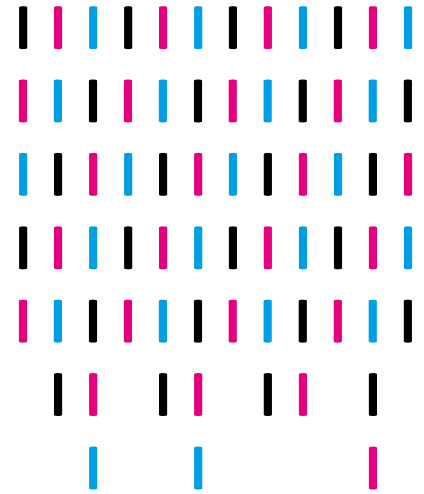




NEPS *SURVEY PAPERS*

Lara Aylin Petersen and Tessa Beyer
**NEPS TECHNICAL REPORT
FOR MATHEMATICS:
SCALING RESULTS
OF STARTING COHORT 8
FOR GRADE 5**

NEPS *Survey Paper* No. 122
Bamberg, August 2025

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).

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 edited by a review board consisting of the scientific management of LifBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section "Publications") and at www.lifbi.de/publications.

Editor-in-Chief: Thomas Bäumer, LifBi

Review Board: Board of Directors, Heads of LifBi Departments, and Scientific Management of NEPS Working Units

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



This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 8 for Grade 5

*Lara Aylin Petersen & Tessa Beyer
Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany*

Email address of the lead author:

lpetersen@leibniz-ipn.de

Bibliographic Data:

Petersen, L. A. & Beyer, T. (2025). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 8 for Grade 5* (NEPS Survey Paper No. 122). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP122:1.0>

Acknowledgements:

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

The present report has been modeled along previous reports published by the NEPS. To facilitate the understanding of the presented results, many text passages (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Duchhardt & Gerdes, 2012; Gnambs, 2025; Petersen & Beyer, 2025).

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 8 for Grade 5

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. To evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) were performed. This report describes the data and scaling procedure for the mathematical competence test for Grade 5 students of Starting Cohort 8. The mathematics test consisted of 24 items representing different content areas as well as different cognitive components and used different response formats. The test was administered to 5,424 students (50% girls) from regular schools. A partial-credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test's dimensionality were evaluated to ensure the quality of the test. The results show that the test exhibited a good reliability (EAP/PV reliability = .82), good item fit statistics, and negligible differential item functioning across different subgroups. Limitations of the test include some difficult items that were missing for the accurate proficiency estimation in the upper ability range. Overall, the results revealed predominantly good psychometric properties of the mathematics test, thus supporting the estimation of an acceptable mathematics competence score. Furthermore, analyses of differential item functioning showed comparable measurement models for the test administered in Starting Cohort 8 and an identical test previously used in Starting Cohort 3 in the same grade. Competence scores from both tests were therefore linked to allow comparisons between the two cohorts. Besides the scaling results, this report also describes the data available in the Scientific Use File and provides the R syntax for scaling the data.

Keywords

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

Content

1	Introduction.....	6
2	Testing Mathematical Competence	6
3	Data and Methods.....	7
3.1	The Design and Procedure of the Study.....	7
3.2	Sample	8
3.3	Missing Responses	8
3.4	Scaling Model	8
3.5	Checking the Quality of the Scale	9
3.6	Software	10
4	Results	10
4.1	Missing Responses	10
4.1.1	Missing Responses per Person	10
4.1.2	Missing Responses per Item.....	14
4.2	Parameter Estimates	15
4.2.1	Item Parameters.....	15
4.2.2	Test Targeting and Reliability.....	16
4.3	Quality of the test	17
4.3.1	Distractor Analyses MC items	17
4.3.2	Item Fit	18
4.3.3	Differential Item Functioning.....	18
4.3.4	Rasch-homogeneity.....	22
4.3.5	Unidimensionality	23
5	Discussion	24
6	Data in the Scientific Use File	24
6.1	Naming Conventions	24
6.2	Linking of the Competence Test.....	24
6.3	Mathematical competence scores	25
7	References.....	27
	Appendix A. Content Areas of Items in Starting Cohort 8 Grade 5.....	29
	Appendix B. R Syntax for Estimating linked WLEs.....	30

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, and information and communication technologies literacy. An overview of the competence domains measured in the NEPS is given by Fuß et al. (2025) as well as Weinert et al. (2019).

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

This report presents the results of these analyses for a test of mathematical competence that was administered in fifth grades of regular schools in Starting Cohort 8 (Wave 1). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of the first wave of Starting Cohort 8 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 (SUF) is presented.

The present report has been modeled on previous reports (Duchhardt & Gerdes, 2012; Gnams, 2025; Petersen & Beyer, 2025). Please note that the analyses of this report are based on the data available some time before data release. Due to ongoing data protection and data cleansing issues, the data set in the SUF may differ slightly from the data set used for the analyses in this report. However, we do not expect fundamental changes in the presented results.

2 Testing Mathematical Competence

The framework and test development for the mathematical competence test are described in Ehmke et al. (2009), Neumann et al. (2013), and Weinert et al. (2019). In the following, specific aspects of the mathematics test will be pointed out that are necessary for understanding the scaling results presented in this report.

The mathematics test administered in the present study consists of 24 items that represented different content-related and process-related components and used different response formats. The administered mathematical test presented respondents with a certain situation, followed by one or two tasks related to it. Each item belongs to one of four content areas (see Table 1).

Each item was constructed in such a way as to primarily address a specific content area (see Appendix A for the allocation of each item to these areas). The framework also describes, as a second and independent dimension, six cognitive components required for solving the tasks. These were distributed across the items.

Table 1: Number of Items by Content Areas

Content area	Number of items
Quantity	8
Space and shape	5
Change and relationships	6
Data and chance	5
Total number of items	24

The mathematics test included three types of response formats: Simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR). In MC items, the test taker had to find the correct response option from several available response options (four or five). In CMC items, a number of subtasks with two response options were presented. SCR items required the test taker to write down an answer in an empty box. Two SCR items (mag5q140_sc8g5_c and mag5d02s_sc8g5_c) each required to write down two answers that were closely related (e.g., a value range). The two answers were scored together for each item and, therefore, these two items were treated as SCR items and not as CMC items. As a consequence, these two SCR items were scored dichotomously. Table 2 shows the distribution of the three response formats across the items of the mathematics test. There was no multi-matrix design regarding the order of the items within the test. All students answered the test items in the same order.

Table 2: Number of Items by Response Formats

Response format	Number of items
Simple multiple-choice	12
Complex multiple-choice	1
Short constructed response	11
Total number of items	24

3 Data and Methods

3.1 The Design and Procedure of the Study

The study was conducted in the last months of Year 2022 in groups of up to 24 students at their respective regular schools and assessed different competence domains. These included reading and mathematical competence. The two tests were presented on paper and also had a common introduction on paper. The introduction was read aloud by trained test administrators, while the students were able to read along in a paper booklet. In order to control for test position effects, the two tests (reading and mathematics) were assigned to

test takers in a different order. Half of the participants were given a booklet that included the reading test followed by the mathematics test, while the other half were given the two tests in reverse order. The allocation of the two booklets was randomized. The students had 28 minutes to answer the mathematical or reading items. After the timeout, the test administrators proceeded to the next competence domain (mathematics or reading). A detailed description of the study design is available on the NEPS website (<http://www.neps-data.de>).

3.2 Sample

A total of 5,424 students received the mathematics test. For 13 respondents, fewer than three valid responses were available (e.g., if the student arrived late and was only able to join the second test). Because no reliable ability scores can be estimated based on such few responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the results presented in this report (see Chapter 4) are based on a sample of 5,411 students in Grade 5, with 2,715 (50.18%) girls.

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) not reached items, c) invalid responses, and d) multiple types of missing responses within CMC and SCR items that are not determined.

Omitted items occurred when students skipped some items. Because of the time limit, not all students reached the last items. In some rare cases, the response was invalid (e.g., when two response options were selected in MC items, although only one was required). Finally, as the partial credit item was aggregated from several subtasks and two SCR items were scored by taking two responses into account (mag5q140_sc8g5_c and mag5d02s_sc8g5_c), different kinds of missing responses or a mixture of valid and missing responses might be found for these items. In rare cases, multiple missings were observed. If the subtasks contained different types of missing responses, the item was coded as a not-determinable missing response. The polytomous item was coded as missing if at least one subtask contained a missing response.

