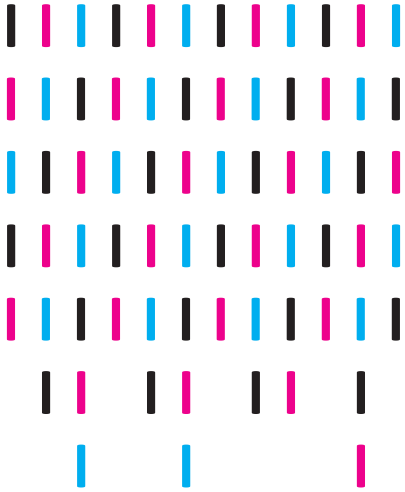

NEPS *SURVEY PAPERS*

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

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 1 FOR 6-YEAR-OLD CHILDREN

NEPS *Survey Paper* No. 74
Bamberg, September 2020; Updated: March 2022

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 1 for Six-Year-Old Children

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. (2020): *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 1 for Six-Year-Old Children* (NEPS Survey Paper No. 74). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP74:2.0

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., Schnittjer & Gerken, 2017).

Please note: This NEPS Survey Paper has been modified in March 2022. You will find the documentation of the modifications on the last page, following the appendix. The original paper can still be found using https://doi.org/10.5157/NEPS:SP74:1.0

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 1 for Six-Year-Old Children

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and designs tests for assessing these different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedure for the mathematical competence test of six-year-old children in wave 7 of starting cohort 1 (newborns). The mathematics test contained 25 items with different response formats representing different content areas and cognitive components. The test was administered to 1,989 children. Their responses were scaled using the Rasch model. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test´s dimensionality were evaluated to ensure the quality of the test. These analyses showed that the test exhibited a good reliability (EAP/PV reliability = 0.810) and a good rasch model fit. Furthermore, test fairness could be confirmed for different subgroups. Overall, the mathematics test had acceptable psychometric properties that allowed for an estimation of reliable mathematics competence scores. Besides the scaling results, this paper also describes the data available in the Scientific Use File and provides the ConQuest syntax for scaling the data as well as the 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 across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies (ICT) literacy, metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competence domains measured in the NEPS is given by Weinert et al. (2011) as well as Fuß, Gnambs, Lockl, and Attig (2019).

Most of the competence data are scaled using models that are based on 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 of six-year-old children in wave 7 of starting cohort 1 (newborns). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of starting cohort 1 and the analyses performed 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.

Please note that the analyses of this report are based on the data available some time before data release. Due to data protection and data cleaning issues, the data in the Scientific Use File (SUF) may differ slightly from the data set used for analyses in this paper. However, fundamentally different results are not expected.

2 Testing Mathematical Competence

The framework and test development for the test of mathematical competence are described in Weinert et al. (2011), Neumann et al. (2013), and Ehmke et al. (2009). In the following, we briefly describe specific aspects of the mathematics test that are necessary for understanding the scaling results presented in this paper.

The items are not arranged in units. Thus, in the test, students usually face a certain situation followed by only one task related to it; sometimes there are two tasks. Each of the items belongs to one of the following five 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. The framework also describes as a second and independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

The mathematics test includes four types of response formats. These are simple multiple-choice (MC), short constructed response (SCR), matching (M), and sorting (S). The most common response format for this age group is the short constructed response (SCR). SCR items require the test-taker to give mostly one-word answers, such as a number. All SCR items were scored dichotomously. Simple multiple-choice items (MC) are items where the children have to find the correct answer from several, usually three or four, response options presented as pictures. Another response format that was given was to sort objects into their correct order (S). Items with this response format were scored dichotomously as well as there is only one correct order in each item. In matching items (M) the children were asked to match some picture cards to given response options. These tasks were constructed in such a way to enable clear dichotomous scoring.

3 Data

3.1 The Design of the Study

The study assessed different competence domains including, among others, mathematical competence and other (non-verbal) cognitive basic skills. All participants received the same mathematics items in the same order. The test was conducted as an individual tablet-based test and was administered at the child's home. The children's answers were recorded by an interviewer.

The mathematics test included 25 items which represented different content-related and process-related components and used different response formats. The characteristics of the 25 items are depicted in the following tables. Table 1 shows the distribution of the five content areas (see Appendix C for the assignment of the items to the content areas), whereas Table 2 shows the distribution of the response formats.

Table 1: Number of Items by Content Areas.

Content area	Frequency
Sets, numbers, and operations	10
Units and measuring	5
Space and shape	4
Change and relationships	4
Data and chance	2
Total number of items	25

Table 2: Number of Items by Response Formats.

Response format	Frequency
Short Constructed Response	15
Simple Multiple-Choice	7
Sorting	2
Matching	1
Total number of items	25

One simple multiple-choice item (man7d021_c) was excluded from the analyses due to severe misfit, resulting in a test of 24 items.

3.2 Sample

Overall, the test was administered to 1,989 children. For 21 respondents less than three valid item responses were available. 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 are based on a sample of 1,968 test takers. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) items that test takers did not reach, and c) missings that are produced when the test administrator aborted the testing.

In this study, all children received the same set of items. As a consequence, there were no items that were not administered to a person. There were also no invalid answers, as the interviewer recorder the answers. Omitted items occurred if the child did not respond to an item. After three consecutively omitted items, the test was not continued. All subsequent item were coded as “not reached”. Due to reasons like exhaustion or sudden and consistent refusal to participate, it may have occurred that some children did not finish the test and the test had to be aborted without three consecutively omitted items. All responses after the test abortion are rated as “test aborted”. There was no time limit for the test.

