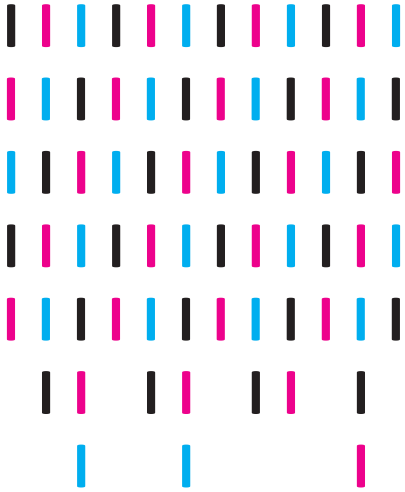

NEPS SURVEY PAPERS

Lara Aylin Petersen and Anna-Lena Gerken

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

NEPS Survey Paper No. 45
Bamberg, November 2018; Updated: September 2021

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS *Survey Paper* series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS *Survey Papers* are edited by a review board consisting of the scientific management of LIfBi and NEPS.

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

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

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 1 for Four-year old Children

Lara Aylin Petersen[1] and Anna-Lena Gerken[1]

[1]IPN – Leibniz Institute for Science and Mathematics Education at Kiel University

Email address of the lead author:

lpetersen@ipn.uni-kiel.de

Bibliographic Data:

Petersen, L. A. & Gerken, A.-L. (2018): *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 1 for Four-year old Children* (NEPS Survey Paper No. 45). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP45:3.0

Acknowledgements:

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

The present report has been modeled along previous reports published by NEPS. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduces *verbatim* from previous working papers (e.g., Schnittjer, 2018; Schnittjer & Gerken, 2017).

Please Note:

This NEPS Survey Paper has been modified in September 2021. You will find the documentation of the modifications on the last pages, following the appendix. The original paper can still be found using https://doi.org/10.5157/NEPS:SP45:1.0

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 1 for Four-year old children

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span and develops 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 procedures for a mathematical competence test that was administered in wave 5 (four year old children) of starting cohort 1 (newborns). The descriptive statistics for the data, the scaling model applied to estimate competence scores, and analyses performed to investigate the quality of the scale as well as the results of these analyses are explained. The mathematics test consists of 20 items which represent different content areas as well as different cognitive components and use different response formats. The test was administered to 2,138 four-year old children. A partial-credit model was used for scaling the data. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These results show that the items exhibited good item fit and measurement invariance across various subgroups. Moreover, the test shows a good reliability. As the correlations between the five content areas are high in a multidimensional model, the assumption of unidimensionality seems adequate. Overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. This paper describes the data available in the scientific use file and provides ConQuest-Syntax for scaling the data.

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

Contents

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, 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) and Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models 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 test. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scales are described in Pohl and Carstensen (2012).

In this paper, the results of these analyses are presented for mathematical competence of four-year old children (wave 5) in starting cohort 1 (newborns). First, the main concepts of the mathematical competence test are introduced. Then, the mathematical competence data of the fifth 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.

Please note that the analyses of this report are based on the data set available at some time different from data release. Due to data protection and data cleaning issues, the data set in the SUF may differ slightly from the dataset used for analyses in this paper. However, major changes in the presented 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 this paper, we briefly describe specific aspects of the mathematics test that are necessary for understanding the scaling results.

In the test, the items are not arranged in units. Thus, in the test, children usually face a certain situation followed by only one task related to it; in one instance there are three subtasks. Each of the items belongs to one of the following content areas:

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

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 five types of response formats: simple multiple-choice (MC), complex multiple choice (CMC), short constructed response (SCR), matching (M), and

sorting (S). In MC items, the test taker had to find the correct response option from several, usually four, available response options. In CMC items, a number of subtasks with two response options were presented. SCR items required the test taker to give mostly one-word answers, such as numbers. In sorting items, selection possibilities should be sorted into their correct order. Items with this response format were scored dichotomously as well for there is only one true order in each item. In matching items, the test taker was asked to match or puzzle some picture cards to given response options. The tasks were constructed in such a way as to enable a clear dichotomous scoring. Only one item was not scored dichotomous.