Missing responses provide information on how well the test worked (e.g., 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 students were coping with the test. Missing responses per item were examined 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). The CMC item mag5v01s_sc8g5_c consisted of a set of subtasks that were aggregated to a polytomous variable, indicating the number of correctly responded subtasks within that item. Response categories of the polytomous item with few responses were collapsed in order to avoid possible estimation problems (Pohl & Carstensen, 2012). These categories were collapsed in the same way as in Grade 5 of Starting Cohort 3 (Duchhardt & Gerdes, 2012) to

ensure the comparability of the starting cohorts. Thus, the lower two categories of item mag5v01s_sc8g5_c were collapsed (see Appendix B).

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 and SCR items were scored dichotomously with 0 for an incorrect and 1 for the correct response (see Haberkorn et al., 2016, 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 is described in Chapter 6.

3.5 Checking the Quality of the Scale

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

The MC items consisted of one correct response option and several distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined to evaluate whether the distractors were predominantly chosen by respondents with a lower ability rather than by those who gave a correct response. We therefore calculated the point-biserial correlation between selecting an incorrect response option for a given item and the total correct score for the remaining items. Negative correlations indicate good distractors, while correlations between .00 and .05 were considered acceptable and correlations above .05 were considered problematic distractors (Pohl & Carstensen, 2012).

Before aggregating the subtasks of CMC items to polytomous variables, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC and the SCR items using 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 were they used to construct the polytomous CMC variables that were included in the final scaling model.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC items, the polytomous CMC items, and the dichotomous SCR items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 ($|t\text{-value}| > 6$) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 ($|t\text{-value}| > 8$) were judged as having a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total correct score greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. The 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, measurement bias was investigated for the variables gender, the number of books at home (as a proxy for cultural capital), migration background, school type, and test position

(first versus second). In addition, we compared the mathematical test administered in Grades 5 of Starting Cohort 3 (Duchhardt & Gerdes, 2012) and Starting Cohort 8 to examine whether competence scores could be compared between the two cohorts. To test for measurement invariance, DIF was estimated using a multi-group IRT model, in which the main effects of the subgroups as well as differential effects of the subgroups on item difficulty were estimated. 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 logits as considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 logits as small but not severe, and differences smaller than 0.4 logit as negligible DIF. Additionally, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

The mathematics test was scaled using the PCM (Masters, 1982) because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nevertheless, 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 test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations, as indicated by Yen's Q3 (1984), suggest unidimensionality. Because the Q3 tends to be slightly negative in case of locally independent items, we report the corrected Q3, which has an expected value of 0. According to Yen (1993), values of Q3 falling below .20 indicate essential unidimensionality. In addition, the assumption of unidimensionality was also 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; see Appendix A). The number of nodes in the multidimensional model using quasi-Monte Carlo integration was chosen in such a way as to obtain stable parameter estimates (10,000 nodes). The correlations between the subdimensions and the differences in model fit between the unidimensional and multidimensional models were used to evaluate the unidimensionality of the test.

3.6 Software

All IRT models were estimated with the TAM package Version 4.2-21 (Robitzsch et al., 2024) in R version 4.5.0 (R Core Team, 2025).

4 Results

4.1 Missing Responses

4.1.1 Missing Responses per Person

Figure 1 illustrates the number of missing responses per person when students omitted items. As can be seen, 51.04% did not omit any of the items, while 16.85% omitted one item, 10.39% omitted two items, 6.93% omitted three items, 5.17% omitted four items, and 9.61% omitted more than five items.

Another source of missing responses is items that students did not reach because they aborted the test (e.g., lack of motivation) or due to time constraints (the study had a timeout after 28 minutes). These missing values refer to items after the last valid response. As shown in Figure 2, 75.16% of the students reached the last item. However, 6.71% of the students did not reach one item, two to ten items were not reached by 16.13% of the students, and 11 or more items were not reached by 2.00% of the students.

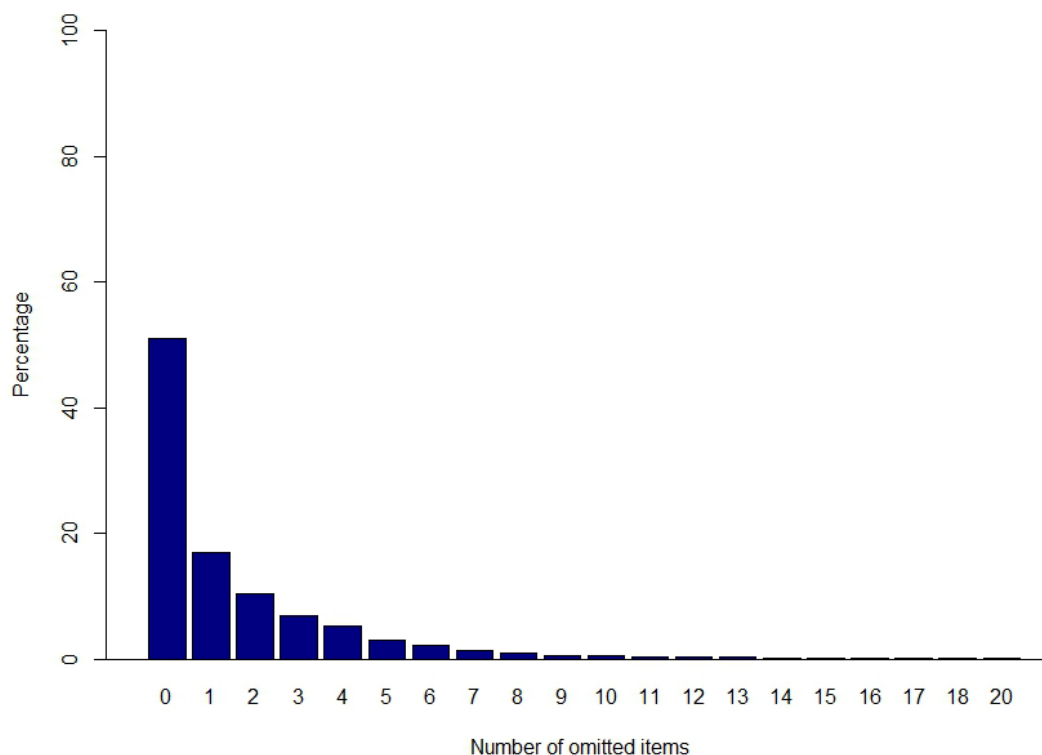


Figure 1. Number of Omitted Items.

The number of invalid responses per person is shown in Figure 3. Not valid responses occurred, for example, if a student had not marked an MC item answer field correctly, but had placed a cross between two answer fields. The number of invalid responses was rather low, with 94.27% of the students having no invalid responses at all. Only 4.92 % of the students had one invalid response, and 0.81% of the students had two or more invalid responses.

For two items, responses were coded as not determinable missing if the subtasks of the CMC item mag5v01s_sc8g5_c and the SCR item mag5q140_sc8g5_c contained different types of missing responses. Figure 4 shows the percentage of not determinable missing responses in the mathematics test. 99.83% of the students did not have a single not determinable missing response.

In Figure 5, the total number of missing responses per person is depicted. 40.86% of the students had no missing responses at all, 15.19% of the students had one missing response, 8.83% of the students had two missing responses, 29.66% had three to ten missing responses, and 5.45% of the students had 11 or more missing responses.

In sum, the number of missing responses was rather small, even though the omitted items and not-reached items accounted for the greatest impact on the total number of missing responses. This indicates that for most students the test functioned as intended.