Missing responses provide information on how well the test worked (e.g., time limits, exhaustion, understanding of instructions). Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the children were coping with the test. Missing responses per item were examined in order to evaluate how well the items functioned.

3.4 Scaling Model

Item and person parameters were estimated using a Rasch model (Rasch, 1960). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

All items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Mathematical competencies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF are described in section 6.

3.5 Checking the Quality of the Scale

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

The fit of the dichotomous variables to the Rasch model (Rasch, 1960) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (t-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.2 (t-value > |8|) were judged as a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total correct score (equal to the discrimination value as computed in ConQuest) greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall, judgment of the fit of an item was based on all fit indicators.

The mathematical competence test should measure the same construct for all participants. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) was examined using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 as small and not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The competence data in the NEPS are usually scaled assuming Rasch-homogeneity. The Rasch (1960) model was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a two-parametric logistic model (2PL; Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the mathematics test was evaluated by specifying a five-dimensional model based on the five content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, TAM in R was used (Kiefer, Robitzsch, & Wu, 2017). The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (9,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The IRT models were estimated in ConQuest version 4.2.5 (Wu, Adams, & Wilson, 2015). The 2PL model was estimated in mdltm (Matthias von Davier, 2005). The multi-dimensional-model was estimated in TAM version 2.8-21 (Kiefer, Robitzsch, & Wu, 2017) in R version 3.4.2 (R Core Team, 2017).

4 Results

4.1 Missing Responses

4.1.1 Missing responses per person

Missing responses may occur when a child does not respond to an item (omit). The number of omitted responses per person is depicted in Figure 1. It shows that 66.2 % of the children omitted no item and only 2.0 % of the children omitted five or more items.

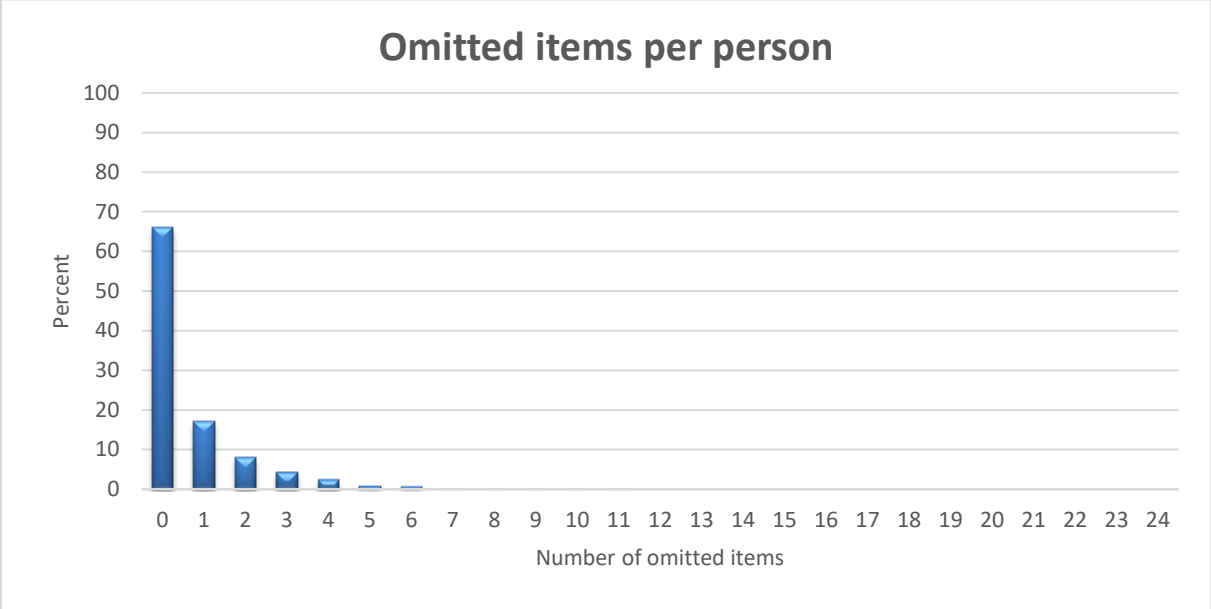

Figure 1: Number of omitted items.

All missing responses after the last valid response are defined as not reached. Figure 2 shows the number of items that were not reached by a person. Nearly all children reached the end of the test. Only 0.6 % didn't reach the end of the test.

