NEPS *SURVEY PAPERS*

Lara Aylin Petersen, Tessa Beyer
and David Bednorz

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 1 FOR EIGHT-YEAR-OLD CHILDREN (WAVE 9)

NEPS *Survey Paper* No. 97
Bamberg, June 2022

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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 1 for Eight-Year-Old Children (Wave 9)

Lara Aylin Petersen, Tessa Beyer, David Bednorz

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

Email address of the lead author:

lpetersen@leibniz-ipn.de

Bibliographic Data:

Petersen, L. A., Beyer, T., & Bednorz, D. (2022). *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 1 for Eight-Year-Old Children (Wave 9)* (NEPS Survey Paper No. 97). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP97:1.0

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., Gnambs, 2022; Kock, Litteck, & Petersen, 2021).

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 1 for Eight-Year-Old Children (Wave 9)

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 for 8-year-old children of starting cohort 1 (newborns). The mathematics test consists of 20 items that represent different content areas as well as different cognitive components and use different response formats. The test was administered to 1,632 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 (EAP/PV reliability = .76) and that the items satisfactorily fitted the model. Furthermore, comparable measurements 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 R syntax for scaling the data.

Keywords

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

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured 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 Fuß et al. (2021) as well as Weinert et al. (2011).

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

In this paper, the results of these analyses are presented for mathematical competence for 8-year-old-children (ninth wave) of Starting Cohort 1 (newborns). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of the ninth wave of Starting Cohort 1 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 (Gnambs, 2022; Kock, Litteck, & Petersen, 2021; Pohl & Carstensen, 2012). 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 Ehmke et al. (2009), Neumann et al. (2013), and Weinert et al. (2011). 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. Each item belongs to one of the following content areas:

- sets, numbers, and operations,
- units and measuring,
- space and shape,
- change and relationships,
- data and chance.

Each item was constructed in such a way as to primarily address a specific content area (see Appendix A). 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 sorting (S). In MC items, the test taker had to find the correct response option from several, usually four (but sometimes more), available response options. In CMC items a number of subtasks with two response options were presented. In sorting items, the test taker had to put objects in the correct order (e.g., coins and numbers in ascending order).

3 Data and Psychometric Analyses

3.1 The Design of the Study

The study was conducted in summer 2020 and assessed different competence domains including reading speed, early reading competence, and mathematical competence. The test for mathematical competence was always presented third after reading speed and early reading competence (see Gnambs, 2022). Procedural metacognition was measured both after early reading competence and mathematic competence. There was no rotation design, thus, all students received the tests in the same order. A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

The mathematics test consisted of 20 items that represented different content-related and process-related components. Table 1 shows the distribution of the five content areas (see Appendix A for the assignment of the items to the content areas), whereas Table 2 shows the distribution of the three response formats.

Table 1.

Number of Items by Content Areas

Content area	**Frequency**
Sets, numbers, and operations	6
Units and measuring	3
Space and shape	3
Change and relationships Data	5
and chance	3
Total number of items	20

Table 2.

Number of Items by Response Formats

Response format	**Frequency**
Simple Multiple-Choice	15
Complex Multiple-Choice	4
Sorting	1
Total number of items	20

Initially, the study was conducted with a personal interview and computer-based testing using dedicated tablets in the student's household (proctored computerized test, CBT, comparably to previous assessments in Starting Cohort 1). Due to the beginning of the COVID-19 pandemic, only 34 participants could be surveyed this way. Hence, the administration mode had to be changed and was switched to a proctored web-based format (WBT). Here, the interviewer accompanied the computer-based testing via phone. The results reported in this technical report refer only to the students who were tested via WBT administration. The WBT procedure is described in more details below.

A couple of weeks before the test date a telephone interview was conducted with a parent to discuss the necessary computer equipment in the household that would allow the child to take the WBT. Although tablet devices were preferred (to keep as comparable as possible to the previous assessments), laptops with a minimum screen size were allowed as alternative assessment devices. At a prearranged test date and time, a trained test administrator called the parent by phone to assist in setting up the tablet or laptop (e.g., positioning the device on the table) and starting the web-based test (e.g., opening the browser, entering the correct link and password). Then, the children worked alone on the WBT. During the test administration, the test administrators supervised the child's progress on the test remotely using a dashboard that showed in real time the test page a child was currently visiting. Assistance and verbal support to the children were provided by phone. Thus, the test administrators had a continuous means of communication with the children during the entire test procedure. Although the test administrators could not directly see the child or the specific testing conditions, they could monitor the child's progress in the test, listen to voiced problems or background noise, and talk to the children. By design, direct assistance through test administrators was rarely required because the web-based test used standardized video instructions that introduced the different tests with prerecorded demonstrations and, thus, allowed a high level of standardization. The role of the test administrators was primarily limited to assisting in starting the test, motivating children between different tests, and helping with unforeseen problems during the test.

3.2 Sample

A total of 1,632 students received the mathematics test. For 120 respondents less than three valid responses were available (e.g., if serious problems were observed during the test administration like interference by a parent or technical problems). Because no reliable ability scores can be estimated based on such few responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper (see Chapter 4) are based on a sample of 1,512 students with 50.45 % girls. They were, on average, 8 years old and about 8.60 % of them had a migration background. 83.33 % of the students used a device with touch functionality (tablet or laptop with touch display) and 16.67 % of the students used a laptop or computer without touch functionality.