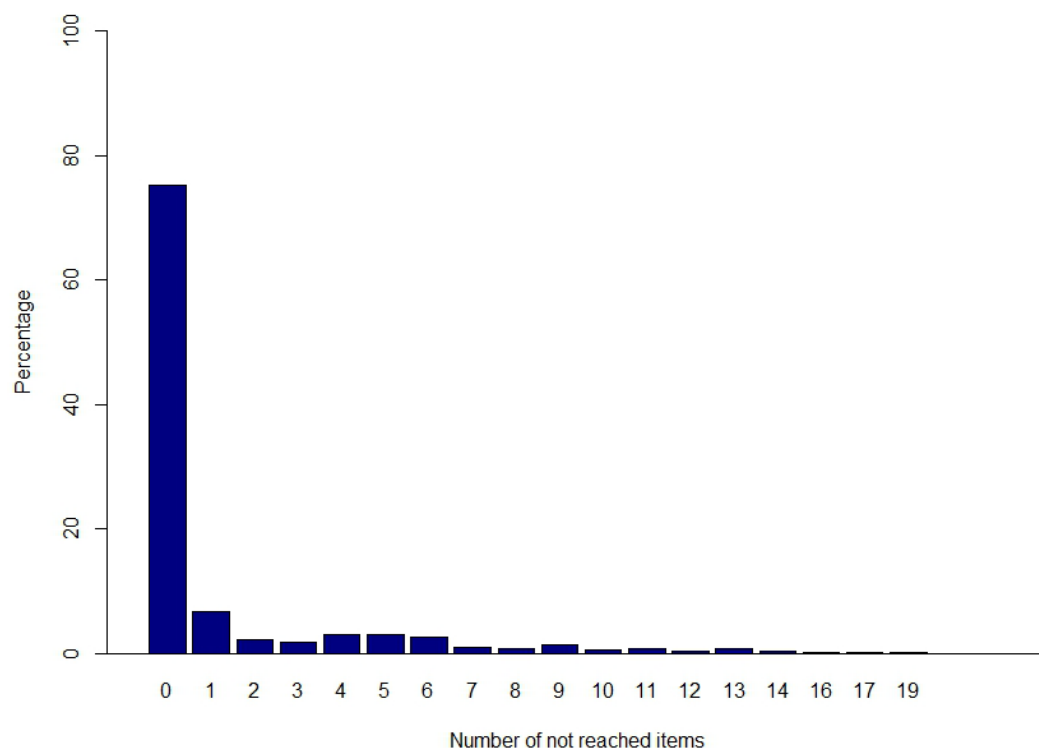


Figure 2. Number of Not Reached Items per Person (e.g., Timeout).

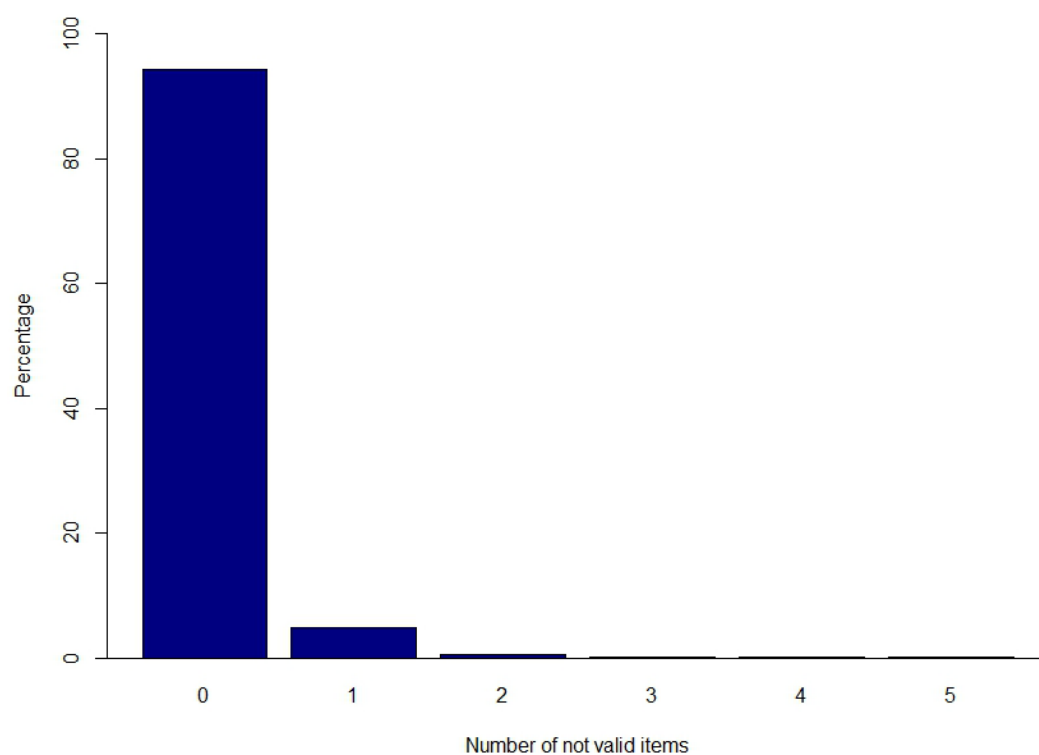


Figure 3. Number of Not Valid Responses.

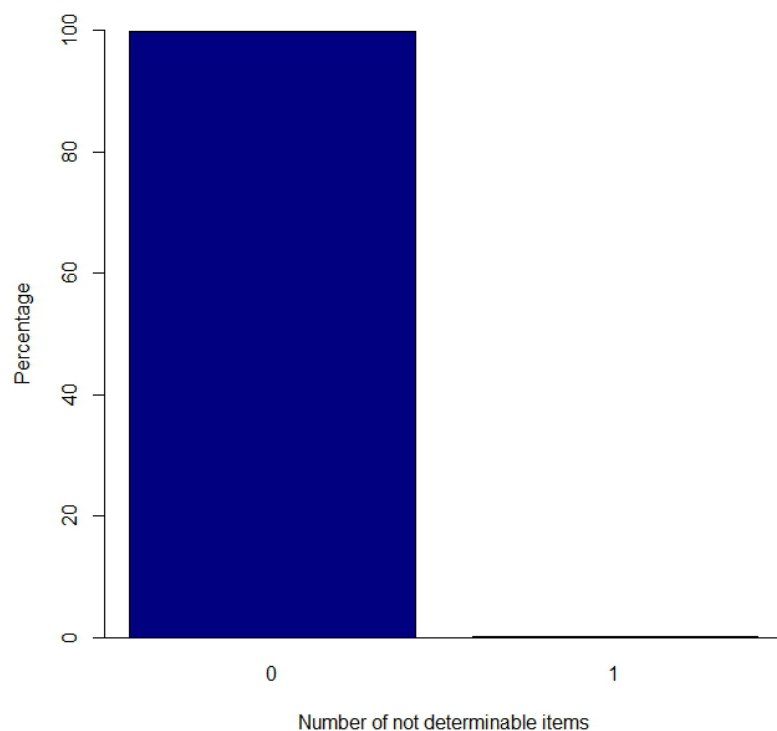


Figure 4. Number of Not Determinable Items.

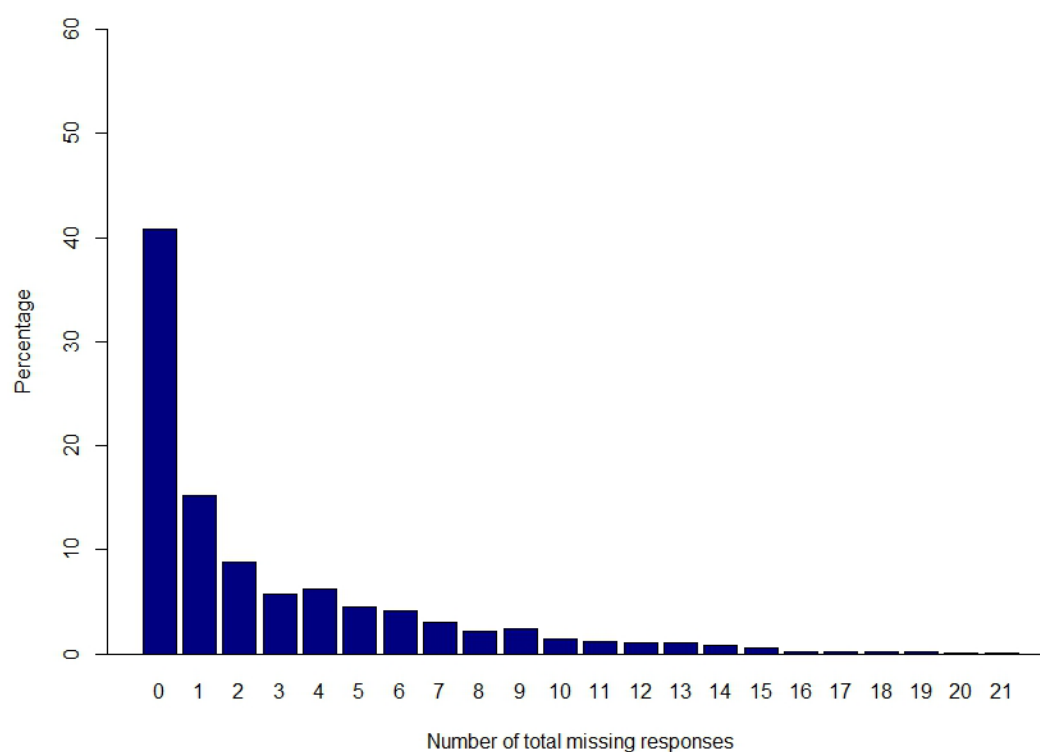


Figure 5. Total Number of Missing Responses.