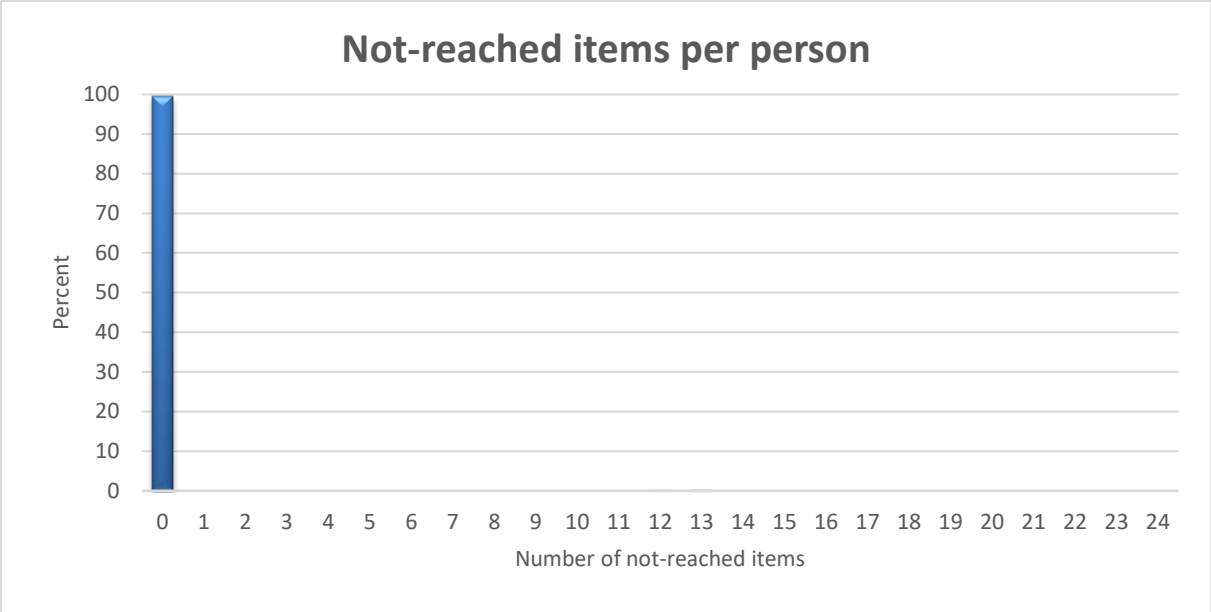

Figure 2: Number of not-reached items.

Figure 3 shows the number of test-aborted items which were defined in case the test administrator had to abort the test. In 99.9 % of all cases, no interruption was necessary. In only 0.1 % of the cases, the test had to be aborted.

Figure 3. Number of test-aborted items.

Figure 4 shows the total number of missing responses per person, which is the sum of omitted, not-reached, and test-aborted missing responses. In total, 66.1 % of the test takers showed no missing response, whereas only 1.7 % showed more than five missing responses. Overall, there was a negligible amount of missings.

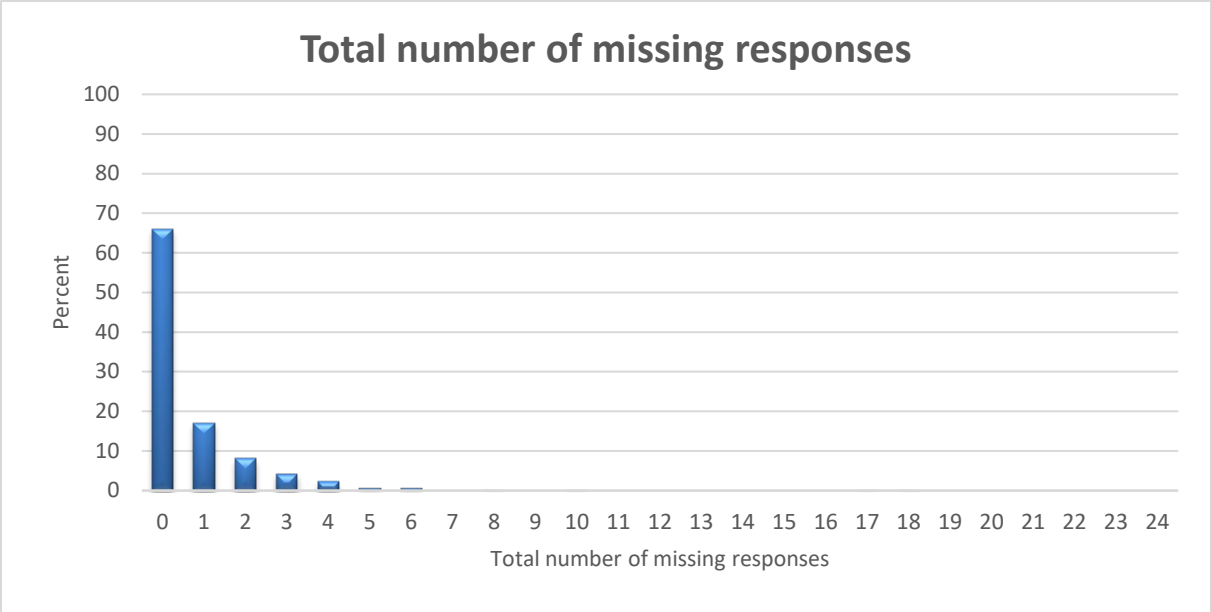

Figure 4. Total number of missing responses.

4.1.2 Missing responses per item

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

The omission rates were good, except for one noticeable item with an omission rate higher than 10 %. An omission rate of 11.33 % occurred for item man7z101_c. The number of persons that did not reach an item increased with the position of the item in the test up to 0.61 %. The percentage of test-aborted items also increased with the position of the item in the test up to 0.15 %. The total number of missing responses per item varied between 0.20 % (man7g061_c) and 11.33 % (man7z101_c).

Table 3: Percentage of Missing values.