3 Data

3.1 The Design of the Study

The study assessed, among others, mathematical competence and executive functions. The test for mathematics competence was administered to all participants immediately after the test for executive functions. No multi-matrix design was applied regarding the choice and order of the items within the mathematics test. 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.

Table 1: Number of Items by Content Area

Content area	Frequency
Sets, numbers, and operations	8
Units and measuring	3
Space and shape	3
Change and relationships	3
Data and chance	3
Total number of items	20

Table 2: Number of items by Response Format

Response format	Frequency
Simple Multiple Choice	7
Complex Multiple Choice	1
Short Constructed Response	9
Matching	2
Sorting	1
Total number of items	20

The mathematics test included 20 items that represented the five content areas (see Appendix B) and the process-related components and used different response formats. The characteristics of the items are presented in the following tables. Table 1 shows the distribution of the items across the five content areas. Table 2 shows the distribution of the items across the five response formats described above.

3.2 Sample

Overall, the test was administered to 2,138 children. 114 of them gave fewer than three valid responses. Because no reliable competence 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 2,024 test takers (50% girls). A detailed description of the study design, the sample, and the administered instrument can be found on the NEPS website (http://www.neps-data.de).

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach and d) 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. Invalid responses occurred in just one item, where it was possible to start the next item although the current item hasn't been solved completely yet. Omitted items occurred if the child did not respond to an item. After three consecutively omitted items, the test administrator was instructed to abort the test. All subsequent items were coded as "not reached". Due to reasons like exhaustion or sudden consistent refusal, it may have occurred that not all children finished 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 persons were coping with the test. Missing responses per item were examined in order to evaluate how well the items functioned.

3.4 Scaling Model

Item and person parameters were estimated using the partial credit model (PCM; Masters, 1982) with Gauss-Hermite quadrature (15 nodes). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

Complex multiple choice (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. If at least one of the subtasks contained a missing response, the partial credit item was scored as missing. Categories of polytomous variables with less than N = 200 responses were collapsed in the analyses in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items. For the CMC item (mak1r14s_c) the lowest two categories were collapsed. Due to unsatisfactory step parameter it was scored dichotomously (see Table 4b). Simple MC items, M items, and S items were scored dichotomously as 0 for an incorrect and 1 for the correct response. Nearly all SCR items were scored dichotomously as well. Only the first Item (mak1m17s_c) was scored in four categories. To estimate item and person parameters, a scoring of 0.5 points for each category was applied (see Pohl & Carstensen, 2013 for studies on the scoring of different response formats).

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

3.5 Checking the Quality of the Scale

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

Before aggregating the subtasks of item mak1m17s_c to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC variables that were included in the final scaling model. The MC items consisted of one correct response option and three distractors (i.e., incorrect response options). The quality of the distractors within MC items 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 (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous and polytomous item to the partial credit model (Masters, 1982) was evaluated using four indices. Items with a WMNSQ > 1.15 or WMNSQ < .85 (*t*-value > $|6|$) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 or WMNSQ < .80 (*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 older ConQuest Versions than for this paper was used) greater than 0.30 were considered as good, greater than 0.20 as acceptable, and below 0.20 as problematic. Lastly, the fit was verified by using item characteristic curves. 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 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 the main effects of the subgroups as well as differential effects of the subgroups on item difficulty were estimated. Differences in the estimated item difficulties between the subgroups were evaluated. 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, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

The dimensionality of the mathematics test was evaluated by specifying a five-dimensional model based on the five content areas (see chapter 3.1). Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Quasi Monte Carlo integration in TAM (Kiefer, Robitzsch, & Wu, 2017) was used. To guarantee the compatibility with the multidimensional model, the unidimensional model was estimated in TAM as well. The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (15,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015). The two-parametric logistic model (2PL) was estimated in MDLTM (Matthias von Davier, 2005). To check the multidimensionality, the IRT models were estimated in TAM version 2.8-21 (Kiefer et al., 2017) in R version 3.4.3 (R Core Team, 2017).

