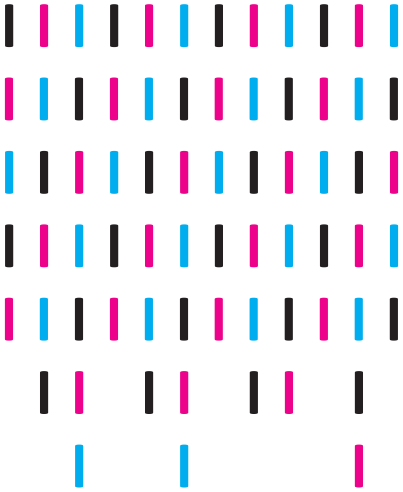

NEPS *SURVEY PAPERS*

Anna-Lena Kock, Kristin Litteck, and Lara Aylin Petersen

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING KOHORT 2 IN SEVENTH GRADE

NEPS *Survey Paper* No. 83
Bamberg, March 2021

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).

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.

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 Seventh Grade

Anna-Lena Kock, Kristin Litteck, and Lara Aylin Petersen
Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the lead author:

alkock@leibniz-ipn.de

Bibliographic Data:

Kock, A.-L., Litteck, K., & Petersen, L. A. (2021): *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 2 in Seventh Grade* (NEPS Survey Paper No. 83). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP83:1.0

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.

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 in Seventh Grade

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 paper describes the data and scaling procedure for the mathematical competence test in grade 7 of starting cohort 2 (kindergarten). The mathematics test consists of 28 items that represent different content areas as well as different cognitive components and use different response formats. The test was administered to 2,616 students. 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, good item fit, and that the items satisfactorily fitted the model. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were some recognizable gaps at the upper end of the scale's item difficulties. Overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. 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 longitudinal linking parameters.

Keywords

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

Content

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, 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 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).

In this paper, the results of these analyses are presented for mathematical competence in grade 7 (ninth wave) of starting cohort 2 (kindergarten). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of the ninth wave 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 (SUF) is presented.

The present report has been modeled on previous reports (Pohl et al., 2012; Haberkorn et al., 2016). 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 paper. 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 Weinert et al. (2011), Neumann et al. (2013), and Ehmke et al. (2009). In the following, specific aspects of the mathematics test will be pointed out 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 faced a certain situation followed by a single task related to it; in one instance there were two tasks. Each item 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 were distributed across the items.

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, usually four, available response options. In CMC items a number of subtasks with two response options were presented. SCR items required the test taker to write down an answer into an empty box.

3. Data

3.1 The Design of the Study

The study was conducted in 2018/19 and assessed different competence domains including scientific literacy, reading competence, and mathematical competence. Each student was individually tested at home and received two of the three tests, with the test domains being assigned randomly. To control for the effect of test position, the students received the mathematics test first as in previous studies of this starting cohort (see Table 1). To measure the students' competence with great accuracy, the tests for mathematical competence and reading competence were available in two difficulty levels. The students were assigned either to the easy or the difficult mathematics test based on their estimated mathematics competence in the previous assessment in grade 4 (Schnittjer et al., 2020). Both mathematics tests consisted of 21 items that represented different content-related and process-related components and used different response formats (see Table 2). There were 14 common items in the two mathematics tests.

Table 1

Design of the study

Position	Competence domain
1	Mathematics easy/difficult or Science
2	Reading easy/difficult or Science

Overall, 28 different items with different response formats were used. The characteristics of the 28 items are depicted in the following tables. Table 2 shows the distribution of the four content areas (see Appendix A for the assignment of the items to the content areas), whereas Table 3 shows the distribution of the response formats. One SCR item (mag4q060_sc2g7_c) had several subtasks but was scored dichotomously because the subtasks are interdependent. One subtask of the CMC item mag7r02s_sc2g7_c was excluded from the analyses due to an unsatisfactory item fit (see Section 4.3.1).

Table 2

Number of Items by Content Areas

Content area	Frequency
Quantity	8
Space and shape	6
Change and relationships	8
Data and chance	6
Total number of items	28

Table 3

Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	22
Complex Multiple-Choice	2
Short Constructed Response	4
Total number of items	28

3.2 Sample

A total of 2,616 students received the mathematics test. For one respondent less than three valid responses were available. Because no reliable ability scores can be estimated based on such few responses, this case was excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 2,615 test-takers. Of these, 1,026 students received the easy test, whereas 1,589 students received the difficult test version. 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.

Invalid responses occurred, for example, when two response options were selected where only one was required. Omitted items occurred when test-takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response were coded as not-reached. Because of the two difficulty levels, some items were not administered to all students. 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 for 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 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 items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. Categories of polytomous variables with less than $N = 200$ responses were collapsed to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items. For item mag7d06s_c the lowest three categories had to be collapsed and for item mag7r02s_sc2g7_c the lowest two categories were collapsed as in previous studies (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 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 is described in section 6.

3.5 Checking the Quality of the Scale

The mathematics test was specifically constructed to be implemented in the NEPS. To ensure appropriate psychometric properties, the quality of the test was examined in several analyses. All analyses were conducted for the whole test and the different booklets, respectively.

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, they were used to construct the polytomous CMC variables that were included in the final scaling model.

The MC items consisted of one correct response option and three or four distractors (i.e., incorrect response options). The quality of the distractors within MC items, that is, whether they were chosen by students with lower ability rather than by those with higher ability, was evaluated using the point-biserial correlation between selecting an incorrect response option 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 a number. All SCR items were scored dichotomously even if there was more than one response required.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC items, the polytomous CMC items, and the 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*-value| > 6) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (|*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 0.30 were considered as good, greater than 0.20 as acceptable, and below 0.20 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 (i.e., 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, migration background, the HISEI (Highest International Socio-Economic Index of Occupational Status), and school type (see Pohl & Carstensen, 2012, for a description of these variables). Moreover, differential item functioning (DIF) was also examined for the administered test version. 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 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, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

The competence data in NEPS are scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To

test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The mathematics test was constructed to measure a unidimensional competence score. The assumption of unidimensionality 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, TAM in R was used. To ensure that the results are comparable with those of 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 (15,000 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.

We ran all analyses separately for the two booklets and with the combined data. Because the analyses for both booklets showed good fit, only the analyses of the combined data are presented here.