Pos	Item	Number of valid responses	Percentage of omitted responses	Percentage of not-reached items	Percentage of test-aborted items	Total number of missing responses
1	**man7z211_c**	1,946	1.12	0.00	0.00	1.12
2	**man7z201_c**	1,961	0.36	0.00	0.00	0.36
3	**man5v181_sc1n7_c**	1,961	0.36	0.00	0.00	0.36
4	**man7z101_c**	1,745	11.33	0.00	0.00	11.33
5	**man7r111_c**	1,943	1.27	0.00	0.00	1.27
6	**man7g051_c**	1,953	0.76	0.00	0.00	0.76
7	**man7g061_c**	1,964	0.20	0.00	0.00	0.20
8	**man7v011_c**	1,920	2.44	0.00	0.00	2.44

9	**man7r151_c**	1,787	9.20	0.00	0.00	9.20
10	**man7g131_c**	1,922	2.34	0.00	0.00	2.34
11	**man7z041_c**	1,883	4.27	0.05	0.00	4.32
12	**man7d071_c**	1,858	5.34	0.25	0.00	5.59
13	**man7g191_c**	1,930	1.52	0.41	0.00	1.93
14	**man7r121_c**	1,954	0.20	0.51	0.00	0.71
15	**man7z081_c**	1,916	2.13	0.51	0.00	2.64
16	**man7v091_c**	1,890	3.40	0.51	0.05	3.96
17	**man7z171_c**	1,903	2.64	0.51	0.15	3.30
18[e]	**man7d021_c**	-	-	-	-	-
19	**man7z221_c**	1,784	8.59	0.61	0.15	9.35
20	**man5z081_sc1n7_c**	1,855	4.98	0.61	0.15	5.74
21	**man7g031_c**	1,953	0.00	0.61	0.15	0.76
22	**man7z231_c**	1,912	2.08	0.61	0.15	2.85
23	**man7r181_c**	1,953	0.00	0.61	0.15	0.76
24	**man7v161_c**	1,921	1.63	0.61	0.15	2.39
25	**man7z141_c**	1,914	1.98	0.61	0.15	2.74

Note. [e]Excluded from the analyses due to unsatisfactory item fit.

4.2 Parameter Estimates

4.2.1 Item parameters

In order to get a first descriptive measure of the item difficulties and check for possible estimation problems, the relative frequency of the responses was evaluated before performing any IRT analyses. The percentage of persons correctly responding to an item (relative to all valid responses) varied between 20.48 % and 89.80 % across all items. On average, the rate of correct responses was 58.18 % (*SD* = 20.00 %).

From a descriptive point of view, the items covered a wide range of difficulties. The estimated item difficulties are depicted in Table 4. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.58 (man7z201_c) and 1.64 (man7g051_c) with a mean of -0.45. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were very small (*SE*(ß) ≤ 0.08).

Table 4: Item Parameters.

	Item	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.
1	**man7z211_c**	69.58	-1.01	0.059	0.95	-2.1	0.52	1.10
2	**man7z201_c**	89.80	-2.58	0.083	0.99	-0.2	0.36	0.87
3	**man5v181_sc1n7_c**	83.07	-1.91	0.070	1.02	0.4	0.39	0.74
4	**man7z101_c**	71.63	-1.03	0.063	0.96	-1.4	0.49	1.01
5	**man7r111_c**	66.39	-0.83	0.058	1.14	5.5	0.34	0.45
6	**man7g051_c**	20.48	1.64	0.065	0.95	-1.5	0.43	1.28
7	**man7g061_c**	26.99	1.21	0.060	0.89	-4.2	0.52	1.51
8	**man7v011_c**	79.79	-1.65	0.067	0.99	-0.3	0.43	0.81
9	**man7r151_c**	50.20	-0.01	0.058	1.15	7.4	0.32	0.40
10	**man7g131_c**	36.89	0.67	0.057	1.07	3.5	0.38	0.57
11	**man7z041_c**	62.72	-0.60	0.058	1.01	0.4	0.46	0.83
12	**man7d071_c**	54.63	-0.19	0.057	1.05	2.6	0.42	0.61
13	**man7g191_c**	55.49	-0.25	0.056	1.02	1.1	0.45	0.70
14	**man7r121_c**	29.22	1.08	0.059	1.15	6.0	0.25	0.30
15	**man7z081_c**	59.92	-0.47	0.057	0.92	-3.9	0.55	1.17
16	**man7v091_c**	62.43	-0.59	0.058	0.90	-4.9	0.57	1.30
17	**man7z171_c**	64.37	-0.70	0.058	0.94	-2.9	0.53	1.05
18[e]	**man7d021_c**	-	-	-	-	-	-	-
19	**man7z221_c**	32.62	0.92	0.061	0.93	-3.1	0.52	1.30
20	**man5z081_sc1n7_c**	82.59	-1.83	0.071	0.93	-1.7	0.48	1.10
21	**man7g031_c**	77.01	-1.46	0.064	0.94	-1.9	0.50	1.03
22	**man7z231_c**	38.60	0.59	0.057	0.97	-1.6	0.48	0.99
23	**man7r181_c**	58.47	-0.41	0.056	1.16	7.7	0.30	0.36
24	**man7v161_c**	40.76	0.48	0.057	0.92	-4.5	0.54	1.25
25	**man7z141_c**	82.65	-1.86	0.070	1.01	0.3	0.40	0.78

Note. Difficulty = Item difficulty, SE = Standard error of item difficulty, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a two-parametric logistic model (2PL).
The item-total correlation corresponds to the point-biserial correlation between the correct response and the total score (discrimination value as computed in ConQuest).
[e]Excluded from the analyses due to unsatisfactory item fit.