4.1.2 Missing Responses per Item

Table 3 shows the percentage of different missing responses per item. Overall, the number of omitted responses per item was small, varying between 0% (mag5v091_sc8g5_c) and 13.60% (mag5q140_sc8g5_c). The percentage of missing responses due to not reached items varied between 0% (the first five items) and 24.84% (last item mag5v091_sc8g5_c). The number of missing responses due to not valid responses varied from 0% (mag5q291_sc8g5_c, mag5q292_sc8g5_c, mag5q231_sc8g5_c, mag5q221_sc8g5_c, mag5q131_sc8g5_c, and mag5v321_sc8g5_c) to 2.25% (mag5r191_sc8g5_c).

Table 3: Percentage of Missing Values per Item

Pos.	Item	N	OM	NR	NV
1	mag5d041_sc8g5_c	5314	1.66	0.00	0.13
2	mag5q291_sc8g5_c	5021	7.21	0.00	0.00
3	mag5q292_sc8g5_c	4941	8.69	0.00	0.00
4	mag5v271_sc8g5_c	4858	10.15	0.00	0.07
5	mag5r171_sc8g5_c	5138	4.79	0.00	0.26
6	mag5q231_sc8g5_c	4697	13.12	0.07	0.00
7	mag5q301_sc8g5_c	5231	3.23	0.07	0.02
8	mag5q221_sc8g5_c	5153	4.66	0.11	0.00
9	mag5d051_sc8g5_c	5318	1.53	0.15	0.04
10	mag5d052_sc8g5_c	5215	1.48	0.15	2.00
11	mag5q140_sc8g5_c	4645	13.60	0.44	0.04
12	mag5q121_sc8g5_c	4801	10.07	1.09	0.11
13	mag5r101_sc8g5_c	5119	3.10	1.35	0.94
14	mag5r201_sc8g5_c	5153	2.64	2.00	0.13
15	mag5q131_sc8g5_c	4933	6.30	2.53	0.00
16	mag5d02s_sc8g5_c	4776	7.76	3.95	0.02
17	mag5d023_sc8g5_c	4723	7.97	4.73	0.02
18	mag5v024_sc8g5_c	4384	13.32	5.64	0.02
19	mag5r251_sc8g5_c	4545	7.78	8.13	0.09
20	mag5v01s_sc8g5_c	4392	7.26	11.14	0.33
21	mag5v321_sc8g5_c	4114	9.81	14.16	0.00
22	mag5v071_sc8g5_c	4473	1.15	15.97	0.22
23	mag5r191_sc8g5_c	4221	1.61	18.13	2.25
24	mag5v091_sc8g5_c	4058	0.00	24.84	0.17

Note. Pos. = Item position within the test. N = Number of valid responses, OM = Percentage of respondents that omitted the item, NR = percentage of students that did not reach an item, NV = percentage of invalid responses.

4.2 Parameter Estimates

4.2.1 Item Parameters

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 items as single variables, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 22.40% (mag5q121_sc8g5_c) and 91.60% (mag5d051_sc8g5_c) across all items. On average, the rate of correct responses was 59.10% ($SD = 18.69\%$).

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) for the data with aggregated subtasks of the CMC item are depicted in Table 4a. The step parameters for the polytomous variable are presented in Table 4b. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.89 (mag5d051_sc8g5_c) and 1.56 (mag5q121_sc8g5_c) with a mean of -0.55 ($SD = 1.09$). Due to the large sample size, the standard errors of the estimated item difficulties (Table 4a, column SE) were small ($SE(\beta) \leq 0.05$).

Table 4a: Item Parameters

No.	Item	Item format	PC %	Diffic.	SE	WMN SQ	t	r_{it}	Discr.	aQ_3
1	mag5d041_sc8g5_c	MC	63.17	-0.68	0.03	1.00	0.02	.49	1.18	.08
2	mag5q291_sc8g5_c	SCR	70.46	-1.05	0.03	0.99	-0.73	.47	1.19	.11
3	mag5q292_sc8g5_c	SCR	64.54	-0.70	0.03	1.03	2.13	.45	1.02	.12
4	mag5v271_sc8g5_c	MC	35.69	0.72	0.03	1.10	6.64	.40	0.76	.09
5	mag5r171_sc8g5_c	MC	51.32	-0.06	0.03	1.03	2.70	.47	1.03	.09
6	mag5q231_sc8g5_c	SCR	42.67	0.41	0.03	1.00	-0.31	.50	1.18	.08
7	mag5q301_sc8g5_c	SCR	35.94	0.76	0.03	0.94	-4.78	.55	1.51	.08
8	mag5q221_sc8g5_c	SCR	81.97	-1.84	0.04	0.99	-0.49	.41	1.21	.08
9	mag5d051_sc8g5_c	MC	91.56	-2.89	0.05	0.92	-2.15	.40	1.90	.07
10	mag5d052_sc8g5_c	MC	68.32	-0.96	0.03	0.96	-3.01	.51	1.41	.08
11	mag5q140_sc8g5_c	SCR	67.02	-0.84	0.04	0.93	-4.85	.54	1.56	.08
12	mag5q121_sc8g5_c	MC	22.37	1.56	0.04	1.01	0.68	.43	1.04	.08
13	mag5r101_sc8g5_c	MC	54.11	-0.20	0.03	1.12	9.33	.40	0.75	.10
14	mag5r201_sc8g5_c	MC	71.96	-1.18	0.04	1.00	0.18	.45	1.15	.08
15	mag5q131_sc8g5_c	SCR	74.01	-1.26	0.04	0.99	-0.65	.45	1.19	.08
16	mag5d02s_sc8g5_c	SCR	82.60	-1.87	0.04	0.93	-2.85	.46	1.62	.08
17	mag5d023_sc8g5_c	SCR	56.79	-0.29	0.03	1.03	1.86	.47	1.07	.09
18	mag5v024_sc8g5_c	SCR	56.46	-0.27	0.03	0.98	-1.43	.51	1.26	.08

No.	Item	Item format	PC %	Diffic.	SE	WMN SQ	<i>t</i>	<i>r</i> _{it}	Discr.	aQ ₃
19	mag5r251_sc8g5_c	MC	48.82	0.09	0.03	0.99	-0.98	.51	1.22	.08
20	mag5v01s_sc8g5_c	CMC	n.a.	-1.22	0.02	0.94	-3.12	.53	0.74	.07
21	mag5v321_sc8g5_c	SCR	34.76	0.87	0.04	1.03	2.10	.46	1.05	.08
22	mag5v071_sc8g5_c	MC	90.27	-2.69	0.05	1.01	0.36	.32	1.10	.07
23	mag5r191_sc8g5_c	MC	53.83	-0.16	0.04	1.00	0.13	.50	1.15	.09
24	mag5v091_sc8g5_c	MC	40.54	0.52	0.04	0.99	-0.65	.50	1.17	.08

Note. No. = Number in scaling data set, PC % = Percentage correct answers for MC and SCR, Diffic. = 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, aQ₃ = Adjusted average absolute residual correlation for item (Yen, 1993).

Percentage correct scores are not informative for polytomous CMC item scores. Therefore, 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.

Table 4b: Step Parameters (with Standard Errors) of the Polytomous Item

Item	Step 1	Step 2
mag5v01s_sc8g5_c	-0.53 (0.03)	0.63 (0.04)

Note. The last step parameter is a constrained parameter and not reported in the table.

4.2.2 Test Targeting and Reliability

Test targeting was investigated to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In Figure 6, the item difficulties of the mathematics items and the ability of the students are plotted on the same scale. The distribution of the estimated students' ability is mapped onto the left side, whereas the right side shows the distribution of item difficulties. The respective difficulties ranged from -2.89 (mag5d051_sc8g5_c) to 1.56 (mag5q121_sc8g5_c) with a mean of -0.55 (*SD* = 1.09), corresponding to Table 4a. Therefore, a rather broad range with easy and difficult items was spanned. The mean of the ability distribution was constrained to be zero. However, the average item difficulty is negative. Thus, although the items covered a wide range of the ability distribution, on average, the items were slightly too easy for the students. As a consequence, students with a low or medium ability will be measured relatively precisely, while students with a high mathematical competence will have a larger standard error. The variance was estimated to be 1.37, which implies a good differentiation between the students. The reliability of the test was good (EAP/PV reliability = .82, WLE reliability = .79).