3.6 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams et al., 2015). The generalized partial credit model and the multi-dimensional model were estimated in TAM version 3.5-19 (Robitzsch et al., 2020) in R version 4.0.2 (R Core Team, 2020).

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, 93.1 % of test-takers gave no invalid response at all. Only 0.3 % of the respondents had more than one invalid response.

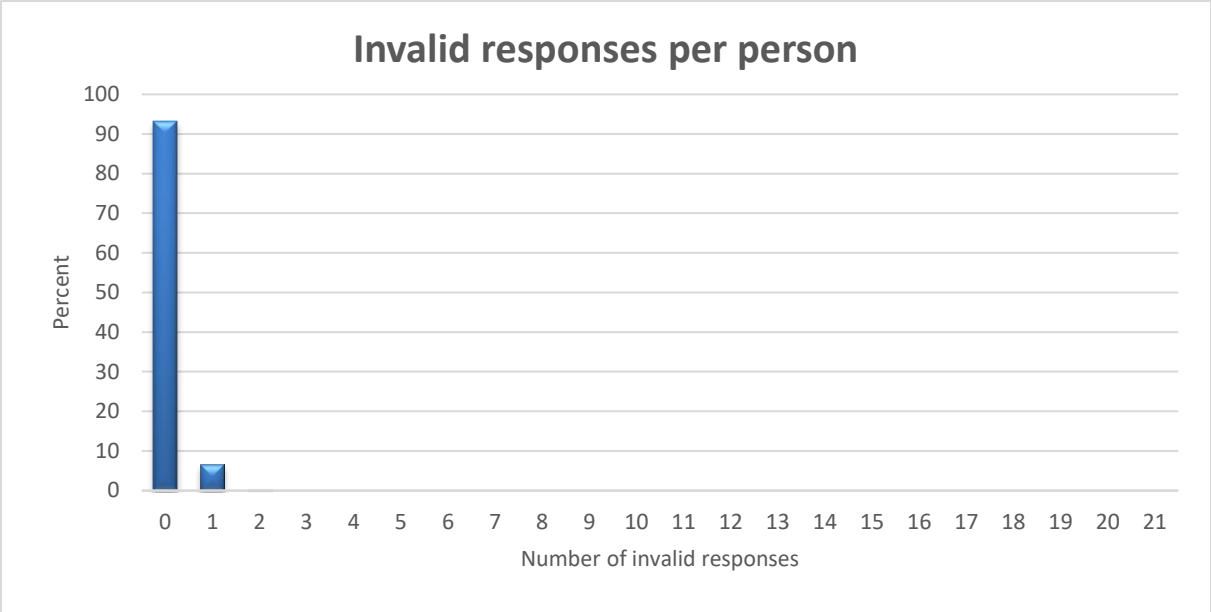

Figure 1. Number of invalid responses

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 49.3 % of the respondents omitted no item, whereas 2.9 % of the respondents omitted more than 5 items.

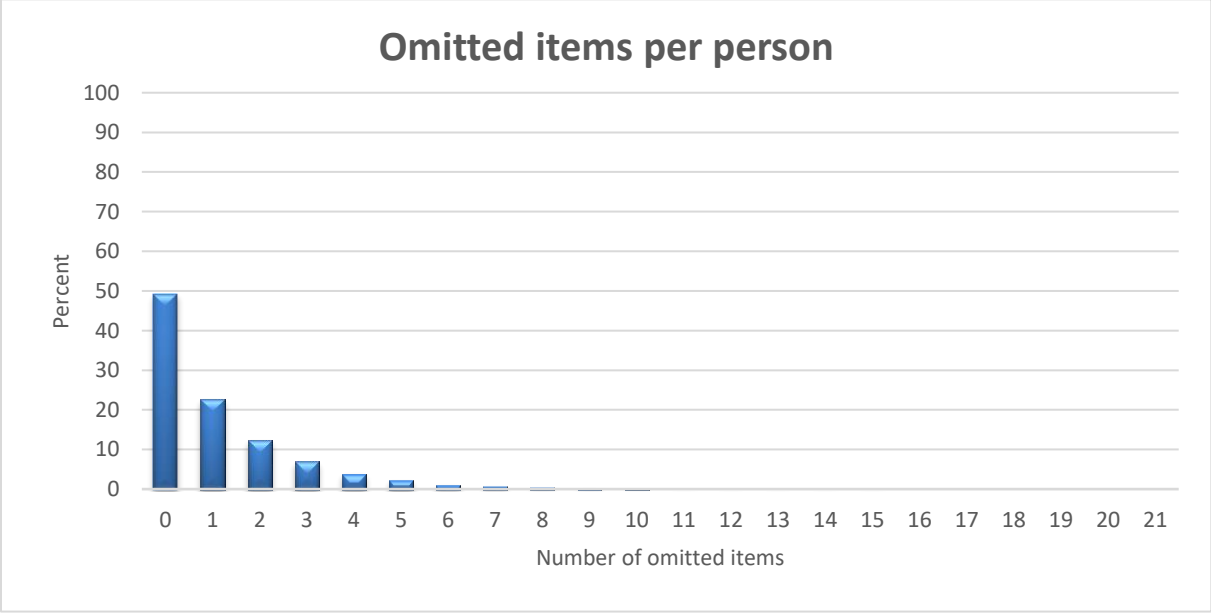

Figure 2. Number of omitted items.

All missing responses after the last valid response are defined as not reached. Figure 3 shows the number of items that were not reached by a person. As can be seen, 59.4 % reached the end of the test, whereas 23.1 % of the test takers did not reach one to five items. 17.6 % of students did not reach more than five items.