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person's abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 5, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas 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 -2.584 (item man7z201_c) to 1.639 (item man7g051_c). Therefore, a rather broad range was spanned. The variance was estimated to be 1.108, which implies good differentiation between the test takers. The reliability of the test (EAP/PV reliability = 0.810, WLE reliability = 0.790) was good. In addition to the wide range of the ability distribution, there was also a somewhat equal distribution of easy and difficult items. Therefore, the measurement of mathematical competence of persons with high and low abilities should be relatively precise.

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

Figure 5: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 11.3 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4).

4.3 Quality of the test

4.3.1 Distractor analyses

To investigate how well the distractors of the MC items performed in the test, the point-biserial correlations between selecting each incorrect response (distractor) and the child's total correct scores was evaluated. The point-biserial correlations for the distractors ranged from -0.39 to 0.00 with a mean of -0.13. 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 0.23 to 0.43 with a mean of 0.33 indicating that high-performing children 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	0.33	-0.13
Minimum	0.23	-0.39
Maximum	0.43	0.00

4.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the Rasch model, as all items were scored dichotomously. Altogether, item fit can be considered to be good (see Table 4). Values of the WMNSQ were close to 1 with the lowest value being 0.89 (man7g061_c) and the highest being 1.16 (man7r181_c). This was the only item with a noticeable WMNSQ above |1.15| and a noticeable *t*-value of 7.7.

All ICC showed a good or very good fit of the items. Overall, there was no indication of severe item over- or underfit. The correlations of the item scores with the total scores varied between 0.25 (man7r121_c) and 0.57 (man7v091_c) with an average correlation of 0.44.

4.3.3 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, migration background, and the number of books at home (see Pohl & Carstensen, 2012, for a description of these variables). Table 6 shows the differences between the estimated difficulties of the items in these different subgroups. Female versus male, for example, indicates the difference in difficulty between boys and girls, β(male) – β(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females.

Overall, 975 (49.5 %) of the test takers were female and 993 (50.5 %) were male. On average, male children exhibited a higher mathematical competence than female children (main effect = -0.16 logits, Cohen's *d* = 0.15). There was no item with a considerable gender DIF above 0.6 logits. The only items for which the difference in item difficulties between the two

groups exceeded 0.4 logits were items mag5v181_sc1n7_c (0.42), man7g061_c (-0.48), man7g031_c (0.42), and man7z231_c (-0.58).

There were 1,440 (73.2 %) participants without migration background, 480 (24.4 %) participants with migration background, and 48 (2.4 %) children with unknown migration status. Group differences and DIF were investigated by using the first two groups. On average, participants with migration background performed considerably worse in the mathematics test than those without migration background (main effect = -0.46 logits, Cohen's *d* = 0.45). DIF exceeding 0.4 logits occurred only for item man7z201_c with 0.41 logits.

The number of books at home was used as a proxy for socioeconomic status. There were 549 (27.9 %) test takers with 0 to 100 books at home, 1,308 (66.5 %) test takers with more than 100 books at home, and 111 (5.6 %) test takers without a valid response. Group differences and DIF were investigated by using the first two groups. On average, participants with 100 or fewer books at home exhibited a considerably lower mathematical competence than participants with more than 100 books (main effect = 0.66, Cohen's *d* = 0.66). The only item for which the difference in item difficulties between the two groups exceeded 0.4 logits was item man7z141_c (0.41 logits).

Table 6: Differential Item Functioning.

	Item	Gender	Migration status	Number of books
		male vs. female	without vs. with	<=100 vs. >100
1	**man7z211_c**	0.17	0.12	-0.02
2	**man7z201_c**	0.09	0.41	-0.16
3	**man5v181_sc1n7_c**	0.42	0.08	-0.08
4	**man7z101_c**	-0.32	0.13	0.16
5	**man7r111_c**	0.07	0.11	-0.40
6	**man7g051_c**	-0.36	-0.01	0.31
7	**man7g061_c**	-0.48	-0.01	0.14
8	**man7v011_c**	0.03	-0.25	0.22
9	**man7r151_c**	0.24	-0.10	-0.08
10	**man7g131_c**	0.36	0.05	-0.03
11	**man7z041_c**	0.20	-0.05	-0.08
12	**man7d071_c**	-0.06	-0.08	0.03
13	**man7g191_c**	-0.19	-0.08	-0.14
14	**man7r121_c**	-0.02	-0.09	-0.21
15	**man7z081_c**	0.21	-0.02	0.02
16	**man7v091_c**	0.34	-0.09	0.12

17	**man7z171_c**	0.11	0.10	0.07
18[e]	**man7d021_c**	-	-	-
19	**man7z221_c**	-0.30	-0.09	0.15
20	**man5z081_sc1n7_c**	0.09	0.19	0.01
21	**man7g031_c**	0.42	-0.07	0.31
22	**man7z231_c**	-0.58	-0.05	-0.10
23	**man7r181_c**	-0.09	0.14	-0.25
24	**man7v161_c**	-0.12	-0.19	0.40
25	**man7z141_c**	-0.13	0.21	-0.41
Main effects:				
DIF model		-0.16	-0.46	0.66
Main effect model		-0.16	-0.46	0.66

Note. [e]Excluded from the analyses due to unsatisfactory item fit.