4 Results

4.1 Missing Responses

4.1.1 Missing responses per person

The number of invalid responses per person was negligible, as can be seen in Figure 1. In fact, 99.4 % of the test takers gave no invalid response.

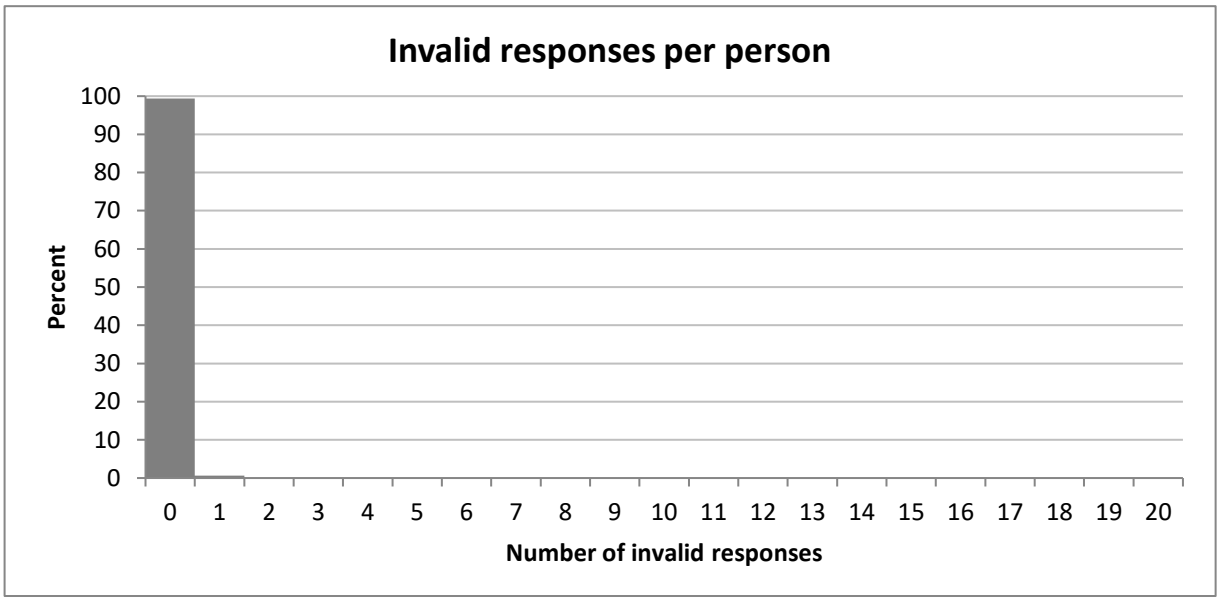

Figure 1. Number of invalid responses.

Missing responses may also occur when a child does not respond to an item (omit). The number of omitted responses per test taker is presented in Figure 2. It shows that 24.1 % of the children omitted no item and 6.2 % of the children omitted more than five items.