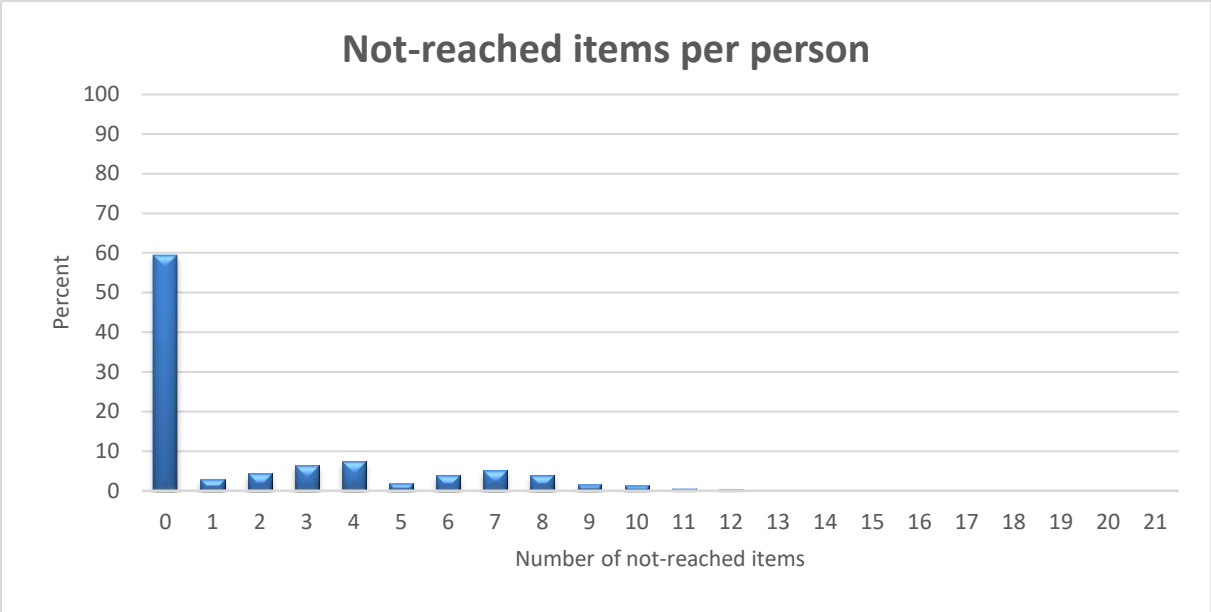

Figure 3. Number of not-reached items.

Figure 4 shows the total number of missing responses per person, which is the sum of invalid, omitted, and not-reached missing responses. In total, 35.3 % of the test takers showed no missing response, whereas 26.9 % showed more than five missing responses.

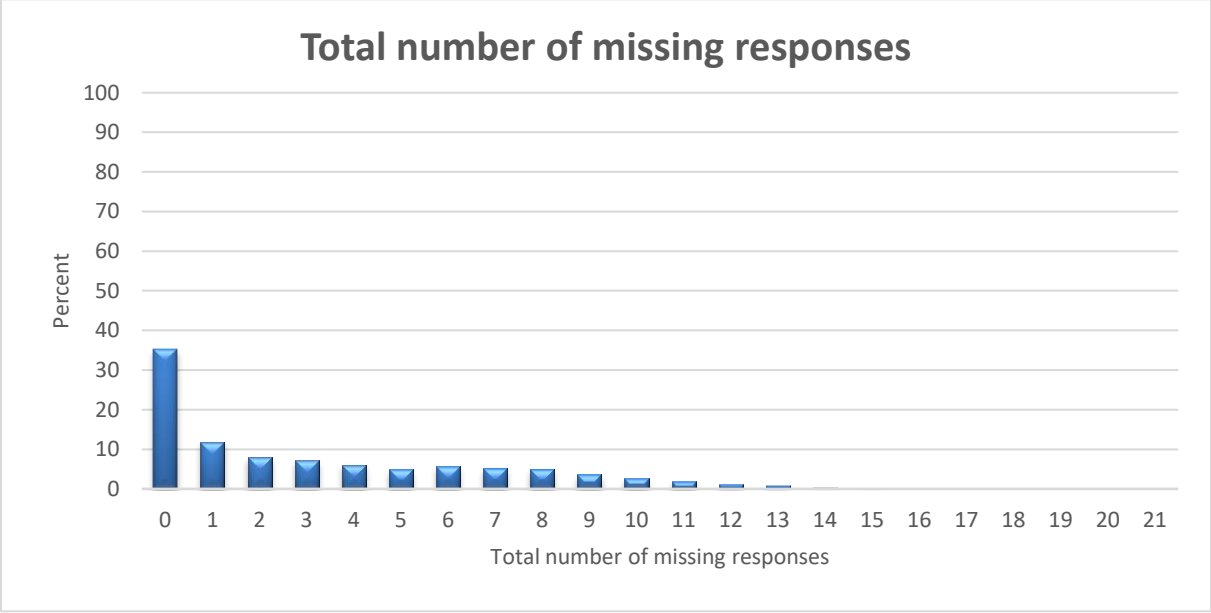

Figure 4. Total number of missing responses.

In sum, the amount of invalid and omitted missing responses is acceptably small. The number of not reached items is, however, rather large and has the greatest impact on the total number of missing responses.

4.1.2 Missing responses per item

Tables 4 and 5 show the number of valid responses for each item in the two booklets, as well as the percentage of missing responses. Overall, the number of omitted responses per item was small, varying between 0.82 % (items mag9d151_sc2g7_c and mag5q301_sc2g7_c, difficult booklet) and 13.55 % (items mag7d061_sc2g7_c and mag4v111_sc2g7_c, easy booklet), except for one item that had an omission rate of 17.05% (item mag9v091_sc2g7_c, difficult booklet). The number of persons that did not reach an item increased with the position of the item in the test up to 29.34 % for the easy booklet and up to 47.95 % for the difficult booklet. The percentage of invalid responses varied from 0.00 % (various items in both booklets) to 5.75 % (mag5r191_sc2g7_c, easy booklet). Multiple missings only occurred for item mag7d06s_c from the easy booklet (0.10 %) and item mag4q060_sc2g7_c from the difficult booklet (0.06 %).