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) technical difficulties, and c) test abortion (not reaching all items).

Omitted items occurred when students skipped some items. Because the test was administered on the private devices of the students, unforeseen technical errors might have prevented the correct presentation of some items or the whole test (e.g., internet problems or technical problems with the device). In some rare cases, the test had to be aborted (e.g., due to problems that prevented the processing of further items). Therefore, not all items have valid responses. 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 (multiple missing). In this study, multiple missing within polytomous items were not observed.

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 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 2 % of the sample responses (N = 30) were collapsed. This procedure deviated slightly from previous technical reports that used N = 200 for category consolidation to avoid estimation problems because the sample size in the present study was rather small which would have required many categories to be collapsed, even though the polytomous variables did not show a misfit and the polytomous item construction was important in terms of item content. This usually occurred for the lower categories of polytomous items. For items man9d11s_c and mag2g12s_sc1n9_c the lowest three categories had to be collapsed.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items and Sorting 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. All analyses were conducted for the whole test and all students (i.e., both types of devices together).

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 Sorting 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), except for one item. Item man9g071_c which had one correct response and ten distractors (response on an 11-step measurement scale). 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 Sorting item required the test taker to sort objects by order. The Sorting item was scored dichotomously.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC items, the polytomous CMC items, and the Scoring item 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 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 (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, measurement bias was investigated for the variables sex, migration background, the HISEI (Highest International Socio-Economic Index of Occupational Status), and school month (before and first two weeks of summer holidays or last week of summer holidays and new school year). 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 five-dimensional model based on the five different content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, R was used (see Chapter 3.6). The number of nodes in the multidimensional model in R was chosen in such a way as to obtain stable parameter estimates (10,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

All IRT models were estimated with the TAM package version 3.7-16 (Robitzsch et al., 2021) in R version 4.1.3 (R Core Team, 2022).

4 Results

4.1 Missing Responses

4.1.1 Missing responses per person

Missing responses may occur when students skip (omit) some items. The total number of omitted responses is depicted in Figure 1. As can be seen, 71.76 % did not omit any of the items, while 19.64 % omitted one item. Two or more items were omitted by 8.60 %.

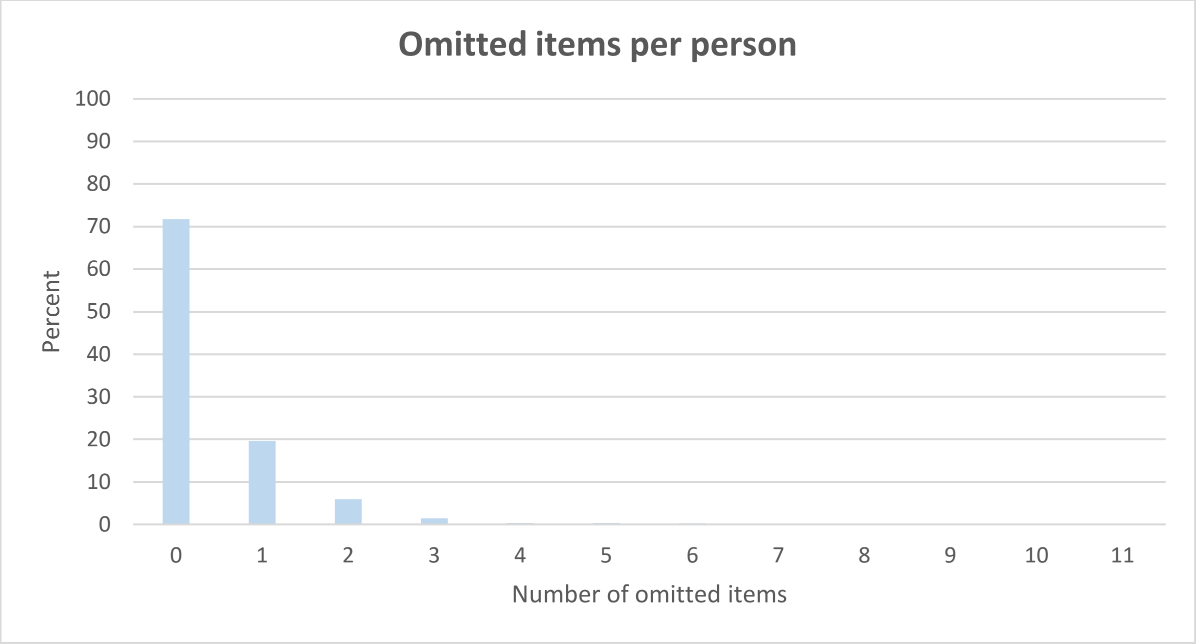

Figure 1. Number of omitted items

Furthermore, missing responses occurred due to technical difficulties. The total number of technical missing is depicted in Figure 2. As can be seen, 96.56 % had no technical difficulties. However, 2.45 % were unable to respond to one item due to technical difficulties. In 0.99 % of the cases, more than one item could not be answered for this reason.