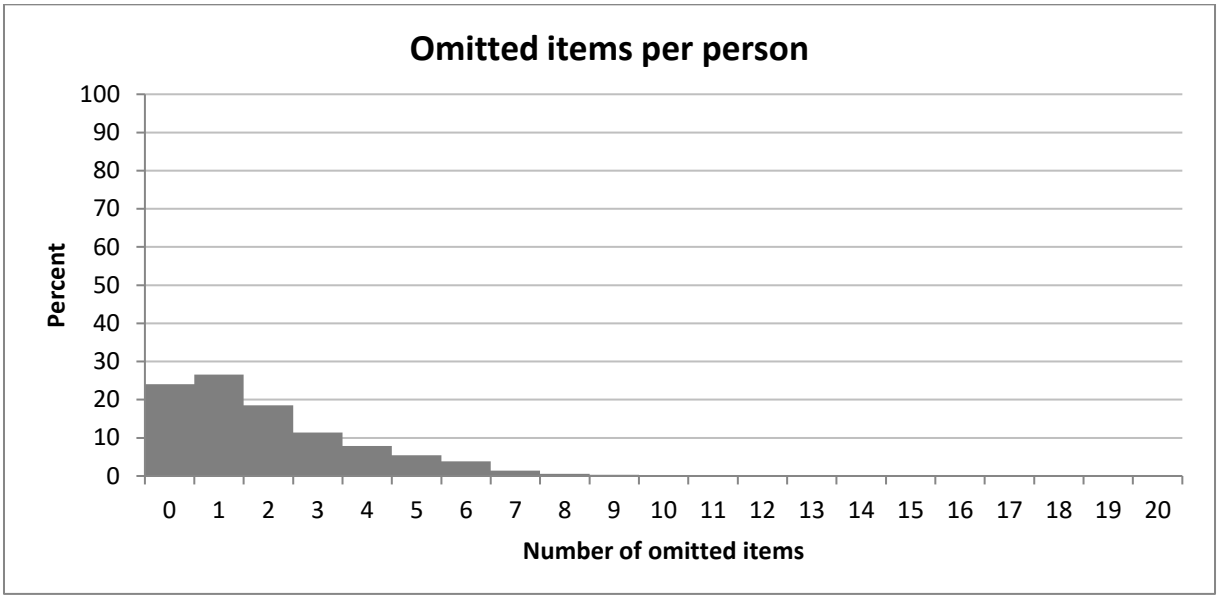

Figure 2. Number of omitted items.

All missing responses after three consecutively omitted items were defined as not-reached. Figure 3 shows the number of items that were not reached by a person. As can be seen, 97.6 % reached the end of the test. Therefore, only 2.4 % of the subjects did not reach the last item.

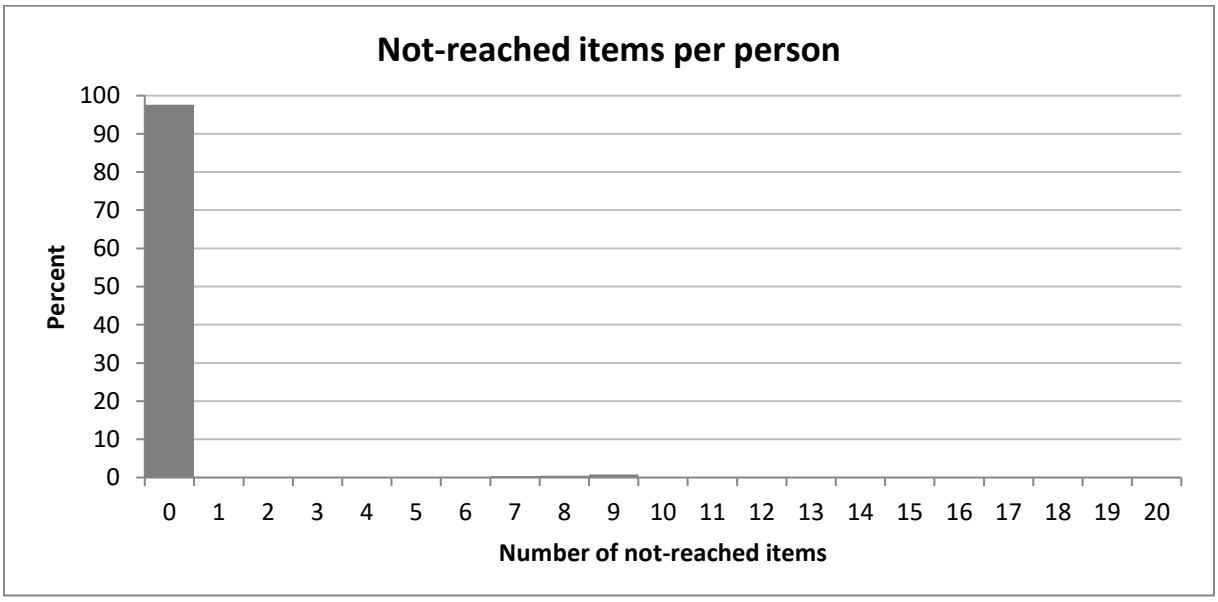

Figure 3. Number of not-reached items.