Overall, testfairness could be confirmed for all analysed subgroups. In Table 7, we compared the models that only included main effects to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for the variables gender and books. Merely for the migration variable the model estimating only the main effect was favored. 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 economical models including only the main effects were preferred over the more complex DIF models for all variables.

Table 7: Comparison of Models with and without DIF.

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	main effect	49,944.356	26	49,996.36	50,141.56
	DIF	49,804.244	50	49,904.24	50,183.48
Migration status	main effect	48,679.684	26	48,731.68	48,876.25
	DIF	48,656.640	50	48,756.64	49,034.64
Books	main effect	47,087.273	26	47,139.27	47,282.97
	DIF	47,027.541	50	47,127.54	47,403.88

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. In order to test this assumption, a two-parametric logistic model (2PL) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4), ranging from 0.30 (item man7r121_c) to 1.51 (item man7g061_c). The average discrimination was 0.90. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 49,383.47,

BIC = 49,729.73, number of parameters = 62) as compared to the 1PL model (AIC = 50,000.57, BIC = 50,212.80, number of parameters = 38). Despite the empirical preference for the 2PL model, the 1PL model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the 1PL model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

Note that these calculations could not be made by conquest 4.2.5 so that we had to use MDLTM (see 3.6, Davier, 2005). As a consequence, the results for AIC and BIC using the 1PL model might differ slightly from the later results (see 4.3.5) comparing multi-dimensionality to unidimensionality of the test, estimated in R (see 3.6).

4.3.5 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 15,000 nodes.

The variances and correlations of the five dimensions are shown in Table 8. Four of the five dimensions exhibit a substantial variance. In dimension three (space and shape), three of the four items showed difficulties ranging from -0.830 to -0.09, so the difficulties were relativley homogenous in this dimension. Additionaly, point-biserial correlations of all four items were rather low (see Table 4). This might explain the rather small variance of 0.611 in dimension three.

For four of the five dimensions, the correlations between the dimensions were rather high, and varied between 0.697 and 0.941. For dimension five (data and chance), the correlations were slightly smaller, varying from 0.593 to 0.795. Due to the fact that dimension five only includes one item, these values seem plausible. Initially, this dimension included two items, but item man7d021_c had been excluded from the analyses due to unsatisfactory item fit.

According to model fit indices, the five-dimensional model fitted the data slightly better (AIC = 49,477.26, BIC = 49,695.07, number of parameters = 39) than the unidimensional model (AIC = 50,003.23, BIC = 50,142.85, number of parameters = 25).

These results indicate that the five content areas measure a common construct, although it is not completely unidimensional. Model fit between the unidimensional and the five-dimensional model is compared in Table 9. Because the mathematics test was constructed to measure a single dimension, a unidimensional mathematics competence score was estimated.

Table 8: Results of Five-Dimensional Scaling.

	Sets, numbers and operations	Units and measurement	Space and shape	Change and Relationship	Data and chance
Sets, numbers and operations (10 items)	1.367				
Units and measurement (5 items)	0.871	1.228			
Space and shape (4 items**)**	0.697	0.869	0.611		
Change and relationships (4 items)	0.896	0.941	0.807	1.406	
Data and chance (1 items)	0.593	0.705	0.795	0.603	1.433

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

Table 9: Comparison of the Unidimensional and the Five-Dimensional Model.

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	49,953.2	25	50,003.23	50,142.85
Five-dimensional	49,399.3	39	49,477.26	49,695.07

***Note*.** Contrary to the calculations for the 1PL and 2PL models, results in this table were achieved by using TAM in R (see 3.6).

5 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test for six-year-old children in starting cohort 1 and at describing how the mathematics competence score had been estimated.

The amount of different kinds of missing responses was evaluated and the number of most kinds of missing responses was rather low. Furthermore, item as well as test quality were examined. As indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, ICC – the items exhibited a good item fit. Also, discrimination values of the items (either estimated in a 2PL model or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. No considerable DIF was shown for any of these variables, indicating that the test was fair for the examined subgroups. The test had a good reliability and distinguished well between test takers, as indicated by the test's variance. The item distribution along the ability scale was good.

Fitting a five-dimensional Rasch model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model than the unidimensional model. Nevertheless, high correlations between the four dimensions indicate that the unidimensional model described the data well. However, predominantly high correlations between the five dimensions indicated that the unidimensional model described the data reasonably well.

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

6 Data in the Scientific Use File

6.1 Naming conventions

The data in the Scientific Use File contains 25 items, that were all scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. Items that were already administered in other waves kept their original names (‘man5v181…’, ‘man5z081…’). For reasons of identification, a suffix was added in front of the ‘…_c’ to specify the current test administration (‘sc1n7’ referring to Starting Cohort 1, newborns, wave 7).

6.2 Linking of competence scores

In starting cohort 1, the mathematics competence tests administered in wave 5 (four-year-old children) and in wave 7 (six-year-old children) include primarly different items that were constructed in such a way that allows an accurate measurement of mathematical competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competencies as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competencies across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). The process of linking combines adjacent measurement points on the same scale. As such, the first wave of each competence scale within a cohort is used as a reference scale that all subsequent measurement waves will refer to.

