4.352   /* mag5d041_sc2g4_c */
2       6.911   /* mag4q101_c */
3       5.127   /* mag4r021_c */
4       5.799   /* mag5v271_sc2g4_c */
5       6.575   /* mag4q011_c */
6       5.618   /* mag4r071_c */
7       1.781   /* mag4d131_c */
8       5.791   /* mag5q231_sc2g4_c */
9       5.671   /* mag5q301_sc2g4_c */
10      4.552   /* mag4v121_c */
11      3.038   /* mag5d051_sc2g4_c */
12      7.250   /* mag4q060_c */
13      5.167   /* mag4d031_c */
14      4.594   /* mag5q140_sc2g4_c */
15      6.755   /* mag4v111_c */
16      1.996   /* mag4r041_c */
17      3.837   /* mag4r042_c */
18      3.394   /* mag4q051_c */
19      3.739   /* mag4q091_c */
20      4.621   /* mag4q092_c */
21      3.897   /* mag4d14s_c */
22      2.897   /* mag5v071_sc2g4_c */
23      5.306   /* mag5r191_sc2g4_c */
24      3.110   /* mag4d081_c */
25      4.620   /* correcting for kindergarten to grade 4*/
```

Appendix C: Content Areas of Items in the Mathematics Test for Grade 4

Position	Item	Content area
1	mag5d041_sc2g4_c	Data and chance
2	mag4q101_c	Quantity
3	mag4r021_c	Space and shape
4	mag5v271_sc2g4_c	Change and relationships
5	mag4q011_c	Quantity
6	mag4r071_c	Space and shape
7	mag4d131_c	Data and chance
8	mag5q231_sc2g4_c	Quantity
9	mag5q301_sc2g4_c	Quantity
10	mag4v121_c	Change and relationships
11	mag5d051_sc2g4_c	Data and chance
12	mag4q060_c	Quantity
13	mag4d031_c	Data and chance
14	mag5q140_sc2g4_c	Quantity
15	mag4v111_c	Change and relationships
16	mag4r041_c	Space and shape
17	mag4r042_c	Space and shape
18	mag4q051_c	Quantity
19	mag4q091_c	Quantity
20	mag4q092_c	Quantity
21	mag4d14s_c	Data and chance
22	mag5v071_sc2g4_c	Change and relationships
23	mag5r191_sc2g4_c	Space and shape
24	mag4d081_c	Data and chance

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