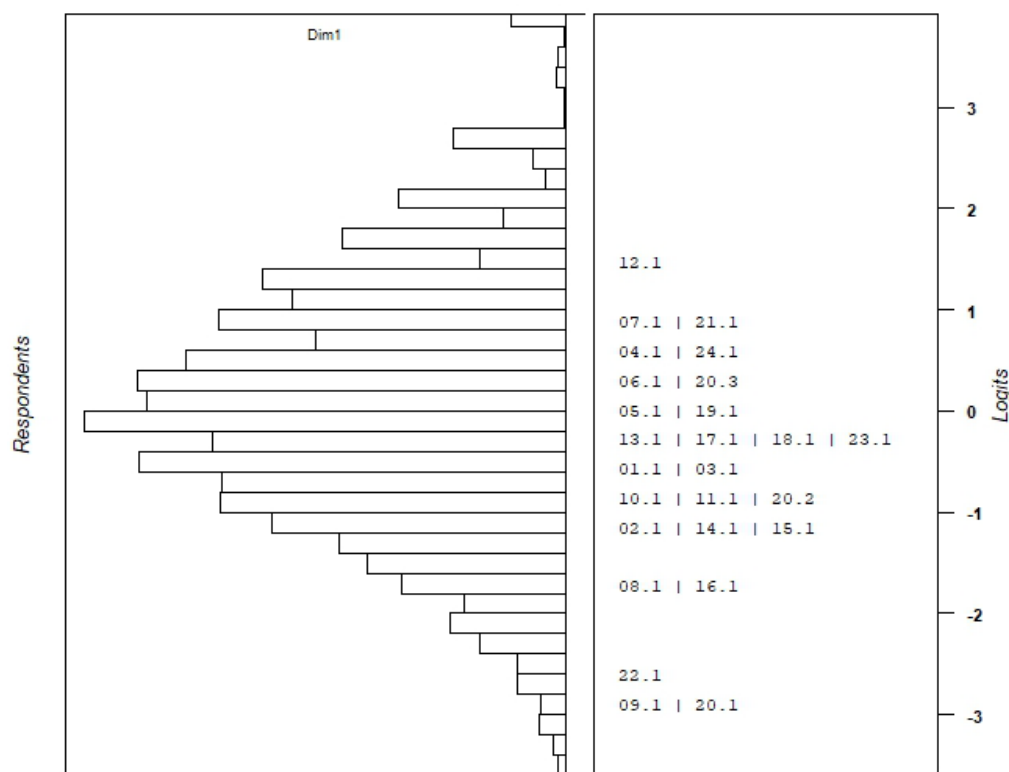


Figure 6. Wright map. The distribution of person ability in the sample is depicted on the left-hand side of the graph. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item corresponding to the number in Table 4a (the second number indicating the threshold).

4.3 Quality of the test

4.3.1 Distractor Analyses MC items

To investigate how well the distractors of the MC items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the students' total correct scores for the remaining items. Table 5 shows the summary of the point-biserial correlations. These results indicate that the distractors worked well.

Table 5: Point Biserial Correlation of Correct and Incorrect Response Options (only MC items)

Parameter	Correct responses	Incorrect responses
Mean	0.36	-0.32
Minimum	0.25	-0.57
Maximum	0.42	-0.13

4.3.2 Item Fit

The evaluation of the item fit was performed based on the final scaling model, the partial credit model, using the items of all response formats. Overall, the item fit was good (see Table 4a). The values of the WMNSQ were close to 1, with the lowest value being 0.92 (mag5d051_sc8g5_c) and the highest being 1.12 (mag5r101_sc8g5_c). Only two t -values were over $|6|$ (mag5v271_sc8g5_c and mag5r101_sc8g5_c). However, we did not indicate a serious misfit. Moreover, all ICCs showed a good fit of the items. Thus, there was no indication of a severe item over- or underfit. The correlations of the item scores with the total scores varied between 0.32 (mag5v071_sc8g5_c) and 0.55 (mag5q301_sc8g5_c), and thus, can all be interpreted as good. Overall, the items showed an average correlation of 0.47.

4.3.3 Differential Item Functioning

We examined measurement bias for several subgroups by estimating differential item functioning (DIF). DIF was investigated for the variables gender, migration background, number of books at home (as a proxy for cultural capital), school type, and test position. A description of these variables can be found in Pohl and Carstensen (2012). Table 6 shows the difference between the estimated difficulties of the items in different subgroups. For example, the column “male vs. female” reports the differences in item difficulties between male and female students. A positive value would indicate that the test was more difficult for males, whereas a negative value would indicate that the test was more difficult for females. Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 7) using Akaike’s (1974) information criterion (AIC) and Bayesian information criterion (BIC, Schwarz, 1978). The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents an overparameterization of models.

Gender: Overall, 2,715 (50.18%) of the students were female, 2,695 (49.82%) were male, and one participant had no information (was excluded for this analysis). On average, male students exhibited a slightly higher mathematical competence than female students (main effect = -0.24 logits, Cohen’s $d = -0.22$). DIF between 0.40 and 0.60 logits occurred only for the item mag5v01s_sc8g5_c. Two items slightly exceeded 0.60 logits (mag5q131_sc8g5_c and mag5q291_sc8g5_c). An overall test for DIF (see Table 7) also suggested a better fit for the model with DIF compared to a model that only estimated the main effect (but ignored potential DIF). Because an evaluation of the item content did not suggest an explanation for the observed DIF in the two items and the item mag5v01s_sc8g5_c did not show DIF in Starting Cohort 3 (see Duchhardt & Gerdes, 2012), no gender-specific item effects were acknowledged in the scaling of the test.

Migration: The sample included 4,644 (85.83%) students without a migration background, 232 (4.29%) students with a migration background, and 535 (9.89%) students with no information on their migration background (“n.a.”). All three groups were used to investigate DIF of migration. There was a considerable difference between the average performance of students with and without a migration background. Students without a migration background had a higher mathematics ability than students with a migration background (main effect = -0.68 logits, Cohen’s $d = -0.62$). On average, students with no migration background exhibited a slightly higher mathematics ability than students with no information (main effect = -0.29

logits, Cohen's $d = -0.26$). There was also a considerable difference between students with migration background and students with no information (main effect = 0.60 logits, Cohen's $d = 0.51$), that reflected better mathematics competencies for students with no migration background information. DIF between 0.40 and 0.60 logits occurred for the items mag5d041_sc8g5_c, mag5q292_sc8g5_c, mag5d051_sc8g5_c, and mag5r201_sc8g5_c. DIF over 0.60 logits occurred for the items mag5v271_sc8g5_c, mag5q121_sc8g5_c, and mag5d02s_sc8g5_c. The largest difference in item difficulties was 0.95 logits (mag5v271_sc8g5_c). Although this appears to be a rather large difference, it may be a consequence of the uncertainty in the estimation due to the small number of students with a migration background. Furthermore, the items with over 0.60 logits DIFs showed no noticeable DIF in Starting Cohort 3 Grade 5 (Duchhardt & Gerdes, 2012). However, the overall tests for DIF using the BIC also favored the main effects models that did not include item-level DIF (Table 7).

Number of Books: The number of books at home was used as a proxy for cultural capital. There were 1,476 (27.28%) students with up to 100 books at home, 1,796 (33.19%) students with more than 100 books at home, and 2,139 (39.53%) students with no information on the number of books at home ("n.a."). All three groups were used to investigate DIF of books at home. Students with more than 100 books at home performed on average 0.71 logits (Cohen's $d = 0.66$) higher on the mathematics test than students with up to 100 books. Only a small difference between the mathematics competencies was observed between students with up to 100 books and students with no information (main effect = 0.11 logits, Cohen's $d = 0.11$). On average, students with more than 100 books at home performed higher than students with no information (main effect = -0.59 logits, Cohen's $d = -0.54$). All items in these group analyses for the variable Books had DIFs under 0.40 logits. Correspondingly, the overall tests for DIF using the BIC also favored the main effects models that did not include item-level DIF (Table 7). Thus, no severe DIF was observed for the variable books at home.