Figure 4 shows the number of test-aborted items which were defined in case the test administrator had to abort the test without three consecutively omitted items. In 98.7 % of all cases, no interruption was necessary. In only 1.3 % of the cases, the test had to be aborted.

Figure 4. Number of test-aborted items.

Figure 5 shows the total number of missing responses per person, which is the sum of invalid, omitted, not-reached and test-aborted missing responses. In total, 23.7 % of the subjects show no missing response. 8.7 % show more than five missing responses. Overall, it seems a typical amount of invalid, omitted, not reached and test-aborted items, considering the age of the children.

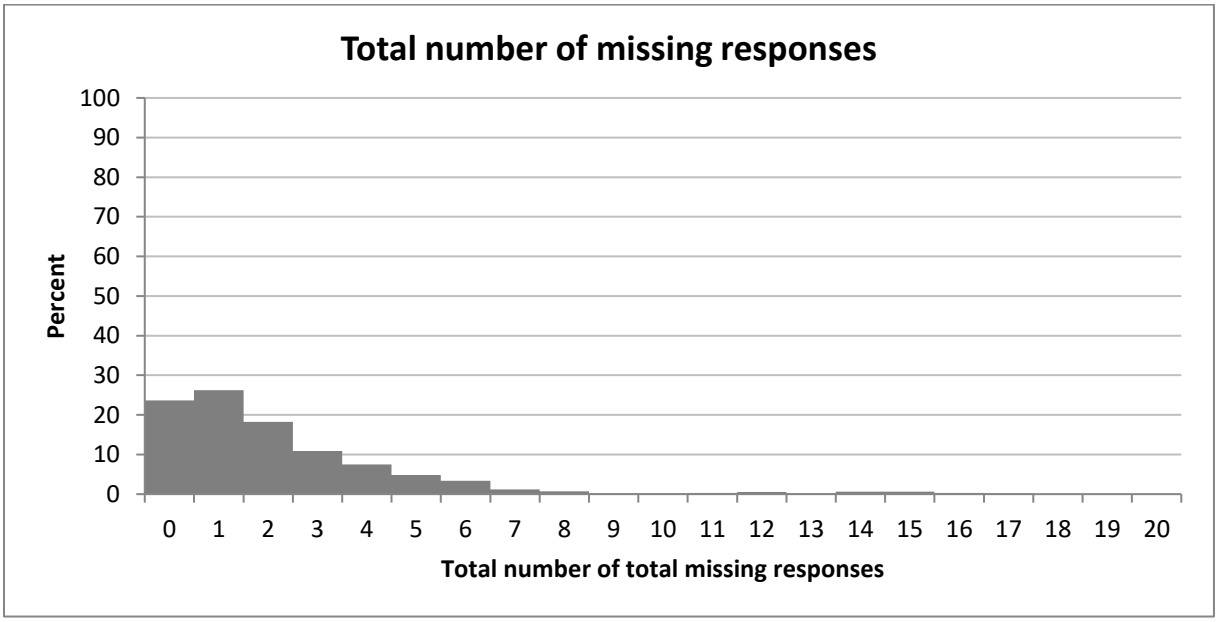

Figure 5. Total number of missing responses.

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 four types of missing responses. Overall, only one item had invalid responses, so this number is negligible.

The omission rates were good, except for eight items with an omission rate higher than 10.00 % (mak1z17s_c, mak1z021_c, mak1z051_c, mak1r131_c, mak1g111_c, mak1v041_c, mak1z081_c, mak1z011_c). The items mak1z17s_c (15.56 %) and mak1z021_c (10.38 %) were the first two items in the test. Considering the age of the four-year old children, it might be an age group specific behavior to react more reserved, indeed shy, at the beginning of the test. The highest omission rates occurred for item mak1z081_c (49.21 %) and item mak1g111_c (26.53 %). As these items have the highest difficulties (see Table 4a), the children might have preferred to skip the item rather than to guess. The other items mak1z051_c (13.29 %), mak1r131_c (12.25 %), mak1v041_c (14.38 %) and mak1z011_c (14.28 %) are also difficult items. All in all, the omission rates vary between 0.20 % (the easiest item mak1d091_c) and 49.21 %.