For the domain of mathematical competence, linking typically is achieved using overlapping items between tests (also known as common items). In this case, we are following an anchor-group design, because of the large competence growth from wave 5 (four-year-old children) to wave 7 (six-year-old children). All items from the mathematics competence tests of wave 5 and wave 7 were administered to an independent link sample, including five- and six-year-old children, that were not part of starting cohort 1, within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 1 across the two waves. An empirical study that evaluated different link methods with regard to the appropriateness of linking NEPS data (Fischer et al., 2016) showed that the method of mean/mean linking (see Kolen & Brennan, 2004) is appropriate for the NEPS tests. For more information on the selection of link samples and the method for linking the tests of mathematical competence see Fischer et al. (2016).

6.2.1 Samples

In starting cohort 1, a subsample of 1,696 children (50.2 % female) participated at both measurement occasions in wave 5 (four-year old children) and wave 7 (six-year old children). Consequently, these respondants were used to link the two tests across both waves (see Fischer et al., 2016). Due to the large competence growth from wave 5 to wave 7, there were no common items in the two tests (except for the two items man5v181_sc1n7_c and man5z081_sc1n7_c, that were not intended to be used as common items). Therefore, to link those two measurements, an independent link sample of 507 five- and six-year-old children (50.3 % female) was selected. Both competence tests, the test for four-year-old children as well as the test for six-year-old children of starting cohort 1, were administered to these children. The two tests were presented in a random order to avoid order effects.

6.2.2 Results

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

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

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	21,155.6	47	21,249.64	21,448.38
Two-dimensional	22,939.4	49	23,037.43	23,244.63

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

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared the the longitudinal subsample of starting cohort 1. The differences in item difficulties between the link sample and starting cohort 1 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 11.

Analyses of differential item functioning between the link sample and starting cohort 1 were calculated and showed difference in logits from *Min* = -1.300 to *Max* = 1.002 in the test for four-year-old children, and from *Min* = -0.704 to *Max* = 0.484 in the test for six-year-old children. Since the differences in logits in the test for four-year-old children were very large, we only used items for calculating the linking correction term *c*, that had no difference in logits greater than |.4|. Therefore, 11 items of the test for four-year-old children and 3 items of the test for six-year-old children were excluded from calculating the linking correction term *c* (see Table 11; items marked with " * " were not used for the calculation). The mathematical

competence tests administered in the two waves were linked using the “mean/mean” method for the anchor-group design (see Fischer et al., 2016).

The correction term for wave 5 and 7 was calculated as *c* = 2.575. This correction term was subsequently added to each item difficulty parameter estimated in the test for six-year-old children in wave 7 (see Table 4) to derive the linked item parameters. The link error, reflecting the uncertainty in the linking process, was calculated according to equation 4 in Fischer et al. (2016) as 0.248 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

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

	Four-year old children				Six-year old children			
	Item	**Δσ**	$SE_{\Delta\sigma}$	***F***	**Item**	**Δσ**	$SE_{\Delta\sigma}$	***F***
1	mak1z17s_c*	-0.717	0.248	8.4	man7z211_c*	0.465	0.129	13.0
2	mak1z021_c	-0.058	0.167	0.1	man7z201_c	0.194	0.158	1.5
3	mak1v181_c	-0.090	0.134	0.4	man5v181_sc1n7_c	-0.246	0.145	2.9
4	mak1z161_c	-0.113	0.206	0.3	man7z101_c	0.353	0.137	6.7
5	mak1r14s_c*	0.654	0.156	17.5	man7r111_c	-0.119	0.128	0.9
6	mak1d191_c*	0.885	0.128	47.6	man7g051_c	0.314	0.184	2.9
7	mak1z051_c	0.289	0.131	4.9	man7g061_c	0.337	0.166	4.1
8	mak1g151_c*	0.672	0.160	17.8	man7v011_c	0.045	0.139	0.1
9	mak1r131_c	0.412	0.130	10.1	man7r151_c	-0.357	0.132	7.4
10	mak1g111_c*	1.040	0.143	52.6	man7g131_c	-0.290	0.136	4.5
11	mak1z121_c	-0.270	0.359	0.6	man7z041_c	-0.047	0.128	0.1
12	mak1v041_c	0.226	0.132	2.9	man7d071_c*	-0.724	0.130	31.1
13	mak1z081_c*	-1.294	0.162	63.8	man7g191_c	-0.224	0.128	3.1
14	mak1d091_c*	-1.044	0.596	3.1	man7r121_c	-0.202	0.144	1.9
15	mak1z201_c*	-0.931	0.165	31.8	man7z081_c	0.040	0.129	0.1
16	mak1g101_c	-0.050	0.254	0.0	man7v091_c	0.082	0.130	0.4
17	mak1z011_c*	-0.565	0.150	14.2	man7z171_c	-0.047	0.129	0.1
18	mak1r071_c	-0.219	0.160	1.9	man7d021_c*	-	-	-
19	mak1d031_c	0.313	0.129	5.9	man7z221_c	-0.049	0.152	0.1
20	mak1v061_c	-0.444	0.157	8.0	man5z081_sc1n7_c	0.346	0.141	6.0
21					man7g031_c*	0.417	0.132	10.1
22					man7z231_c	0.087	0.141	0.4
23					man7r181_c	-0.202	0.126	2.6
24					man7v161_c	-0.159	0.135	1.4

25	man7z141_c	0.137	0.141	0.9

***Note*.** $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in wave 5 / wave 7 and the link sample (positive values indicate easier items in the link sample); $SE\Delta\sigma$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is *F0154*(1,2201) = 56.60.
A non-significant test indicates measurement invariance.
* These items were excluded from calculating the correction term und link error.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs, man7_sc1 and man7_sc1u, including their respective standard errors, man7_sc2 and man7_sc2u. The WLE scores provided in man7_sc1u are linked to the underlying reference scale of newborns and can be used for longitudinal comparisons between the measurement points. In contrast, the WLE scores in man7_sc1 are not linked to the underlying reference scale of wave 5 and therefore should be used only for cross-sectional research questions. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A, the fixed item parameters for estimating the uncorrected WLE scores are provided in Appendix B. Students that did not take part in the test or those that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE scores for mathematical competence.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

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

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A., & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne – Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze & M. Grüßing (Eds.). *Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (pp. 313-327). Münster: Waxmann.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2012). Incorporating different response formats in the IRT-scaling model for competence data. Manuscript submitted for publication.