School type: 2,843 (52.54%) of the participants were high school students ("Gymnasium"), whereas 2,568 (47.46%) attended different types of school. On average, high school students showed a higher mathematical competence than the other students (main effect = 1.24, Cohen's $d = 1.31$). There was no item with a considerable DIF. This main effect was identically found in Starting Cohort 3 in Grade 7 (Schnittjer & Gerken, 2017). Correspondingly, the overall test for DIF using the AIC or BIC slightly favored the main effects model with DIF over the model that did not include item-level DIF (Table 7). As no items showed DIF over 0.40 logits, we did not include the effect in the scaling of the test.

Test position: The mathematical competence test was administered in two different positions (first vs. second; see Section 2.3 for the design of the study). There were 2,645 (48.88%) respondents who received the mathematics test first before the reading competence test, and 2,766 (51.12%) respondents who received the mathematics test in second position after the reading test. Participants were randomly assigned to one of the two groups. The results show an almost non-existent main effect of test position (main effect = 0.08, Cohen's $d = 0.07$). There was no DIF over 0.40 logits. Also, the overall test for DIF using the AIC or BIC (see Table 7) showed a better fit for the model that ignored potential DIF.

In summary, based on the DIF results, we assume measurement invariance for all tested variables.

Table 6: Differential Item Functioning

Item	Gender	Migration			Number of Books			School Type	Rotation
	male vs. female	without vs. with	without vs. n.a.	with vs. n.a.	< 100 vs. > 100	< 100 vs. n.a.	> 100 vs. n.a.	other type vs. high school	second vs. first
mag5d041_sc8g5_c	0.38 (0.34)	-0.43 (-0.39)	-0.27 (-0.24)	0.16 (0.13)	0.32 (0.30)	0.03 (0.02)	-0.30 (-0.27)	0.11 (0.12)	0.08 (0.07)
mag5q291_sc8g5_c	-0.61 (-0.54)	0.11 (0.10)	0.26 (0.23)	0.07 (0.06)	0.01 (0.01)	-0.11 (-0.11)	-0.13 (-0.12)	0.23 (0.25)	-0.07 (-0.06)
mag5q292_sc8g5_c	-0.23 (-0.20)	0.49 (0.45)	0.35 (0.32)	-0.27 (-0.23)	-0.11 (-0.11)	-0.19 (-0.18)	-0.08 (-0.07)	0.04 (0.04)	-0.08 (-0.07)
mag5v271_sc8g5_c	-0.18 (-0.16)	0.95 (0.87)	0.21 (0.19)	-0.75 (-0.63)	-0.23 (-0.22)	0.03 (0.03)	0.26 (0.24)	-0.44 (-0.46)	0.10 (0.09)
mag5r171_sc8g5_c	0.25 (0.23)	0.17 (0.15)	-0.02 (-0.02)	-0.27 (-0.23)	-0.09 (-0.09)	-0.01 (-0.01)	0.08 (0.07)	0.05 (0.05)	-0.13 (-0.11)
mag5q231_sc8g5_c	-0.11 (-0.10)	0.02 (0.02)	-0.03 (-0.03)	-0.08 (-0.07)	0.01 (0.00)	0.02 (0.02)	0.02 (0.02)	-0.07 (-0.07)	0.03 (0.03)
mag5q301_sc8g5_c	0.16 (0.14)	0.14 (0.13)	0.10 (0.09)	0.30 (0.26)	0.21 (0.20)	0.18 (0.16)	-0.04 (-0.03)	0.02 (0.02)	0.00 (0.00)
mag5q221_sc8g5_c	-0.26 (-0.23)	-0.23 (-0.21)	-0.25 (-0.22)	-0.10 (-0.08)	-0.08 (-0.07)	0.01 (0.01)	0.09 (0.08)	0.20 (0.22)	-0.18 (-0.16)
mag5d051_sc8g5_c	-0.01 (-0.01)	-0.44 (-0.41)	-0.29 (-0.26)	0.07 (0.06)	0.19 (0.17)	0.04 (0.04)	-0.15 (-0.13)	0.63 (0.67)	0.15 (0.14)
mag5d052_sc8g5_c	0.19 (0.17)	-0.39 (-0.36)	-0.17 (-0.15)	0.19 (0.16)	0.27 (0.25)	0.09 (0.08)	-0.18 (-0.17)	0.21 (0.23)	0.02 (0.02)
mag5q140_sc8g5_c	0.23 (0.20)	-0.29 (-0.26)	-0.08 (-0.07)	0.19 (0.16)	0.11 (0.10)	0.12 (0.11)	0.01 (0.01)	0.38 (0.40)	-0.06 (-0.06)
mag5q121_sc8g5_c	0.00 (0.00)	0.62 (0.57)	0.13 (0.12)	0.14 (0.12)	-0.11 (-0.10)	-0.04 (-0.04)	0.07 (0.06)	-0.14 (-0.15)	0.09 (0.08)
mag5r101_sc8g5_c	-0.12 (-0.11)	0.37 (0.34)	0.05 (0.05)	-0.43 (-0.36)	-0.15 (-0.14)	-0.02 (-0.02)	0.13 (0.12)	-0.60 (-0.64)	-0.12 (-0.11)
mag5r201_sc8g5_c	0.25 (0.22)	-0.46 (-0.42)	-0.15 (-0.14)	0.30 (0.26)	-0.02 (-0.02)	0.03 (0.03)	0.05 (0.05)	-0.05 (-0.05)	-0.11 (-0.09)
mag5q131_sc8g5_c	-0.67 (-0.59)	0.35 (0.32)	0.15 (0.13)	-0.32 (-0.27)	-0.10 (-0.09)	-0.17 (-0.16)	-0.07 (-0.06)	-0.05 (-0.05)	-0.01 (-0.01)
mag5d02s_sc8g5_c	0.25 (0.23)	-0.65 (-0.59)	-0.21 (-0.19)	0.41 (0.35)	0.22 (0.20)	0.11 (0.11)	-0.10 (-0.09)	0.31 (0.33)	0.07 (0.06)
mag5d023_sc8g5_c	-0.17 (-0.15)	0.08 (0.07)	-0.06 (-0.05)	-0.19 (-0.16)	-0.08 (-0.08)	0.01 (0.01)	0.09 (0.08)	-0.21 (-0.22)	0.04 (0.03)
mag5v024_sc8g5_c	-0.07 (-0.07)	0.12 (0.11)	0.13 (0.12)	-0.05 (-0.04)	-0.08 (-0.07)	0.11 (0.10)	0.19 (0.17)	-0.22 (-0.23)	0.00 (0.00)

Item	Gender	Migration				Books		School Type	Rotation
	male vs. female	without vs. with	without vs. n.a.	with vs. n.a.	< 100 vs. > 100	< 100 vs. n.a.	> 100 vs. n.a.	other type vs. high school	second vs. first
mag5r251_sc8g5_c	0.22 (0.20)	-0.07 (-0.07)	0.06 (0.05)	0.11 (0.09)	0.03 (0.03)	0.16 (0.15)	0.14 (0.12)	0.01 (0.02)	0.04 (0.04)
mag5v01s_sc8g5_c	0.45 (0.40)	0.06 (0.06)	-0.01 (-0.01)	-0.14 (-0.12)	-0.01 (-0.01)	0.08 (0.08)	0.09 (0.09)	-0.17 (-0.18)	0.07 (0.06)
mag5v321_sc8g5_c	-0.18 (-0.16)	-0.14 (-0.13)	-0.17 (-0.15)	0.05 (0.04)	-0.06 (-0.05)	-0.16 (-0.15)	-0.11 (-0.10)	-0.25 (-0.27)	0.04 (0.03)
mag5v071_sc8g5_c	0.19 (0.17)	-0.23 (-0.21)	0.07 (0.07)	0.25 (0.21)	-0.20 (-0.18)	-0.27 (-0.25)	-0.07 (-0.07)	0.00 (0.00)	-0.05 (-0.04)
mag5r191_sc8g5_c	0.36 (0.32)	-0.14 (-0.13)	0.11 (0.09)	0.24 (0.21)	-0.03 (-0.03)	0.10 (0.09)	0.13 (0.12)	0.00 (0.00)	-0.09 (-0.08)
mag5v091_sc8g5_c	-0.32 (-0.29)	0.00 (0.00)	0.08 (0.07)	0.10 (0.09)	-0.01 (-0.01)	-0.13 (-0.12)	-0.13 (-0.11)	-0.01 (-0.01)	0.16 (0.15)
Main effect (DIF model)	-0.24 (-0.22)	-0.68 (-0.62)	-0.29 (-0.26)	0.60 (0.51)	0.71 (0.66)	0.11 (0.11)	-0.59 (-0.54)	1.24 (1.31)	0.08 (0.07)
Main effect (Main effect model)	-0.20 (-0.18)	-0.70 (-0.64)	-0.30 (-0.27)	0.58 (0.49)	0.70 (0.65)	0.13 (0.12)	-0.58 (-0.53)	1.20 (1.27)	0.08 (0.07)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. n.a. = not available answers.