The number of persons that did not reach an item increased with the position of the item in the test up to 2.42 %. The percentage of test-aborted items also increased with the position of the item in the test up to 1.28 %.

Table 3: Percentages of Missing Values per item

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted items	Percentage of not-reached items	Percentage of test-aborted items
mak1z17s_c	1	1,709	0.00	15.56	0.00	0.00
mak1z021_c	2	1,814	0.00	10.38	0.00	0.00
mak1v181_c	3	1,931	0.00	4.59	0.00	0.00
mak1z161_c	4	1,895	0.00	6.32	0.00	0.05
mak1r14s_c	5	1,956	0.00	3.31	0.00	0.05
mak1d191_c	6	1,957	0.00	3.16	0.00	0.15
mak1z051_c	7	1,750	0.00	13.29	0.05	0.20
mak1g151_c	8	2,000	0.00	0.89	0.10	0.20
mak1r131_c	9	1,766	0.00	12.25	0.20	0.30
mak1g111_c	10	1,471	0.00	26.53	0.35	0.44
mak1z121_c	11	1,928	0.00	3.85	0.40	0.49
mak1v041_c	12	1,694	0.00	14.38	1.19	0.74
mak1z081_c	13	979	0.00	49.21	1.68	0.74
mak1d091_c	14	1,962	0.00	0.20	2.03	0.84
mak1z201_c	15	1,882	0.00	4.00	2.03	0.99
mak1g101_c	16	1,938	0.00	1.09	2.03	1.14
mak1z011_c	17	1,671	0.00	14.28	2.03	1.14
mak1r071_c	18	1,868	0.59	3.71	2.17	1.24
mak1d031_c	19	1,816	0.00	6.77	2.27	1.24

mak1v061_c	20	1,871	0.00	3.85	2.42	1.28

4.2 Parameter Estimates

4.2.1 Item parameters

In order to get a first rough descriptive measure of item difficulties and check for possible estimation problems, the relative frequency of the responses given before performing IRT analyses were evaluated. The percentage of persons correctly responding to an item (relative to all valid responses) varied between 12.87 % and 92.66 % across all items. On average, the rate of correct responses was 46.36 % (*SD* = 23.28 %).

The estimated item difficulties are depicted in Table 4a. The step parameter of the polytomous item mak1z17s_c is depicted in Table 4b. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. From a descriptive point of view, the items covered a wide range of difficulties. The estimated item difficulties varied between -2.830 (item mak1d091_c) and 2.279 (item mak1z081_c) with a mean of -0.124. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were small (*SE*(ß) ≤ 0.103).

Table 4a: Item Parameters

Item	Posi-tion	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.
mak1z17s_c	1	n.a.	-0.353	0.061	0.93	-2.6	0.53	0.697
mak1z021_c	2	59.98	-0.407	0.055	0.92	-4.3	0.54	1.304
mak1v181_c	3	35.63	0.706	0.055	1.03	1.7	0.38	0.623
mak1z161_c	4	73.25	-1.143	0.059	0.97	-1.0	0.46	0.985
mak1r14s_c	5	68.20	-0.872	0.056	0.98	-0.7	0.45	0.847
mak1d191_c	6	50.69	-0.026	0.053	1.17	9.8	0.22	0.150
mak1z051_c	7	36.97	0.669	0.057	0.96	-2.2	0.48	0.994
mak1g151_c	8	72.60	-1.121	0.057	1.17	6.2	0.20	0.164
mak1r131_c	9	37.32	0.627	0.057	1.02	1.0	0.39	0.623
mak1g111_c	10	18.76	1.765	0.074	1.10	2.4	0.19	0.204
mak1z121_c	11	89.83	-2.434	0.081	0.94	-1.0	0.39	1.411
mak1v041_c	12	39.02	0.558	0.057	1.02	1.2	0.41	0.695
mak1z081_c	13	12.87	2.279	0.103	0.92	-1.2	0.43	1.829
mak1d091_c	14	92.66	-2.830	0.092	0.98	-0.3	0.31	1.023
mak1z201_c	15	41.34	0.436	0.054	0.87	-7.8	0.61	2.153
mak1g101_c	16	83.44	-1.835	0.068	1.02	0.4	0.34	0.740