Table 4

Percentage of Missing Values for the Easy Booklet

Item position	Item	Number of valid respon-ses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items	Percentage of multiple missings
1	mag7q011_c	1,008	0.00	1.36	0.00	0.00
2	mag7d061_sc2g7_c	886	0.10	13.55	0.00	0.00
3	mag7r071_c	946	0.00	7.80	0.00	0.00
4	mag5v271_sc2g7_c	911	0.00	11.21	0.00	0.00
5	mag4q011_sc2g7_c	991	0.10	3.31	0.00	0.00
6	mag7r081_sc2g7_c	993	0.29	2.92	0.00	0.00
7	mag7v031_sc2g7_c	977	0.00	4.78	0.00	0.00
8	mag7d06s_c	967	0.10	5.46	0.10	0.10
9	mag5q301_sc2g7_c	995	0.10	2.73	0.19	0.00
10	mag7v021_c	971	0.00	4.78	0.58	0.00
11	mag7r02s_sc2g7_c	983	0.10	3.22	0.88	0.00
12	mag4q060_sc2g7_c	894	3.70	7.89	1.27	0.00
13	mag4d031_sc2g7_c	964	0.29	4.09	1.66	0.00
14	mag9q181_sc2g7_c	982	0.00	0.88	3.41	0.00
15	mag4v111_sc2g7_c	808	0.49	13.55	7.21	0.00
16	mag7q041_sc2g7_c	898	0.00	2.73	0.00	0.00
17	mag7d042_sc2g7_c	896	0.00	1.07	11.60	0.00
18	mag5r251_sc2g7_c	797	0.00	4.39	17.93	0.00
19	mag7d031_c	757	0.10	3.51	22.61	0.00
20	mag5v321_sc2g7_c	676	0.19	7.02	26.51	0.00
21	mag5r191_sc2g7_c	666	5.75	0.00	29.34	0.00

Table 5

Percentage of Missing Values for the Difficult Booklet

Item position	Item	Number of valid respon-ses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items	Percentage of multiple missings
1	mag7q041_c	1,459	0.00	8.18	0.00	0.00
2	mag7d061_sc2g7_c	1,415	0.06	10.89	0.00	0.00
3	mag7r071_c	1,512	0.00	4.85	0.00	0.00
4	mag5v271_sc2g7_c	1,387	0.00	12.71	0.00	0.00
5	mag4q011_sc2g7_c	1,546	0.00	2.71	0.00	0.00
6	mag9v091_sc2g7_c	1,315	0.00	17.05	0.19	0.00
7	mag7r081_sc2g7_c	1,532	0.19	3.08	0.31	0.00
8	mag9d151_sc2g7_c	1,566	0.13	0.82	0.50	0.00
9	mag5q301_sc2g7_c	1,565	0.00	0.82	0.69	0.00
10	mag9v121_sc2g7_c	1,485	0.00	5.35	1.20	0.00
11	mag7r091_sc2g7_c	1,518	0.06	2.52	1.89	0.00
12	mag4q060_sc2g7_c	1,397	4.34	3.78	3.90	0.06
13	mag4d031_sc2g7_c	1,437	0.00	3.34	6.23	0.00
14	mag7q051_c	1,241	0.06	10.07	11.77	0.00
15	mag4v111_sc2g7_c	1,103	0.82	12.15	17.62	0.00
16	mag7q041_sc2g7_c	1,185	0.00	2.77	22.66	0.00
17	mag7d042_sc2g7_c	1,177	0.00	1.20	24.73	0.00
18	mag5r251_sc2g7_c	1,024	0.00	2.71	32.85	0.00
19	mag7d031_c	895	0.00	3.40	40.28	0.00
20	mag7v071_sc2g7_c	837	0.06	2.39	44.87	0.00
21	mag5r191_sc2g7_c	822	0.31	0.00	47.95	0.00

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 21.42 % and 92.00 % across all items. On average, the rate of correct responses was 57.70 % (*SD* = 19.31 %).

From a descriptive point of view, the items covered a wide range of difficulties. The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variables) are depicted in Table 6a. The step parameters for polytomous variables are presented in Table 6b. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.29 (mag7d042_sc2g7_c) and 1.62 (mag9v121_sc2g7_c) with a mean of -0.28. Due to the large sample size, the standard errors of the estimated item difficulties (Table 6a, column 5) were small (*SE*(ß) ≤ 0.11).

Table 6a

Item Parameters

Pos.	Item	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	aQ_3
1	mag7q011_c	72.02	-1.70	0.08	1.04	1.2	0.38	0.83	0.03
2	mag7d061_sc2g7_c	43.16	0.33	0.05	1.10	5.6	0.38	0.64	0.04
3	mag7r071_c	43.53	0.34	0.05	0.93	-4.0	0.55	1.43	0.03
4	mag5v271_sc2g7_c	55.79	-0.31	0.05	0.97	-1.8	0.52	1.27	0.03
5	mag4q011_sc2g7_c	43.59	0.32	0.05	0.97	-1.6	0.51	1.17	0.03
6	mag7r081_sc2g7_c	46.85	0.15	0.05	1.04	2.2	0.44	0.89	0.03
7	mag7v031_sc2g7_c	49.95	-0.58	0.08	1.05	1.9	0.43	0.89	0.02
8	mag7d06s_c	n.a.	0.36	0.10	0.90	-3.0	0.52	1.01	0.03
9	mag5q301_sc2g7_c	57.66	-0.37	0.05	0.91	-5.2	0.57	1.63	0.04
10	mag7v021_c	21.42	0.92	0.09	1.01	0.2	0.39	1.01	0.03
11	mag7r02s_sc2g7_c	n.a.	-1.48	0.11	1.03	0.7	0.28	0.42	0.02
12	mag4q060_sc2g7_c	34.19	0.98	0.05	1.05	2.2	0.39	0.80	0.03
13	mag4d031_sc2g7_c	59.81	-0.51	0.05	1.03	1.5	0.45	0.94	0.03
14	mag9q181_sc2g7_c	77.29	-2.04	0.09	0.93	-1.6	0.47	1.62	0.05
15	mag4v111_sc2g7_c	32.44	0.83	0.06	0.99	-0.5	0.47	1.08	0.03
16	mag7q041_sc2g7_c	68.22	-1.04	0.06	0.95	-2.3	0.51	1.42	0.03
17	mag7d042_sc2g7_c	86.16	-2.29	0.07	1.02	0.4	0.33	0.96	0.02
18	mag5r251_sc2g7_c	61.18	-0.68	0.06	1.05	2.3	0.42	0.87	0.03
19	mag7d031_c	35.71	0.55	0.06	0.95	-2.2	0.52	1.31	0.04
20	mag5v321_sc2g7_c	32.84	0.15	0.09	1.00	0.1	0.43	1.07	0.03

21	mag5r191_sc2g7_c	75.87	-1.57	0.07	0.95	-1.4	0.46	1.31	0.04
22	mag7q041_c	68.13	-0.51	0.07	1.11	4.2	0.32	0.48	0.04
23	mag9v091_sc2g7_c	48.14	0.45	0.07	0.93	-3.3	0.54	1.49	0.04
24	mag9d151_sc2g7_c	87.74	-1.91	0.09	0.95	-0.8	0.39	1.44	0.03
25	mag9v121_sc2g7_c	25.93	1.62	0.07	0.97	-0.8	0.45	1.25	0.03
26	mag7r091_sc2g7_c	70.42	-0.66	0.07	0.96	-1.4	0.49	1.36	0.03
27	mag7q051_c	44.64	0.58	0.07	1.10	4.5	0.35	0.58	0.03
28	mag7v071_sc2g7_c	50.78	0.22	0.08	1.04	1.7	0.40	0.81	0.04

Note. Pos. = Item position in the test. 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, *aQ3* = adjusted average absolute residual correlation for item (Yen, 1993).
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).