Table 7: Comparison of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	129,931.8	30	129,991.8	130,189.7
	DIF	129,436.8	53	<i>129,542.8</i>	<i>129,892.4</i>
Migration without vs. with	Main effect	117,329.3	30	117,389.3	<i>117,584.1</i>
	DIF	117,237.1	53	<i>117,343.1</i>	117,687.2
Migration without vs. n.a.	Main effect	124,428.0	30	124,488.0	<i>124,684.6</i>
	DIF	124,381.4	53	<i>124,487.4</i>	124,834.7
Migration with vs. n.a.	Main effect	17,917.1	30	<i>17,977.1</i>	<i>18,116.3</i>
	DIF	17,874.5	53	17,980.5	18,226.5
Books < 100 vs. > 100	Main effect	80,776.3	30	80,836.3	<i>81,019.1</i>
	DIF	80,716.0	53	<i>80,822.0</i>	81,144.9
Books < 100 vs. n.a.	Main effect	86,488.1	30	<i>86,548.1</i>	<i>86,733.9</i>
	DIF	86,448.7	53	86,554.7	86,883.0
Books > 100 vs. n.a.	Main effect	92,034.9	30	92,094.9	<i>92,283.2</i>
	DIF	91,972.2	53	<i>92,078.2</i>	92,410.9
School Type	Main effect	128,539.7	30	128,599.7	128,797.6
	DIF	128,260.4	53	<i>128,366.4</i>	<i>128,716.0</i>
Rotation	Main effect	130,001.4	30	<i>130,061.4</i>	<i>130,259.2</i>
	DIF	129,963.3	53	130,069.3	130,418.9

Note. The AIC and BIC values of the best-fitting model are shown in italics.

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model and also the partial credit model (PCM; Masters, 1982) is that all item discrimination parameters are equal. To test this assumption of Rasch-homogeneity, we also fitted a GPCM (Muraki, 1992) to the data. The estimated discrimination parameters are depicted in Table 4a ("Discr."). They varied between 0.74 (mag5v01s_sc8g5_c) and 1.90 (mag5d051_sc8g5_c). The average discrimination parameter was 1.18 ($SD = 0.26$). Low discrimination parameters below 0.50 (de Ayala, 2009) were not observed for any item.

In addition, a visual inspection of the respective item characteristic curve of the PCM indicated an adequate fit. However, model fit indices suggested a better model fit of the GPCM (AIC = 129,481.5, BIC = 129,811.3, number of parameters = 50) as compared to the PCM (AIC = 129,929.3, BIC = 130,107.4, number of parameters = 27). Despite the empirical preference for the GPCM, the PCM 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 PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

4.3.5 Unidimensionality

The unidimensionality of the test was investigated by specifying a four-dimensional model based on the four different content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, the Quasi-Monte Carlo estimation implemented in R in the package “TAM” was used. The number of nodes per dimension was chosen as 10,000 to enable stable parameter estimation. The variances, correlations, and EAP reliabilities of the four dimensions are shown in Table 8. All four dimensions exhibited substantial variances. The correlations among the four dimensions were rather high and varied between 0.91 and 0.95. However, the AIC and BIC favored the four-dimensional model (Table 9). Additionally, for the unidimensional model, the average absolute residual correlations as indicated by the adjusted Q_3 statistic (Table 4a) were quite low ($M = .08$, $SD = .01$) — the largest individual residual correlation was .12 — and, thus, indicated an essentially unidimensional test. Because the mathematics test was constructed to measure a single dimension and the correlations between the dimensions were rather high, a unidimensional mathematics competence score was estimated.

Table 8: Results of Four-Dimensional Scaling

	Quantity	Change and relationships	Data and chance	Space and shape
Quantity (8 items)	1.29			
Change and relationships (6 items)	.95	1.14		
Data and chance (5 items)	.91	.94	1.39	
Space and shape (5 items)	.91	.95	.94	1.09
EAP reliability	.79	.80	.78	.77

Note. Variances of the dimensions are given 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	130,189.7	27	130,243.7	130,421.8
Four-dimensional	129,713.1	36	<i>129,785.1</i>	<i>130,022.6</i>

Note. The AIC and BIC values of the best-fitting model are shown in italics.

5 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test which was administered in Starting Cohort 8 of the NEPS (Wave 1) with students in Grade 5. We investigated different kinds of missing responses and examined the item and test parameters to check the quality of the test. We conducted further quality inspections by examining differential item functioning, testing Rasch-homogeneity, and investigating the tests' dimensionality.

The different kinds of missing responses were evaluated, and all kinds of missing responses were rather low and acceptable. As indicated by various fit criteria (WMNSQ, t -value of the WMNSQ, ICC), the items exhibited a good item fit. The range of the item difficulties was good, and the item distribution along the ability scale was satisfactory. Although more difficult items in the mathematical test were needed. As a consequence, ability estimates will be precise for low-performing students but less precise for high-performing students. The reliability was good with .82. Furthermore, different variables were used for testing measurement invariance (DIF). Some items exceeded 0.60 logits DIF, but the overall model tests favored the models without DIF, except for the model comparison for the variable school type. We assumed that no considerable DIF became evident for these variables, indicating that the test was fair for the examined subgroups. Moreover, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with the total score) were good. The high correlations between the four dimensions indicate that the four content areas measure a common construct, although the model indices AIC and BIC favored the four-dimensional model. In sum, the mathematical competence 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

In the data set, 24 items are either scored as dichotomous variables (MC and SCR items), with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (corresponding to the CMC item), indicating the number of correctly answered subtasks. The dichotomous variables are marked with a '_c' behind their variable name, whereas the polytomous variables are marked with a 's_c' at the end of the variable's names (CMC items and SCR items that take two answers into account). In the scaling model, the collapsed polytomous variable mag5v01s_sc8g5_c is scored in steps of 0.5 – 0 for the lowest category and 1.5, denoting the highest (see Appendix B). All items were already administered in other waves in the NEPS and kept their original names. However, for reasons of identification, a suffix was added to specify the current test administration ('sc8g5' referring to Starting Cohort 8, Wave 1). Further details on the naming conventions of the variables can be found in Fuß et al. (2025).

6.2 Linking of the Competence Test

The same mathematical competence test was administered in Starting Cohort 3 (Duchhardt & Gerdes, 2012) and Starting Cohort 8 in Grade 5. Therefore, $N = 5,193$ students from Starting Cohort 3 were used to evaluate whether the item difficulties were comparable between the two cohorts. As shown in Table 10, the mean DIF effect (i.e., the difference in item difficulties)

was zero. The highest item DIF was -0.41 logits for item mag5d052_c. No DIF effects were observed for the remaining items. However, an overall test for DIF suggested a better fit for the model with DIF (AIC = 262,776.9, BIC = 263,162.1, number of parameters = 53) compared to a model without DIF (AIC = 263,011.4, BIC = 263,229.5, number of parameters = 30). Finally, the main effect between starting cohorts was -0.03 logits (Cohen's $d = -0.03$), suggesting hardly any differences between the starting cohorts.