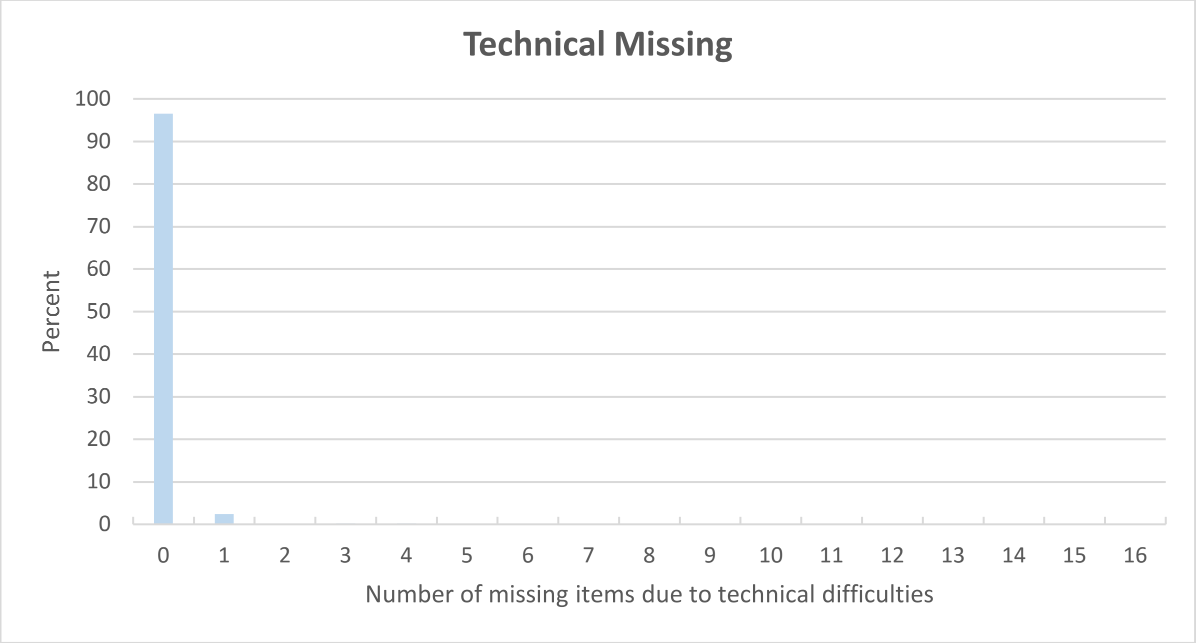

Figure 2. Number of technical missing items per person

In some cases, the test had to be aborted for various reasons (e.g. due to disturbance or persistent technical problems). Figure 3 shows the total number of missing items due to test abortion. In total, only 0.93 % of the tests had to be aborted.

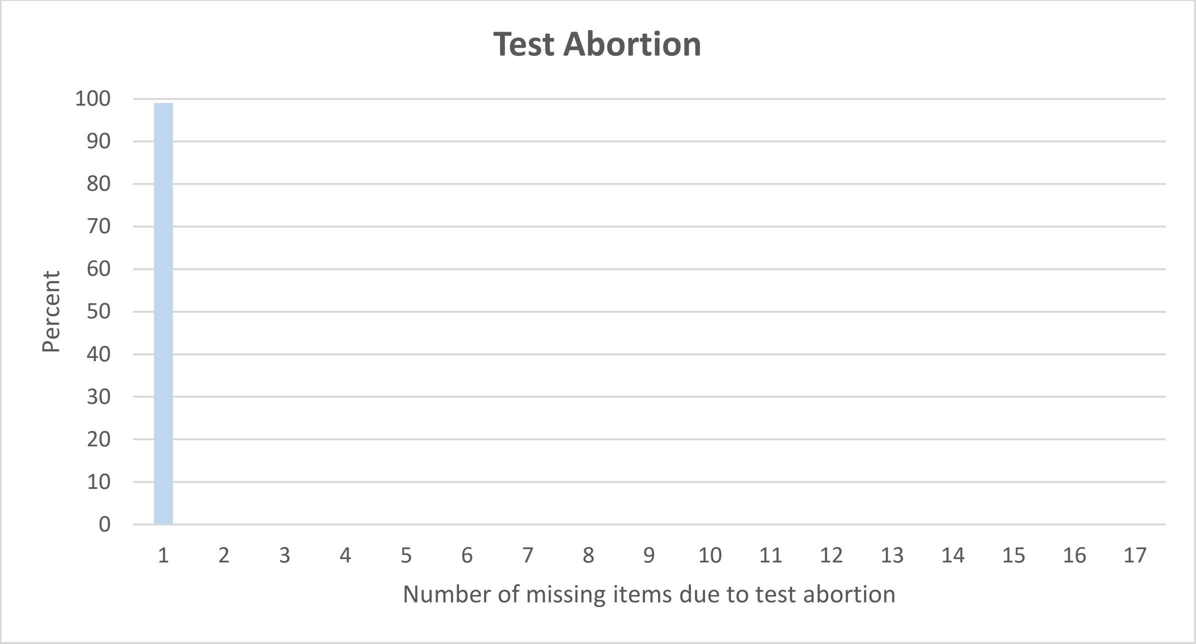

Figure 3. Number of test abortion missing items per person

Figure 4 shows the total number of missing responses per person, which is the sum of omitted items, and missing responses due to technical difficulties and test abortion. In total, 68.25 % of the students showed no missing response, whereas 1.39 % showed more than five missing responses.