Table 6b

Step Parameters (with Standard Errors) of Polytomous Items

Item	step 1	step 2
mag7d06s_c	-0.22 (0.07)	0.22
mag7r02s_sc2g7_c	-1.18 (0.07)	1.18

Note. The last step parameter is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

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 5, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The respective difficulties ranged from -2.29 (item mag7d042_sc2g7_c) to 1.62 (item mag9v121_sc2g7_c). Therefore, a rather broad range was spanned. However, there was just one very difficult item. As a consequence, subjects with a low or medium ability will be measured relatively precisely, while subjects with a high mathematical competence will have a larger standard error. The variance was estimated to be 1.150, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = 0.769, WLE reliability = 0.737) was good.

Scale in logits	Person ability	Item difficulty
3		
	X	
	X	
	X	
	X	
	X	
2	X	
	XX	
	XXX	25
	XXX	
	XXXXXX	
	XXXXX	
	XXXXX	
1	XXXXXXX	10 12
	XXXXXXX	15
	XXXXXXXXX	27
	XXXXXXXXX	19 23
	XXXXXXXXXX	2 3 5 8
	XXXXXXXXX	6 20 28
	XXXXXXXXX	
0	XXXXXXXXX	
	XXXXXXXX	
	XXXXXXXX	4 9
	XXXXXXX	13 22
	XXXXXXXX	7 18 26
	XXXXXXXX	
-1	XXXXXXX	
	XXXXXX	16
	XXXX	
	XXXX	
	XXX	11 21
	XXX	1
	XX	
-2	XX	24
	X	14
	X	
	X	17
-3	X	

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 14.7 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 6a).

4.3 Quality of the test

4.3.1 Fit of the subtasks of complex multiple-choice items

Before the responses to the subtasks of the CMC items 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 in a Rasch model. Counting the subtasks of CMC items separately, there were 35 variables in total.

The rates of correct responses given to the subtasks of the CMC items varied from 43.98 % to 92.00 %. With one exception, the subtasks showed a good item fit with the WMNSQ ranging between 0.89 and 1.15 and the respective *t*-values between -3.1 and 5.5. Only one subtask of the item mag7r02s_sc2g7_c exhibiting unsatisfactory item fit (WMNSQ of 1.26, *t*-value of 8.6 and a respective item discrimination of -0.11) and was excluded from further analyses. The good model fit of the other subtasks justified their aggregation to polytomous variables for both items mag7d06s_c and mag7r02s_sc2g7_c.

4.3.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating – for the MC items – the point-biserial correlation between each incorrect response (distractor) and the students' total correct scores. This distractor analysis was performed based on preliminary analyses treating all subtasks of the CMC item as single items.

Table 7 shows a summary of point-biserial correlations between correct and incorrect responses and the number correct scores for MC items (only the items where subjects were asked to choose between distractors). The point-biserial correlations for the distractors ranged from -0.32 to 0.09 with a mean of -0.14. Although some distractors 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 0.18 to 0.41 with a mean of 0.31 indicating that more proficient students were also more likely to identify the correct response option.

Table 7

Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.31	-0.14
Minimum	0.18	-0.32
Maximum	0.41	0.09

4.3.3 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 6a). The values of the WMNSQ were close to 1 with the lowest value being 0.90 (mag7d06s_c) and the highest being 1.11 (mag7q041_c). 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 0.28 (mag7r02s_sc2g7_c) and 0.57 (mag5q301_sc2g7_c). Overall, the items showed an average correlation of 0.44.

4.3.4 Differential item functioning

We examined test fairness for several subgroups (i.e., measurement invariance) by estimating differential item functioning (DIF). DIF was investigated for the variables gender, migration background, school type (see Pohl & Carstensen, 2012, for a description of these variables), as well as for the HISEI and the difficulty of the booklet. Table 8 shows the difference between the estimated difficulties of the items in different subgroups. For example, the column "female versus male" indicates the difference in difficulty *ß*(male) – *ß*(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females.

Gender: Overall, 1,281 (49.0 %) of the test takers were female, 1,258 (48.1 %) were male, and 76 (2.9 %) did not give a valid response. On average, male students exhibited a higher mathematical competence than female students (main effect = -0.35 logits, Cohen's *d* = 0.33). There was one item (mag7r02s_sc2g7_c) with a considerable gender DIF above 0.6 logits. DIF exceeding 0.4 logits occurred for the items mag7v021_c, mag5r191_sc2g7_c, mag9v091_sc2g7_c, and mag9v121_sc2g7_c.

Migration: There were 1,765 (67.5 %) participants without a migration background, 772 (29.5 %) participants with a migration background, and 78 (3.0 %) students that gave no valid answer. On average, participants without migration background performed considerably better in the mathematics test than those with a migration background (main effect = 0.24 logits, Cohen's *d* = 0.23). There was no item with DIF exceeding 0.4 logits.

HISEI: The HISEI was calculated for the whole starting cohort 2 and divided in two categories (lower and higher HISEI) using a median-split. Overall, 1,049 (40.1 %) of the test takers were assigned to a lower HISEI whereas 1,471 (56.3 %) of the test takers were assigned to a higher HISEI, and for 95 (3.6 %) students no assignment could be calculated. Students with a higher HISEI performed better than persons with a lower HISEI (main effect = 0.66 logits, Cohen's *d* = 0.65). There was no item with DIF exceeding 0.4 logits.

School: Overall, 1,661 students (63.5 %) who took the mathematics test attended secondary school (German: "Gymnasium") whereas 757 (28.9 %) were enrolled in other school types. Subjects in secondary schools showed a higher mathematics competence on average (main effect = 0.69 logits, Cohen's *d* = 0.68) than subjects in other school types. There was no item with DIF exceeding 0.4 logits.