Table 10: Differential Item Functioning Analyses between the Starting Cohort 3 and Starting Cohort 8

No.	Item	Estimate	No.	Item	Estimate
1	mag5d041_c	-0.26 (-0.25)	13	mag5r101_c	-0.04 (-0.04)
2	mag5q291_c	0.04 (0.04)	14	mag5r201_c	0.10 (0.09)
3	mag5q292_c	0.15 (0.14)	15	mag5q131_c	0.19 (0.18)
4	mag5v271_c	-0.18 (-0.17)	16	mag5d02s_c	0.29 (0.27)
5	mag5r171_c	0.06 (0.05)	17	mag5d023_c	0.18 (0.17)
6	mag5q231_c	-0.04 (-0.03)	18	mag5v024_c	-0.10 (-0.09)
7	mag5q301_c	0.18 (0.17)	19	mag5r251_c	-0.09 (-0.09)
8	mag5q221_c	0.07 (0.07)	20	mag5v01s	0.01 (0.01)
9	mag5d051_c	-0.34 (-0.32)	21	mag5v321_c	-0.09 (-0.08)
10	mag5d052_c	-0.41 (-0.38)	22	mag5v071_c	-0.07 (-0.07)
11	mag5q140_c	-0.07 (-0.06)	23	mag5r191_c	0.15 (0.14)
12	mag5q121_c	0.05 (0.05)	24	mag5v091_c	0.21 (0.20)
Main effect (DIF model)		-0.03 (-0.03)	Main effect (Main effect model)		-0.03 (-0.03)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's d) in parentheses.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores in the form of WLEs (mag5_sc1), along with their corresponding standard errors "mag5_sc2". The person abilities were estimated using fixed item parameters from Starting Cohort 3 (Duchhardt & Gerdes, 2012). Because no DIF between the cohorts was observed, all items were constrained. The WLE scores were also corrected for the position of the mathematical competence test within the test battery. As a result, the WLE scores in Starting Cohort 8 are linked to the scale of Starting Cohort 3, thus allowing comparisons between cohorts. The R Syntax for estimating the WLE

scores from the items is provided in Appendix B. Students who did not take part in the test or those who 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.

Alternatively, users interested in examining latent relationships can either include the measurement model in their analyses or estimate plausible values. Plausible values for the reading competence test can be estimated using the R package *NEPSscaling*¹ (Scharl et al., 2020; Scharl & Zink, 2022).

¹ <https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values>

7 References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Publications.
- Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for mathematics: Scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper No. 19). University of Bamberg. <https://doi.org/10.5157/NEPS:WP19:1.0>
- 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). Waxmann.
- Fuß, D., Gnambs, T., Lockl, K., Attig, M., & Nusser, L. (2025). *Competence data in NEPS: Overview of Measures and Variable Naming Conventions (Starting Cohorts 1 to 6), Revised Version 2025*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/NEPS_OverviewCompetenceData_2025_en.pdf
- Gnambs, T. (2025). *NEPS technical report for reading: Scaling results of Starting Cohort 8 for Grade 5* (NEPS Survey Paper No. 117). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP117:1.0>
- Haberkorn, K., Pohl, S., & Carstensen, C. (2016). Incorporating different response formats of competence test in an IRT model. *Psychological Test and Assessment Modeling*, 58, 223-252.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/BF02296272>
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. <https://doi.org/10.1177/014662169201600206>
- 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. <https://doi.org/10.25656/01:8426>
- Petersen, L. A., & T., Beyer (2025). *NEPS Technical Report for Mathematics -Scaling Results of Starting Cohort 1 for Ten-Year-Old Children (Wave 11)* (NEPS Survey Paper No. 121). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP121:1.0>
- Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Otto-Friedrich-Universität, Nationales Bildungspanel. <https://doi.org/10.5157/NEPS:WP14:1.0>
- 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. <https://doi.org/10.25656/01:8430>

- R Core Team (2025). R: A language and environment for statistical computing (Version 4.5.0) [Computer software]. <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).
- Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules* (Version 4.2-21) [Computer software]. <https://doi.org/10.32614/CRAN.package.TAM>
- Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6* (NEPS Survey Paper No. 71). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP71:1.0>
- Scharl, A., & Zink, E. (2022). NEPSScaling: Plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-Scale Assessments in Education*, 10. <https://doi.org/10.1186/s40536-022-00145-5>
- Schnittjer, I., & Gerken, A.-L. (2017). *NEPS Technical Report for mathematics: Scaling results of Starting Cohort 3 in grade 7* (NEPS Survey Paper No. 16). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP16:1.0>
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464. <https://doi.org/10.1214/aos/1176344136>
- 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. <https://nbn-resolving.org/urn:nbn:de:gbv:lue4-opus-144121>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450. <https://doi.org/10.1007/BF02294627>
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2019). Development of competencies across the life span. In H. P. Blossfeld & H. G. Roßbach (Eds.), *Edition ZfE. Education as a lifelong process. (Vol 3, pp. 57-81)*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-23162-0_4
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145. <http://doi.org/10.1177/014662168400800201>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>

Appendix A. Content Areas of Items in Starting Cohort 8 Grade 5

Item number	Position in Test	Item	Content area
1	1	mag5d041_sc8g5_c	Data and chance
2	2	mag5q291_sc8g5_c	Quantity
3	3	mag5q292_sc8g5_c	Quantity
4	4	mag5v271_sc8g5_c	Change and relationships
5	5	mag5r171_sc8g5_c	Space and shape
6	6	mag5q231_sc8g5_c	Quantity
7	7	mag5q301_sc8g5_c	Quantity
8	8	mag5q221_sc8g5_c	Quantity
9	9	mag5d051_sc8g5_c	Data and chance
10	10	mag5d052_sc8g5_c	Data and chance
11	11	mag5q140_sc8g5_c	Quantity
12	12	mag5q121_sc8g5_c	Quantity
13	13	mag5r101_sc8g5_c	Space and shape
14	14	mag5r201_sc8g5_c	Space and shape
15	15	mag5q131_sc8g5_c	Quantity
16	16	mag5d02s_sc8g5_c	Data and chance
17	17	mag5d023_sc8g5_c	Data and chance
18	18	mag5v024_sc8g5_c	Change and relationships
19	19	mag5r251_sc8g5_c	Space and shape
20	20	mag5v01s_sc8g5_c	Change and relationships
21	21	mag5v321_sc8g5_c	Change and relationships
22	22	mag5v071_sc8g5_c	Change and relationships
23	23	mag5r191_sc8g5_c	Space and shape
24	24	mag5v091_sc8g5_c	Change and relationships

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).

Appendix B. R Syntax for Estimating linked WLEs

```
library(haven) # contains read_sav function for loading the data
library(TAM) # contains tam.mml and tam.wle functions
library(doby) # to recode variables

## load data
dat <- read_sav(file = "SC8_xTargetCompetencies.sav")

## item names for scaling (24 items)
items <- c("mag5d041_sc8g5_c", "mag5q291_sc8g5_c", "mag5q292_sc8g5_c",
"mag5v271_sc8g5_c", "mag5r171_sc8g5_c", "mag5q231_sc8g5_c", "mag5q301_sc8g5_c",
"mag5q221_sc8g5_c", "mag5d051_sc8g5_c", "mag5d052_sc8g5_c", "mag5q140_sc8g5_c",
"mag5q121_sc8g5_c", "mag5r101_sc8g5_c", "mag5r201_sc8g5_c", "mag5q131_sc8g5_c",
"mag5d02s_sc8g5_c", "mag5d023_sc8g5_c", "mag5v024_sc8g5_c", "mag5r251_sc8g5_c",
"mag5v01s_sc8g5_c2", "mag5v321_sc8g5_c", "mag5v071_sc8g5_c", "mag5r191_sc8g5_c",
"mag5v091_sc8g5_c")

## collapse response categories of polytomous item mag5v01s_sc8g5_c
dat$mag5v01s_sc8g5_c <- doBy::recodeVar(dat$mag5v01s_sc8g5_c, 0:4, c(0, 0, 1, 2, 3))

## define vector with fixed item estimates from Starting Cohort 3 (Duchhardt & Gerdes,
2012)
xsi <- c(-0.370, -1.040, -0.799, 0.922, -0.082, 0.478, 0.603, -1.855, -2.471, -0.500, -0.721,
1.514, -0.117, -1.225, -1.401, -2.094, -0.428, -0.127, 0.219, -0.588, 0.977, -2.549,
-0.271, 0.333)
xsi <- cbind(seq_along(xsi), xsi)

## define Q-matrix for 0.5 scoring of PCM
Q = matrix(1, nrow=24, ncol=1)
Q[20, 1] <- 0.5

## estimate partial credit model
model <- tam.mml(resp = dat[, items], pid = dat$ID_t, Q = Q, irtmodel = "PCM2",
xsi.fixed = xsi)
summary(model)

## item fit
tam.fit(model)

## WLE
WLE = tam.wle(model)
```