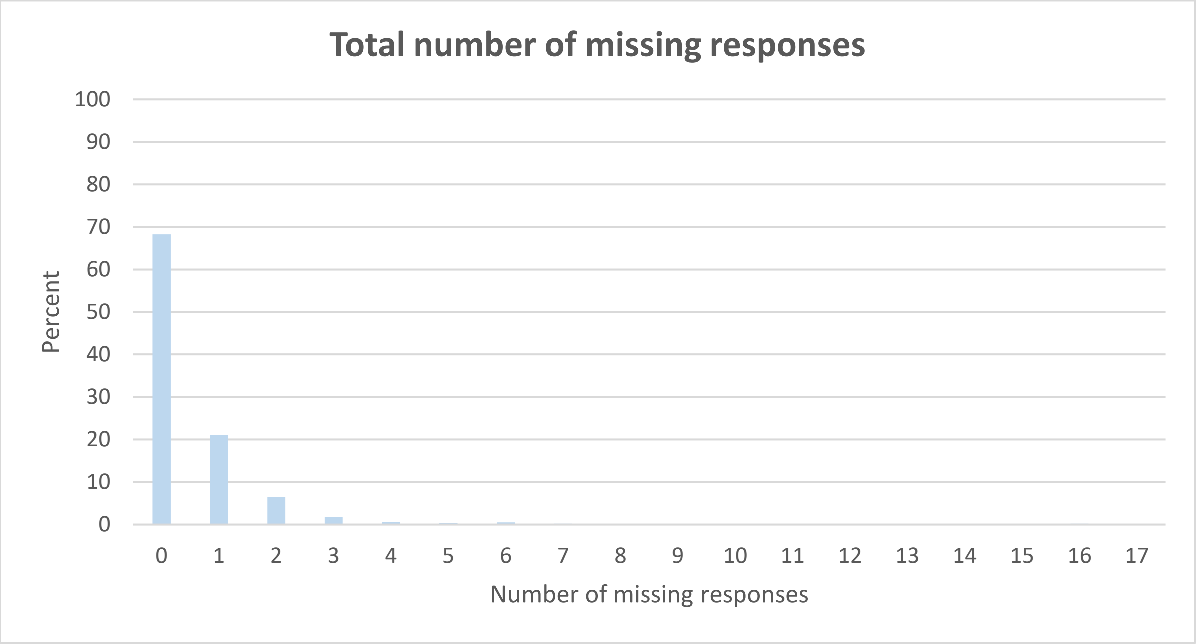

Figure 4. Total number of missing items per person

In sum, the number of omitted missing responses is rather small, even though they account for the greatest impact on the total number of missing responses. This indicates that for most children the test functioned as intended.

4.1.2 Missing responses per item

Table 3 shows the number of valid responses for each item, as well as the percentage of missing responses. Overall, the number of omitted responses per item was small, varying between 0.46 % (item man9d11s_c) and 13.10 % (item man9v011_c). The number of missing responses due to technical difficulties varied from 0.07 % (item mag2v071_sc1n9_c) to 0.86% (item mag2v041_sc1n9_c). The percentage of missing responses due to aborted test varied between 0.07 % (items man9z101_c, man9g071_c, and mag2v121_sc1n9_c) and 0.93 % (items man9z091_c, and mag2v041_sc1n9_c).

Table 3.

Percentage of Missing Values per Item

Pos.	Item	NV	OM	TD	TA
1	mag2v071_sc1n9_c	1,499	0.79	0.07	0.00
2	man9g041_c	1,493	1.12	0.13	0.00
3	mag2r031_sc1n9_c	1,482	1.72	0.26	0.00
4	man9d11s_c	1,500	0.46	0.33	0.00
5	man9z101_c	1,480	1.59	0.46	0.07
6	man9g071_c	1,480	1.85	0.20	0.07
7	mag2v121_sc1n9_c	1,474	2.12	0.33	0.07
8	mag2r111_sc1n9_c	1,471	2.25	0.20	0.26
9	man9z061_c	1,478	1.59	0.33	0.33
10	mag1d09s_sc1n9_c	1,490	0.93	0.20	0.33
11	man9z051_c	1,482	1.32	0.33	0.33
12	mag2g12s_sc1n9_c	1,460	2.38	0.60	0.46
13	man9d03s_c	1,466	2.38	0.20	0.46
14	man9v081_c	1,483	0.79	0.66	0.46
15	mag2r151_sc1n9_c	1,455	2.98	0.26	0.53
16	man9z021_c	1,478	1.46	0.20	0.60
17	mag1z071_sc1n9_c	1,482	0.93	0.40	0.66
18	man9v011_c	1,295	13.10	0.53	0.73
19	man9z091_c	1,460	1.98	0.53	0.93
20	mag2v041_sc1n9_c	1,471	0.93	0.86	0.93

Note. Pos. = Item position within the test. NV = Number of valid responses, OM = Percentage of respondents that omitted the item, TD = Percentage of respondents that had technical difficulties, TA = Percentage of respondents that had a missing due to test abortion.

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 14.80 % and 97.10 % across all items. On average, the rate of correct responses was 73.10 % (*SD* = 19.77 %).

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variables) for the data with aggregated subtasks of CMC items are depicted in Table 4a. The step parameters for polytomous variables 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 -2.99 (man9d03s_c) and 2.10 (man9z021_c) with a mean of -1.04 (*SD* = 1.32). Due to the large sample size, the standard errors of the estimated item difficulties (Table 4a, column 5 *SE*) were small (*SE*(ß) ≤ 0.09).

Table 4a.