mak1z011_c	17	38.12	0.597	0.058	0.90	-5.1	0.56	1.633
mak1r071_c	18	53.85	-0.153	0.054	1.06	3.6	0.36	0.494
mak1d031_c	19	32.32	0.870	0.057	1.02	0.8	0.38	0.649
mak1v061_c	20	46.50	0.185	0.054	0.96	-2.5	0.50	1.023

Note. Difficulty = Item difficulty/location parameter, SE = Standard error of item difficulty/location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).

Percent correct scores are not informative for polytomous item scores. These are denoted by n.a.
For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

Table 4b: Step Parameters of the Polytomous Item

Item	Position in the test	step 1 (SE)	step 2 (SE)	Step 3
mak1z17s_c	1	-0.575 (0.066)	-0.165 (0.068)	0.740

Note: mak1r14s_c was scored dichotomously and therefore cannot be found in Table 4b but in Table 4a.

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person´s abilities (WLEs) to evaluate the appropriateness of the test for the specific target population.

In Figure 6, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers´ ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.762, indicating that the test differentiated well between subjects.

The item difficulties ranged from -2.83 (item mak1d091_c) to 2.28 (item mak1z081_c). Therefore, a rather broad range was covered, which is also illustrated in figure 6. The reliability of the test (EAP/PV reliability = 0.70. WLE reliability = 0.67) was acceptable good. In addition to the wide range of the ability distribution, there was also a somewhat equal distribution of easy and difficult items. Therefore, person abilities in high- and low ability regions should be measured relative precisely.

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

Figure 6: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 10.9 cases. The difficulty of the items is depicted on the right-hand side of the graph. Each number represents one item (see Table 4a).

4.3 Quality of the test

4.3.1 Distractor analyses

To investigate how well the distractors performed in the test, the point-biserial correlations between selecting each incorrect response (distractor) in dichotomous items and the child's total correct scores was evaluated. This distractor analysis was performed on the basis of preliminary analyses treating all subtasks of the CMC item as single items. The point-biserial correlations for the distractors ranged from -0.50 to -0.02 with a mean of -0.24. These results indicate that the distractors worked well. In contrast, the point-biserial correlations between selecting the correct response and the child's total correct scores ranged from 0.07 to 0.50 with a mean of 0.28 indicating that more proficient 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 (dichotomous items only)	Incorrect responses (dichotomous items only)
Mean	0.28	-0.24
Minimum	0.07	-0.50
Maximum	0.50	-0.02

4.3.2 Item fit

The evaluation of the item fit is based on the final scaling model, the partial credit model, using the dichotomous and polytomous items. Altogether, item fit can be considered to be good (see Table 4a and 4b). Values of the WMNSQ were close to one (mean = 1.00) with the lowest value being 0.87 (item mak1z201_c) and the highest 1.17 (item mak1d191_c and item mak1g151_c).

Only one item (mak1d191_c) showed a noticeable *t*-values above |8.0| with a *t*-value of 9.8 and WMNSQ of 1.17, as can be seen in Table 4a. The point-biserial correlation between the item scores and the total scores was adequate (.22). The item characteristic curve (ICC) followed the trend of a little underfit, because it was relatively flat in comparison to the model probability curve. Although item mak1d191_c showed a little underfit and a *t*-value above |8|the item was included in further analyses because the fit indices showed only a little misfit and the associated content area (data and chance) was very important for a balanced representation of the content areas in this age group.

Item mak1g151_c also showed a little deviation of the model probability curve due to the relatively flat item characteristic curve. However, due to the acceptable *t*-value (6.2) and acceptable point biserial correlation, we appreciate the item fit as given.