Booklet: To estimate the participants' proficiency with greater accuracy, the participants received different test versions with low or high difficulty (see section 3.1 for the design of the study). The booklets shared a subset of 14 items. For these common items, we examined potential DIF across the respective versions. A subsample of 1,026 (39.2 %) students received the easy test and 1,589 (60.8 %) persons received the difficult test. Subjects who were administered the difficult test scored on average 1.33 logits (Cohen's d = 1.56) higher on the common items than subjects who received the easy test. There was no noticeable DIF for the common items concerning the test version. The largest difference in difficulties between the two groups was 0.46 logits (item mag5q301_sc2g7_c).

Table 8

Differential Item Functioning

Pos.	**Item**	**Gender**	**Migration status**	**HISEI**	**School**	**Booklet**
		male vs. female	with vs. without	low vs. high	no sec. vs sec.	easy vs. difficult
1	mag7q011_c	-0.23	-0.07	0.18	0.06	
2	mag7d061_sc2g7_c	0.02	-0.03	-0.25	-0.29	-0.32
3	mag7r071_c	-0.09	0.03	0.22	0.19	0.37
4	mag5v271_sc2g7_c	-0.15	-0.06	0.02	-0.16	0.11
5	mag4q011_sc2g7_c	0.06	0.12	0.03	0.16	0.01
6	mag7r081_sc2g7_c	0.19	0.00	0.10	-0.05	-0.21
7	mag7v031_sc2g7_c	0.12	-0.26	-0.27	-0.22	
8	mag7d06s_c	-0.19	0.10	0.20	0.04	
9	mag5q301_sc2g7_c	0.38	0.20	0.22	0.25	0.46
10	mag7v021_c	-0.44	0.33	0.09	0.04	
11	mag7r02s_sc2g7_c	-0.61	0.11	0.10	-0.02	
12	mag4q060_sc2g7_c	-0.32	0.04	-0.20	-0.10	-0.19
13	mag4d031_sc2g7_c	0.08	-0.08	-0.10	-0.08	-0.11
14	mag9q181_sc2g7_c	-0.17	0.33	-0.02	0.02	
15	mag4v111_sc2g7_c	0.12	-0.36	-0.13	-0.10	-0.28
16	mag7q041_sc2g7_c	-0.10	-0.12	0.11	0.06	0.02
17	mag7d042_sc2g7_c	0.32	-0.13	-0.21	-0.27	-0.23
18	mag5r251_sc2g7_c	0.29	-0.14	0.08	0.11	-0.13
19	mag7d031_c	-0.38	0.11	0.13	0.30	0.29
20	mag5v321_sc2g7_c	0.05	0.14	0.24	0.16	
21	mag5r191_sc2g7_c	0.57	0.29	0.03	0.27	0.11

22	mag7q041_c	0.26	0.00	-0.33	-0.29	
23	mag9v091_sc2g7_c	-0.46	-0.11	-0.06	0.12	
24	mag9d151_sc2g7_c	-0.38	0.36	0.31	0.13	
25	mag9v121_sc2g7_c	-0.40	-0.06	-0.33	-0.17	
26	mag7r091_sc2g7_c	-0.02	-0.11	0.33	0.19	
27	mag7q051_c	0.10	-0.16	-0.20	-0.32	
28	mag7v071_sc2g7_c	0.35	0.07	-0.17	-0.10	
Main effect (DIF model)		**-0.34**	**0.24**	**0.65**	**0.70**	**1.33**
Main effect (Main effect model)		**-0.35**	**0.24**	**0.66**	**0.69**	**1.33**

Overall, test fairness could be confirmed for all tested subgroups. In Table 9, we compared the models that only included the main effects to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for the variables gender and HISEI. The variables migration status and school favored models estimating only the main effect.

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 all four variables were preferred over the more complex DIF models.

Table 9

Comparison of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	54,345.31	33	54,411.30	*54,604.01*
	DIF	54,208.17	61	*54,330.16*	54,686.38
Migration status	Main effect	54,334.22	33	*54,400.22*	*54,592.90*
	DIF	54,292.11	61	54,414.11	54,770.28
HISEI	Main effect	53,757.84	33	53,823.84	*54,016.30*
	DIF	53,693.85	61	*53,815.85*	54,171.60

School	Main effect	51,483.16	33	*51,549.16*	*51,740.25*
	DIF	51,429.60	61	51,551.60	51,904.83
Booklet	Main effect	35,628.78	16	35,660.78	*35,754.69*
	DIF	35,551.34	30	*35,611.34*	35,787.41

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

4.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption of Rasch-homogeneity, we also fitted a generalized partial credit model (GPCM; Muraki, 1992) to the data. The estimated discrimination parameters are depicted in Table 6a ("Discr."). They varied between 0.42 (item mag7r02s_sc2g7_c) to 1.63 (item mag5q301_sc2g7_c). The average discrimination parameter fell at 1.07. Model fit indices suggested a slightly better model fit of the generalized partial credit model (AIC = 55,896.34, BIC = 56,260.22, number of parameters = 62) as compared to the Rasch model (AIC = 56,129.12, BIC = 56,379.93, number of parameters = 32). Despite the empirical preference for the generalized partial credit model, the Rasch 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 Rasch model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

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 15,000 nodes.

The variances, correlations, and EAP Reliability of the four dimensions are shown in Table 10. All four dimensions exhibited a substantial variance. The correlations among the four dimensions were rather high and varied between 0.917 and 0.954. Moreover, the AIC and BIC favored the unidimensional model (Table 11). Additionally, for the unidimensional model the average absolute residual correlations as indicated by the adjusted Q_3 statistic (Table 6a) were quite low (M = .03, SD = .01) — the largest individual residual correlation was .05 — and, thus, indicated an essentially unidimensional test. Because the mathematics test was constructed to measure a single dimension, a unidimensional mathematics competence score was estimated.