Item Parameters

Pos.	Item	PC %	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	aQ_3
1	mag2v071_sc1n9_c	86.99	-2.26	0.08	0.95	-0.90	0.38	1.24	0.03
2	man9g041_c	71.40	-1.12	0.06	0.92	-2.72	0.51	1.50	0.04
3	mag2r031_sc1n9_c	84.95	-2.07	0.08	1.02	0.38	0.35	0.92	0.03
4	man9d11s_c	n.a.	-2.60	0.04	0.91	-2.09	0.43	0.99	0.05
5	man9z101_c	51.55	-0.07	0.06	0.92	-3.79	0.56	1.61	0.05
6	man9g071_c	33.65	0.83	0.06	0.97	-0.99	0.50	1.24	0.04
7	mag2v121_sc1n9_c	81.55	-1.79	0.07	0.99	-0.31	0.41	1.14	0.03
8	mag2r111_sc1n9_c	59.62	-0.47	0.06	1.11	4.66	0.37	0.59	0.04
9	man9z061_c	26.93	1.21	0.06	0.98	-0.60	0.47	1.10	0.03
10	mag1d09s_sc1n9_c	n.a.	-1.60	0.03	1.18	5.67	0.36	0.22	0.04
11	man9z051_c	56.55	-0.32	0.06	0.93	-3.15	0.55	1.54	0.06
12	mag2g12s_sc1n9_c	n.a.	-2.18	0.04	0.97	-0.75	0.37	0.68	0.04
13	man9d03s_c	n.a.	-2.99	0.05	0.97	-0.52	0.29	0.62	0.04
14	man9v081_c	62.91	-0.65	0.06	0.98	-0.86	0.48	1.12	0.04
15	mag2r151_sc1n9_c	66.46	-0.84	0.06	1.11	4.09	0.35	0.59	0.03
16	man9z021_c	14.82	2.10	0.08	1.01	0.23	0.39	0.96	0.03
17	mag1z071_sc1n9_c	80.97	-1.74	0.07	0.99	-0.35	0.40	1.14	0.04
18	man9v011_c	81.24	-1.75	0.08	1.10	2.17	0.29	0.58	0.02
19	man9z091_c	52.88	-0.13	0.06	0.98	-0.96	0.51	1.12	0.02
20	mag2v041_sc1n9_c	87.90	-2.35	0.09	0.97	-0.47	0.37	1.31	0.04

Note. Pos. = Item position in the test, PC % = Percentage correct answers, 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. 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 Polytomous Items

Item	step 1	step 2	step 3
man9d11s_c	-0.08 (0.06)		
mag1d09s_sc1n9_c	-1.69 (0.06)	0.16 (0.05)	0.81 (0.06)
mag2g12s_sc1n9_c	0.23 (0.07)		
man9d03s_c	0.20 (0.07)		

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 5, 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 mean of the ability distribution was constrained to be zero. The respective difficulties ranged from -2.99 (item man9d03s_c, item 13) to 2.10 (item man9z021_c, item 16). Therefore, a rather broad range was spanned. However, there was just one very difficult item and most of the items had a difficulty under zero (mean difficulty = -1.04). 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 1.10, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .76, WLE reliability = .73) was good.