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

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

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online*, 5(2), 80-102.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, 5(2), 189-216.

Pohl, S., Haberkorn, K., Carstensen, C.H. (2015). *Measuring competencies across the lifespan – Challenges of linking test scores.* In M. Stemmler, A. von Eye, &W. Wiedermann (EDS), *Dependent data in social science research* (pp.281.308). Berlin, Germany: Springer.

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

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

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

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

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

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & von Maurice, J. (Eds.). *Education as a lifelong process: The German National Educational Panel Study (NEPS).* (pp. 67-86). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Appendix

Appendix A: ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort I – Wave 7

```
Title Starting Cohort I, Wave 7, MATHEMATICS: Rasch Model;
/* load data */
data filename.dat;
format pid 4-10 responses 12-35; /* insert number of columns with data*/
labels << labels.nam;

codes 0,1;

/* scoring */
score (0,1)              (0,1)              !item (1-24);

/* load linked item parameters */
import anchor_parameters << anchor_parameters.prm;

/* model specification */
set constraint=cases;
model item + item*step;

estimate;

/* save results */
show !estimates=latent >> filename.shw;
itanal >> filename.ita;
show cases !estimates=wle >> filename.wle;
```

Appendix B: Fixed Item Parameters

```
1       1.573  /* man7z211_c */
2       -0.006 /* man7z201_c */
3       0.665  /* man5v181_sc1n7_c */
4       1.546  /* man7z101_c */
5       1.748  /* man7r111_c */
6       4.217  /* man7g051_c */
7       3.790  /* man7g061_c */
8       0.927  /* man7v011_c */
9       2.569  /* man7r151_c */
10      3.251  /* man7g131_c */
11      1.982  /* man7z041_c */
12      2.388  /* man7d071_c */
13      2.328  /* man7g191_c */
14      3.662  /* man7r121_c */
15      2.108  /* man7z081_c */
16      1.988  /* man7v091_c */
17      1.880  /* man7z171_c */
18      3.495  /* man7z221_c */
19      0.749  /* man5z081_sc1n7_c */
20      1.117  /* man7g031_c */
21      3.166  /* man7z231_c */
22      2.168  /* man7r181_c */
23      3.054  /* man7v161_c */
24      0.718  /* man7z141_c */
```

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

Position	Item	Content area	Response format
1	man7z211_c	Sets, numbers, and operations	Short Constructed Response
2	man7z201_c	Sets, numbers, and operations	Matching
3	man5v181_sc1n7_c	Change and relationships	Short Constructed Response
4	man7z101_c	Sets, numbers, and operations	Short Constructed Response
5	man7r111_c	Space and shape	Simple Multiple-Choice
6	man7g051_c	Units and measuring	Sorting
7	man7g061_c	Units and measuring	Simple Multiple-Choice
8	man7v011_c	Change and relationships	Simple Multiple-Choice
9	man7r151_c	Space and shape	Simple Multiple-Choice
10	man7g131_c	Units and measuring	Short Constructed Response
11	man7z041_c	Sets, numbers, and operations	Short Constructed Response
12	man7d071_c	Data and chance	Simple Multiple-Choice
13	man7g191_c	Units and measuring	Short Constructed Response
14	man7r121_c	Space and shape	Short Constructed Response
15	man7z081_c	Sets, numbers, and operations	Short Constructed Response
16	man7v091_c	Change and relationships	Short Constructed Response
17	man7z171_c	Sets, numbers, and operations	Short Constructed Response
18[e]	man7d021_c	Data and chance	Simple Multiple-Choice
19	man7z221_c	Sets, numbers, and operations	Short Constructed Response
20	man5z081_sc1n7_c	Sets, numbers, and operations	Short Constructed Response
21	man7g031_c	Units and measuring	Sorting
22	man7z231_c	Sets, numbers, and operations	Simple Multiple-Choice
23	man7r181_c	Space and shape	Short Constructed Response
24	man7v161_c	Change and relationships	Short Constructed Response
25	man7z141_c	Sets, numbers, and operations	Short Constructed Response

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). [e]Excluded from the analyses due to unsatisfactory item fit.

Documentation of the modifications as of March 2022

	Date	Page	Modification
1.	March 2022	Page 12	Correction of decimal places for Discr. of item man7g191_c in Table 4.
2.	March 2022	Page 15-17	Corrected signs for the main effects in the text of Differential Item Functioning, changing description “male vs. female” in Table 6 and correction of rounding decimal places for the first item man7z211_c in Table 6.
3.	March 2022	Page 16	Corrected absolute number of persons for the category “test takers without a valid response” of the DIF variable “Books at home”