Table 10

Results of Four-Dimensional Scaling

	Quantity	**Space and shape**	**Change and Relationship**	**Data and chance**
Quantity (8 items)	(1.195)			
Space and shape (6 items)	0.931	(1.319)		
Change and relationships (8 items)	0.932	0.940	(1.399)	
Data and chance (6 items)	0.954	0.931	0.917	(1.085)
EAP Reliabilty	0.751	0.741	0.739	0.737

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

Table 11

Comparison of the Unidimensional and the Four-Dimensional Model

Model	**Deviance**	**Number of parameters**	**AIC**	**BIC**
Unidimensional	56,128.12	32	*56,192.12*	*56,379.93*
Four-dimensional	56,112.75	41	56,194.75	56,435.38

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 in the seventh grade of 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, the amount of not-reached items was rather high (only 59,35 % reached the end of the test), indicating that the test had too many items for the allocated testing time. Other types of missing responses were acceptably small. 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. The item distribution along the ability scale was good, except for some gaps at the upper end of the scale. Nevertheless, the test had a good reliability and distinguished well between test-takers, as indicated by the test's variance. Moreover, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with the total score) were acceptable. The high correlations

between the four dimensions as well as a lower AIC and BIC indicated that the unidimensional model described the data reasonably well. Different variables were used for testing measurement invariance. Only one item (mag7r02s_sc2g7_c) of the test showed a considerable DIF for the variable gender that slightly exceeded 0.6 logits (see 4.3.4). In sum, the analyses indicated that the test was fair for the examined subgroups.

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

6. Data in the Scientific Use File

6.1 Naming conventions

There are 28 items in the data set that 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 items) indicating the number of correctly answered subtasks. The dichotomous variables are marked with a '_c' at the end of the variable name; the polytomous variables are marked with a 's_c' or 's_sc2g7_c' behind their variable names. Items that were already administered in other grades kept their original names ('mag5v271…', 'mag4q011…', 'mag5q301…', 'mag4q060…', 'mag4d031…', 'mag9q181…', 'mag4v111…', 'mag5r251…', 'mag5v321…', 'mag5r191…', 'mag9v091…', 'mag9d151…', and 'mag9v121…'). However, for reasons of identification a suffix was added in front of the '…_c' to specify the current test administration ('sc2g7' referring to Starting Cohort 2, Grade 7).

6.2 Linking of competence scores

In starting cohort 2, the mathematics competence tests administered in kindergarten, grade 1, grade 2, grade 4, and grade 7 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 compared directly; 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 et al. (2016). The process of linking combines adjacent measurement points on the same scale. Therefore, the first wave of each competence scale within a cohort is used as a reference scale that all subsequent measurement waves will refer to. For the domain of mathematical competence, linking is achieved using overlapping items (also known as common items). For the linking procedure of mathematical competence across kindergarten and grade 1 see Schnittjer and Fischer (2018), across grade 1 and grade 2 see Schnittjer and Gerken (2018), and across grade 2 and grade 4 see Schnittjer et al. (2020).

To link the test of mathematics competence conducted in grade 4 and grade 7, seven items that already were administered in grade 4 were, again, administered in grade 7. An empirical study that evaluated different link methods concerning 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. Six of the seven common items that were

administered in grade 4 and grade 7 were found to be measurement invariant across the two measurement points. Therefore, they served as link items and the anchor-items design as described in Fischer et al. (2016) was used. 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 longitudinal subsample of 2,450 students participated at both measurement occasions (in grade 4 and grade 7). Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016).

6.2.2 Results

To examine whether the two tests administered in the longitudinal sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model. For the two-dimensional model, the common items load on the first dimension and the unique items (i.e., the items included in only one test) load on the second dimension. In both grades, the information criteria slightly favored the two-dimensional model over the one-dimensional model (see Table 12). We also examined the residual correlations for the one-dimensional models. The corrected absolute Q_3 statistics indicated largely unidimensional scales in grade 4 ($M(aQ_3) = 0.00$, $SD(aQ_3) = 0.03$), and grade 7 ($M(aQ_3) = 0.03$, $SD(aQ_3) = 0.02$). This indicates that unidimensional scales can be assumed for the mathematics tests in both grades, although the model test slightly favored the two-dimensional model.

Table 12

Comparison of the Unidimensional and the two-Dimensional Model

Grade	Model	Deviance	Number of parameters	AIC	BIC
Grade 4	Unidimensional	137,415.33	25	137,465.33	137,634.96
	Two-dimensional	137,365.73	27	*137,419.73*	*137,602.93*
Grade 7	Unidimensional	55,163.08	31	55,225.08	*55,406.50*
	Two-dimensional	55,158.16	33	*55,224.16*	55,417.29

Note. The results in this table were achieved by using ConQuest 4.2.5. The AIC and BIC values of the best fitting model are shown in italics.

Table 13

DIF Analyses for the common items used for linking in the tests for mathematical competence in grades 4 and 7

Grade 4	Grade 7	Δσ	$SE_{\Delta\sigma}$	*t*	*F*
mag5v271_sc2g4_c	mag5v271_sc2g7_c	0.04	0.08	0.56	0.31
mag4q011_c	mag4q011_sc2g7_c	0.28	0.08	3.45	11.92
mag5q301_sc2g4_c	mag5q301_sc2g7_c	-0.07	0.07	-1.00	0.99
mag4q060_c	mag4q060_sc2g7_c	0.22	0.10	2.26	5.11
mag4d031_c	mag4d031_sc2g7_c	-0.36	0.07	-4.85	23.54
mag4v111_c	mag4v111_sc2g7_c	-0.11	0.10	-1.11	1.24

Note. Δσ = Difference in item difficulty parameters between grades 4 and 7 (positive values indicate easier items in grade 4); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis; F_{crit} = Critical value for the minimum effects hypothesis test for an α of .05; the degrees of freedom (df_1, df_2) are based on the number of measurement points (df_1 = *k*-1) and the number of test-takers taking both tests (df_2 = n-1). The critical *F* (1, 2449) = 61.54. A non-significant test indicates measurement invariance.

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 comparing grade 4 and grade 7. The differences in item difficulties between the link subsample grade 4 and link subsample grade 7 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 13. Analyses of differential item functioning identified no common items with significant (α = .05) DIF (difference in logits: *Min* = -0.368, *Max* = 0.300).