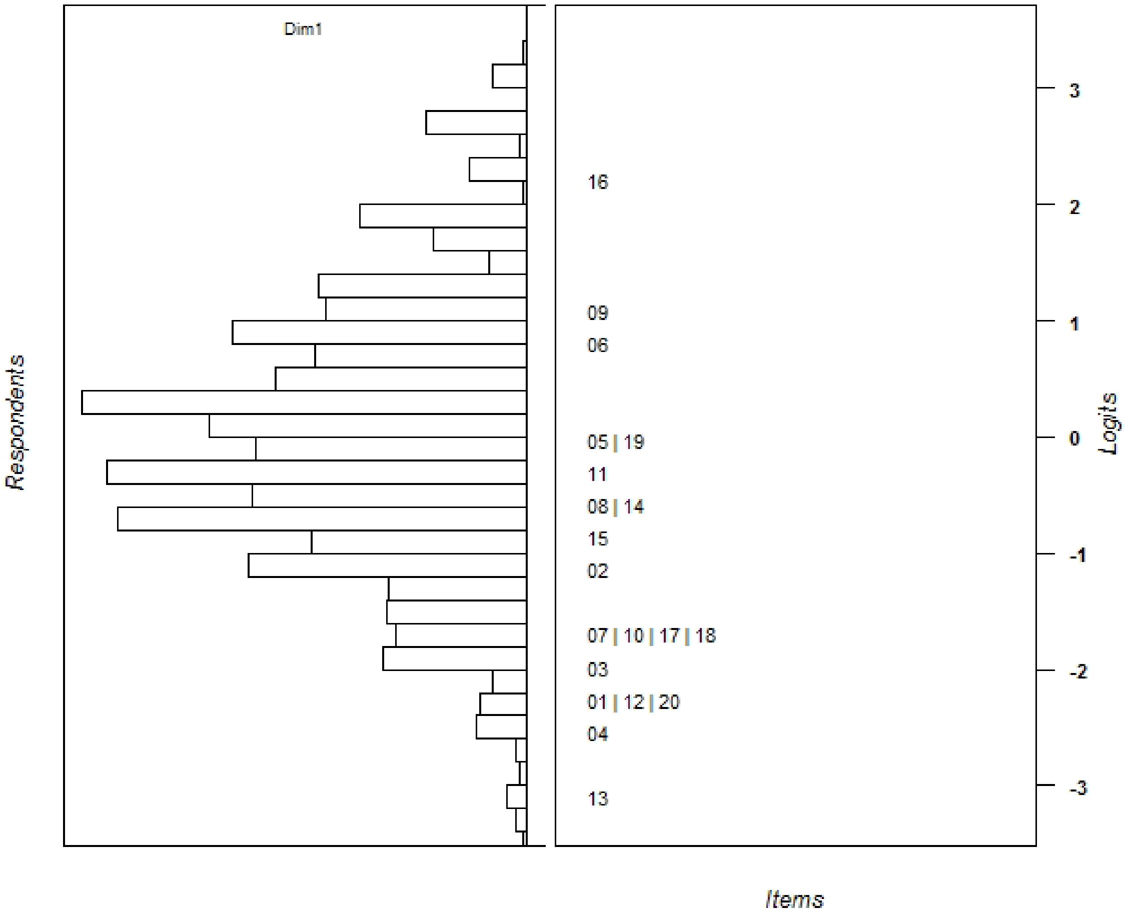

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 4a).

4.3 Quality of the test

4.3.1 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.

Table 5 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 -0.53 to -0.06 with a mean of -0.27. None of the 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 0.16 to 0.42 with a mean of 0.30 indicating that more proficient students were also more likely to identify the correct response option.

Table 5.

Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.30	-0.27
Minimum	0.16	-0.53
Maximum	0.42	-0.06

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 0.91 (man9d11s_c) and the highest being 1.18 (mag1d09s_sc1n9_c). However, the respective *t*-values were under 6 and, thus, 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 0.29 (man9d03s_c and man9v011_c) and 0.56 (man9z101_c). Overall, the items showed an average correlation of 0.42.

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 sex, migration background, highest parental international socioeconomic index (HISEI) (see Pohl & Carstensen, 2012, for a description of these variables), school month, and administration device (tablet versus laptop). 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).

Sex: Overall, 763 (50.46 %) of the students were female, 749 (49.54 %) were male, and no response for the sex variable was missing. On average, male students exhibited a higher mathematical competence than female students (main effect = -0.22 logits, Cohen’s *d* = -0.28). DIF exceeding 0.40 logits occurred for the items mag2r031_sc1n9_c, mag2r111_sc1n9_c, man9z051_c, and man9v081_c. No items exceeded 0.60 logits.

Table 6.

Differential Item Functioning

Pos.	Item	Sex	Migration	HISEI	School month	Device
		male vs. female	without vs. with	low vs. high	prior vs. after holidays	tablet vs. laptop
1	mag2v071_sc1n9_c	-0.22 (-.27)	0.25 (.32)	-0.07 (-.09)	0.13 (.16)	0.08 (.10)
2	man9g041_c	-0.35 (-.45)	-0.10 (-.12)	0.03 (.04)	-0.01 (-.02)	-0.06 (-.08)
3	mag2r031_sc1n9_c	0.52 (.66)	-0.45 (-.58)	0.29 (.37)	0.04 (.05)	-0.14 (-.18)
4	man9d11s_c	-0.07 (-.09)	-0.13 (-.16)	-0.05 (-.06)	0.02 (.03)	-0.07 (-.09)
5	man9z101_c	-0.30 (-.38)	0.03 (.04)	-0.12 (-.16)	0.17 (.21)	-0.19 (-.25)
6	man9g071_c	-0.36 (-.45)	-0.01 (-.01)	-0.08 (-.10)	-0.02 (-.03)	-0.12 (-.16)
7	mag2v121_sc1n9_c	-0.18 (-.23)	0.10 (.13)	0.21 (.27)	-0.10 (-.13)	0.05 (.06)
8	mag2r111_sc1n9_c	0.46 (.59)	0.06 (.07)	0.06 (.08)	0.09 (.12)	0.19 (.24)
9	man9z061_c	0.35 (.44)	-0.21 (-.27)	0.01 (.02)	-0.02 (-.02)	-0.05 (-.06)
10	mag1d09s_sc1n9_c	0.06 (.08)	0.19 (.25)	-0.16 (-.21)	0.11 (.14)	-0.04 (-.05)
11	man9z051_c	-0.45 (-.57)	-0.35 (-.45)	0.15 (.20)	-0.10 (-.13)	-0.09 (-.11)
12	mag2g12s_sc1n9_c	-0.10 (-.13)	0.25 (.32)	-0.20 (-.26)	0.03 (.04)	0.03 (.04)
13	man9d03s_c	0.10 (.12)	0.33 (.43)	-0.23 (-.29)	0.10 (.13)	0.23 (.30)
14	man9v081_c	0.49 (.63)	-0.28 (-.35)	0.18 (.24)	0.10 (.13)	0.04 (.05)
15	mag2r151_sc1n9_c	0.00 (-.01)	0.04 (.05)	-0.04 (-.05)	-0.05 (-.07)	0.02 (.03)
16	man9z021_c	0.05 (.07)	0.09 (.12)	-0.28 (-.36)	-0.13 (-.16)	0.03 (.03)
17	mag1z071_sc1n9_c	0.17 (.22)	0.00 (.00)	0.19 (.25)	-0.22 (-.28)	0.07 (.09)
18	man9v011_c	-0.16 (-.20)	-0.06 (-.07)	0.23 (.29)	-0.13 (-.17)	0.10 (.13)
19	man9z091_c	0.13 (.17)	-0.06 (-.08)	-0.15 (-.19)	0.01 (.01)	0.22 (.28)
20	mag2v041_sc1n9_c	-0.16 (-.21)	0.28 (.37)	0.01 (.01)	-0.01 (-.02)	-0.30 (-.38)
Main effect (DIF model)		**-0.22 (-.28)**	**-0.32 (-.42)**	**0.34 (.44)**	**-0.11 (-.14)**	**0.03 (.04)**
Main effect (Main effect model)		**-0.21 (-.26)**	**-0.25 (-.32)**	**0.25 (.33)**	**-0.06 (-.08)**	**0.03 (.03)**

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. HISEI = Highest international socio-economic index of parents.

Migration: There were 1,382 (91.40 %) participants without a migration background, and 130 (8.60 %) participants with a migration background. No response for the migration variable was missing. On average, participants without migration background performed better in the mathematics test than those with a migration background (main effect = -0.32 logits, Cohen's *d* = -0.42). Only item mag2r031_sc1n9_c exceeding 0.4 logits.

HISEI: The HISEI was calculated for the whole Starting Cohort 1 and divided in two categories (lower ≤ 60.70 and higher > 60.70 HISEI) using a median split. Overall, 555 (36.71 %) of the students had a lower HISEI whereas 957 (63.29 %) of the students had a higher HISEI. Students with a higher HISEI performed better than children with a lower HISEI (main effect = 0.34 logits, Cohen's *d* = 0.44). There was no item with DIF exceeding 0.4 logits.

School month: A total of 902 students (59.66 %) took the mathematics test before summer holidays or in the first two weeks of the holidays ("prior" in Table 6), while 610 (40.34 %) took it during the last four weeks of summer holidays or in a new school year ("after" in Table 6). Students who participated before the summer holidays showed a higher mathematics competence on average (main effect = -0.11 logits, Cohen's *d* = -0.14) than students who took the test during the summer holidays. There was no item with DIF exceeding 0.4 logits.

Device: A total of 1,260 students (83.33 %) used tablets (a device with touch functionality) to answer the items, while 252 students (16.67 %) used laptops or computers without touch functionality. The two groups showed, on average, negligible differences in mathematics competence (main effect = 0.03 logits, Cohen's *d* = 0.04). There was no item with DIF exceeding 0.4 logits.

Table 7.

Comparison of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Sex	Main effect	35,825.34	30	35,885.34	*36,044.98*
	DIF	35,715.10	49	*35,813.10*	36,073.84
Migration	Main effect	35,856.34	30	*35,916.34*	*36,075.98*
	DIF	35,831.42	49	35,929.42	36,190.16
HISEI	Main effect	35,799.62	30	*35,859.62*	*36,019.26*
	DIF	35,766.09	49	35,864.09	36,124.83
School month	Main effect	35,882.45	30	*35,942.45*	*36,102.09*
	DIF	35,869.25	49	35,967.25	36,227.99
Device	Main effect	35,887.03	30	*35,947.03*	*36,106.67*
	DIF	35,873.56	49	35,971.56	36,232.30

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

Overall, measurement invariance could be confirmed for all tested subgroups as the main effects and item DIFs were negligible. This corresponds to the model comparisons in Table 7. In Table 7, 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 main effect models for all variables except sex. 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 variables were preferred over the more complex DIF models.

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 generalized partial credit model (GPCM; Muraki, 1992) to the data. The estimated discrimination parameters are depicted in Table 4a ("Discr."). They varied between 0.22 (item mag1d09s_sc1n9_c) to 1.61 (item man9z101_c). The average discrimination parameter was 1.01. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 34,423.53, BIC = 34,668.31, number of parameters = 46) as compared to the PCM (AIC = 34,704.45, BIC = 34,848.12, 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 five-dimensional model based on the five different content areas. Each item was assigned to one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Quasi-Monte Carlo 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 10,000 nodes. The variances, correlations, and EAP reliability of the five dimensions are shown in Table 8. All five dimensions exhibited a substantial variance. The correlations among the five dimensions were rather high and varied between 0.785 and 0.967. However, the AIC and BIC favored the five-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 = .04, SD = .01) — the largest individual residual correlation was .06 — and, thus, indicated an essentially unidimensional test. Because the mathematics test was constructed to measure a single dimension and the correlation between the dimensions were rather high, a unidimensional mathematics competence score was estimated.

Table 8.

Results of Five-Dimensional Scaling

	Units and measuring	**Sets, numbers and operations**	**Change and relation-ships**	**Data and chance**	**Space and shape**
Units and measuring (3 items)	(1.412)				
Sets, numbers and operations (6 items)	0.967	(1.278)			
Change and relationships (8 items)	0.911	0.922	(1.092)		
Data and chance (6 items)	0.785	0.817	0.875	(1.003)	
Space and shape (6 items)	0.797	0.785	0.867	0.812	(0.813)
EAP reliability	0.733	0.749	0.723	0.617	0.596

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

Table 9

Comparison of the Unidimensional and the Five-Dimensional Model

Model	**Deviance**	**Number of parameters**	**AIC**	**BIC**
Unidimensional	34716,58	27	34,770.58	34,914.25
Five-dimensional	34514,61	41	*34,596.61*	*34,814.78*