In the longitudinal subsample, the mean of the item difficulty parameters for the six common items used for linking was 1.349 in grade 4 and 0.124 in grade 7. Mean/mean linking (Loyd & Hoover, 1980) resulted in a correction term of $c_{4\text{-}7}$ = 1.349 - 0.124 = 1.225. The correction term for linking kindergarten to grade 4 was $c_{KG\text{-}4}$ = 4.620 (Schnittjer et al., 2020). The sum of the correction terms $c_{KG\text{-}4}$ + $c_{4\text{-}7}$ = 5.845 was added to each item difficulty parameter derived in grade 7. The linked item parameters can be seen in Appendix C. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.095 and has to be included in the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs (“mag7_sc1” and “mag7_sc1u”) including their respective standard errors (“mag7_sc2” and “mag7_sc2u”). For “mag7_sc1u”, person abilities were estimated using the linked item difficulty parameters. As a result, the WLE scores provided in “mag7_sc1u” can be used for longitudinal comparisons between kindergarten, grades 1, 2, 4, and 7. The resulting differences in WLE scores can be interpreted as development trajectories across

measurement points. In contrast, the WLE scores in “mag7_sc1” are not linked to the underlying reference scale of kindergarten. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions.

The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix B, the fixed linked item parameters for estimating the uncorrected WLE scores are provided in Appendix C. 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 [Computer Software].

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

http://doi.org/10.1007/978-1-4612-1694-0_16

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, Germany: 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.

https://www.neps-data.de/Portals/0/Survey%20Papers/SP_I.pdf

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/Overview_NEPS_Competence-Data.pdf

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.

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

https://doi.org/10.1007/978-1-4757-4310-4

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

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

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://www.pedocs.de/volltexte/2013/8426/pdf/JERO_2013_2_Neumann_et_al_Modeling_and_assessing_mathematical_competencies.pdf

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.

https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

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://www.pedocs.de/volltexte/2013/8430/pdf/JERO_2013_2_Pohl_Carstensen_Scaling_of_competence_tests.pdf

R Core Team (2020). R: A language and environment for statistical computing (Version 3.2.4) [Software]. Retrieved from https://www.R-project.org/.

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

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test Analysis Modules*. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 3.5-19).

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. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XLVI.pdf

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

https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XLVII.pdf

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.

https://www.neps-data.de/Portals/0/Survey%20Papers/SP_LXIX.pdf

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

https://doi.org/10.1214/aos%2F1176344136

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://pub-data.leuphana.de/frontdoor/index/index/docId/776

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. (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.

https://doi.org/10.1007/s11618-011-0182-7

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

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

Appendix

Appendix A. Content Areas of Items in the Mathematics Test for Grade 7

Position	Item	Content area
1	mag7q011_c	Quantity
2	mag7d061_sc2g7_c	Data and chance
3	mag7r071_c	Space and shape
4	mag5v271_sc2g7_c	Change and relationships
5	mag4q011_sc2g7_c	Quantity
6	mag7r081_sc2g7_c	Space and shape
7	mag7v031_sc2g7_c	Change and relationships
8	mag7d06s_c	Data and chance
9	mag5q301_sc2g7_c	Quantity
10	mag7v021_c	Change and relationships
11	mag7r02s_sc2g7_c	Space and shape
12	mag4q060_sc2g7_c	Quantity
13	mag4d031_sc2g7_c	Data and chance
14	mag9q181_sc2g7_c	Quantity
15	mag4v111_sc2g7_c	Change and relationships
16	mag7q041_sc2g7_c	Quantity
17	mag7d042_sc2g7_c	Data and chance
18	mag5r251_sc2g7_c	Space and shape
19	mag7d031_c	Data and chance
20	mag5v321_sc2g7_c	Change and relationships
21	mag5r191_sc2g7_c	Space and shape
22	mag7q041_c	Quantity
23	mag9v091_sc2g7_c	Change and relationships
24	mag9d151_sc2g7_c	Data and chance
25	mag9v121_sc2g7_c	Change and relationships
26	mag7r091_sc2g7_c	Space and shape
27	mag7q051_c	Quantity
28	mag7v071_sc2g7_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. ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort II - Grade 7

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


data filename.dat;
format pid 1-7 responses 9-36; /* insert number of columns with data*/
labels << labels.nam;


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


recode (0,1,2,3,4,5)   (0,0,0,0,1,2)   !item (8);      /* collapsing the lowest 4 categories */
recode (0,1,2,3)       (0,0,1,2)       !item (11);     /* collapsing the lowest 2 categories */

score (0,1,2)          (0,0.5,1)       !item (8,11);
score (0,1)            (0,1)           !item (1-7,9-10,12-28);


model item + item*step;
set constraint=cases;
estimate;

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

Appendix C. Original and linked item difficulties for the mathematics test in Grade 7.

	item	Common item	Original item difficulties	Linked item difficulties
1	mag7q011_c	no	-1.73	4.12
2	mag7d061_sc2g7_c	no	0.32	6.17
3	mag7r071_c	no	0.32	6.17
4	mag5v271_sc2g7_c	yes	-0.34	5.50
5	mag4q011_sc2g7_c	yes	0.28	6.13
6	mag7r081_sc2g7_c	no	0.12	5.96
7	mag7v031_sc2g7_c	no	-0.60	5.24
8	mag7d06s_c	no	0.33	6.18
9	mag5q301_sc2g7_c	yes	-0.42	5.43
10	mag7v021_c	no	0.92	6.77
11	mag7r02s_sc2g7_c	no	-1.50	4.35
12	mag4q060_sc2g7_c	yes	0.97	6.82
13	mag4d031_sc2g7_c	yes	-0.55	5.30
14	mag9q181_sc2g7_c	no	-2.04	3.81
15	mag4v111_sc2g7_c	yes	0.80	6.64
16	mag7q041_sc2g7_c	no	-1.06	4.78
17	mag7d042_sc2g7_c	no	-2.34	3.50
18	mag5r251_sc2g7_c	no	-0.68	5.16
19	mag7d031_c	no	0.54	6.38
20	mag5v321_sc2g7_c	no	0.12	5.97
21	mag5r191_sc2g7_c	no	-1.64	4.21
22	mag7q041_c	no	-0.51	5.34
23	mag9v091_sc2g7_c	no	0.43	6.27
24	mag9d151_sc2g7_c	no	-1.98	3.86
25	mag9v121_sc2g7_c	no	1.61	7.46
26	mag7r091_sc2g7_c	no	-0.72	5.12
27	mag7q051_c	no	0.57	6.42
28	mag7v071_sc2g7_c	no	0.21	6.05

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Table 6a). Linked item difficulty parameters were derived by adding $c_{KG\text{-}7}$ to the original item parameters.