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 1 of the NEPS (wave 9), on average 8-year old students in second grade. 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 amount of different kinds of missing responses was 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 item distribution along the ability scale was acceptable. The range of the item difficulties were good, although most of the items showed difficulties under zero and were rather low. The test was designed to match the competencies students should have achieved in the first semester of second grade in elementary school. Due to the COVID-19 pandemic, data collection had to be rescheduled from spring to summer and fall. Thus, by changing the time of data collection, the students should have achieved further competencies in mathematics. The students gain about half a standard deviation in mathematics competencies within a school year in elementary school and lower secondary education (Lehmann, 2008; Vom Hofe et al., 2002). Therefore, most of the items were easy to solve for the students. Nevertheless, the test had a good reliability and distinguished acceptable between the students as indicated by the test's variance. Different variables were used for testing measurement invariance (DIF). 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 five dimensions indicate that the five content areas measure a common construct, although the model indices AIC and BIC favored the five-dimensional model.

In sum, the mathematic test had acceptable to 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 20 items in the data set that are either scored as dichotomous variables (MC and Sorting items) with 0 indicating an incorrect response and 1 indicating a correct response or scored as a polytomous variable (corresponding to the CMC items, considering the merging of categories, see Chapter 3.4). The dichotomous variables are marked with '_c' at the end of the variable name; the polytomous variables are marked with 's_c' or 's_sc1n9_c' behind their variable names. Items that were already administered in other waves in the NEPS kept their original names ('mag2v071…', 'mag2r031…', 'mag2v121…', 'mag2r111…', 'mag1d09…', 'mag2g12…', 'mag2r151…', 'mag1z071…', 'mag2v041…'). However, for reasons of identification a suffix was added to specify the current test administration ('sc1n9' referring to Starting Cohort 1, wave 9).

6.2 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided as WLE ("man9_sc1") including the corresponding standard errors ("man9_sc2"). The WLE scores in "man9_sc1" are not linked to the underlying reference scales of kindergarten (first and second data collection). As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. Currently no linked item difficulty parameters for longitudinal comparisons between kindergarten 4-year old, kindergarten 6-year old, and grade 2 (8 -year old) are available. Due to the COVID-19 pandemic the data collection of the linking study had to be rescheduled. Please consider the SUF updates on the NEPS homepage to be informed about additions of linked WLEs to the SUF of wave 9 in Starting Cohort 1.

The R Syntax for estimating the WLE scores from the items is provided in Appendix B. Students that did not take part in the test or those that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE scores for mathematical competence.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. Plausible values for competence tests administered in the NEPS can be estimated using the R package *NEPSscaling*[1] (Scharl, Carstensen, & Gnambs, 2020).

[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,* 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). Waxmann.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2021). *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

Gnambs, T. (2022). *NEPS Technical Report for Early Reading Competence: Scaling Results of Starting Cohort 1 (Wave 9)* (NEPS Survey Paper No. 96). Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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.

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). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP83:1.0

Lehmann, R. (2008). *Erhebung zum Lese- und Mathematikverständnis: Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin* (ELEMENT) (Version 1). IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_ELEMENT_v

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

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://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://doi.org/10.25656/01:8430

R Core Team (2022). R: A language and environment for statistical computing (Version 4.1.3) [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. (2021). *TAM: Test Analysis Modules*. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 3.7-16).

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

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:luen4-opus-144121

Vom Hofe, R., Pekrun, R., Kleine, M., & Götz, T. (2002). Projekt zur Analyse der Leistungsentwicklung in Mathematik (PALMA): Konstruktion des Regensburger Mathematikleistungstests für 5. bis 10. Klassen. In *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (pp. 83-100). Beltz. (Zeitschrift für Pädagogik, Beiheft; 45) https://doi.org/10.25656/01:3940

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).* VS Verlag für Sozialwissenschaften.

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

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

Appendix A. Content Areas of Items in the Mathematics Test for 8-year old students (Grade 2)

Position	Item	Content area
1	mag2v071_sc1n9_c	Change and relationships
2	man9g041_c	Units and measuring
3	mag2r031_sc1n9_c	Space and shape
4	man9d11s_c	Data and chance
5	man9z101_c	Sets, numbers, and operations
6	man9g071_c	Units and measuring
7	mag2v121_sc1n9_c	Change and relationships
8	mag2r111_sc1n9_c	Space and shape
9	man9z061_c	Sets, numbers, and operations
10	mag1d09s_sc1n9_c	Data and chance
11	man9z051_c	Sets, numbers, and operations
12	mag2g12s_sc1n9_c	Units and measuring
13	man9d03s_c	Data and chance
14	man9v081_c	Change and relationships
15	mag2r151_sc1n9_c	Space and shape
16	man9z021_c	Sets, numbers, and operations
17	mag1z071_sc1n9_c	Sets, numbers, and operations
18	man9v011_c	Change and relationships
19	man9z091_c	Sets, numbers, and operations
20	mag2v041_sc1n9_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 fitting the partial credit model in Starting Cohort 1 wave 9

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

### load data
dat <- read_sav(file = "SC1_xTargetCompetencies_D_9-0-0.sav")
items <- c( [add the names of the items provided in Appendix A] )

### Fit the model
model <- tam.mml(resp = dat[, items], pid = dat$ID_t)
summary(model)

### Estimate WLEs
wle <- tam.wle(model, Msteps = 1000)
```