Although Item mak1z201_c had an overfit WMNSQ, it showed an ordinary item characteristic curve (ICC) and the highest point-biserial correlation (.61), so we appreciate the item fit as good. The other item characteristic curves support a good fit of the other items.

Overall, the point-biserial correlations between the item scores and the total scores ranged from 0.19 (item mak1g111_c) to 0.61 (item mak1z201_c) with a mean of 0.41. Due to the large sample size and 0.87 ≤ WMNSQ ≤ 1.17, items showed satisfactory item fit in the test.

4.3.3 Differential item functioning

We examined test fairness for different groups (i.e. measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables gender and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 6 shows the difference between the estimated difficulties of the items in different subgroups. For example, the column Gender "male vs. female" indicates the differences in difficulty β(male) – β(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females.

Table 6: Differential Item Functioning

Item	Position	Gender	Migration status
		male vs. female	without vs. with
mak1z17s_c	1	0.386	-0.202
mak1z021_c	2	0.120	0.040
mak1v181_c	3	0.216	0.138
mak1z161_c	4	-0.030	-0.032
mak1r14s_c	5	-0.094	-0.302
mak1d191_c	6	-0.248	0.188
mak1z051_c	7	-0.040	0.124
mak1g151_c	8	-0.038	0.314
mak1r131_c	9	-0.066	0.092
mak1g111_c	10	-0.472	-0.166
mak1z121_c	11	0.136	-0.476
mak1v041_c	12	-0.006	-0.182
mak1z081_c	13	-0.478	-0.216
mak1d091_c	14	-0.802	-0.344
mak1z201_c	15	0.016	-0.172
mak1g101_c	16	0.004	-0.332
mak1z011_c	17	-0.046	-0.038
mak1r071_c	18	-0.158	0.008
mak1d031_c	19	0.086	0.304
mak1v061_c	20	0.496	0.346
Main effect (model with DIF)		**0.168**	**-0.296**
Main effect (model without DIF)		**0.168**	**-0.302**

Gender: Overall, 1,009 (49.85 %) of the children were male and 1,015 (50.15 %) were female. On average, female children exhibited a slightly higher mathematical competence than male children (main effect = 0.168 logits, Cohen´s *d* = -0.193). Four items showed DIFs greater than 0.4 logits (mak1g111_c, mak1z081_c, mak1d091_c, mak1v061_c). However, with only one DIF being above 0.6 logits (-0.802 for item mak1d091_c), the differences between the two groups were not considered severe.

Migration: There were 1,424 (70.36 %) participants without migration background, 550 (27.17 %) participants with migration background, and 50 (2.47 %) participants without a valid response. Only the first two groups were used for investigating DIF of migration. On average, children without migration background performed better in the mathematics test than those with migration background (main effect = -0.296 logits, Cohen's *d* = 0.352). There was no considerable DIF comparing the two groups. However, one item mak1z121_c showed a small DIF of 0.48 between the groups.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models, which allow DIFs to those that only estimate main effects without allowing DIFs (see Table 7). Akaike's (1974) information criterion (AIC) favored the models estimating DIF for both DIF variables. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents from overparametrization of models. Using BIC, the more parsimonious models including only the main effects were preferred for all DIF variables.

Table 7: Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	42521.782	24	42569.78	42704.49
	DIF	42430.391	44	42518.39	42765.36
Migration	Main effect	41416.185	24	41464.18	41598.29
	DIF	41355.025	44	41443.02	41688.89

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; Birnbaum, 1968; also known as generalized partial credit model) was fitted to the data. The estimated discrimination parameters are depicted in Table 4a. They ranged between 0.15 (item mak1d191_c) to 2.15 (item mak1z201_c). Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 41879.88, BIC = 42194.19, number of parameters = 56) as compared to the 1PL Rasch model (AIC = 42572.01, BIC = 42774.07, number of parameters = 36). Despite the empirical preference for the 2PL model, the Rasch model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the Rasch model (also known as partial credit model, 1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework. Note that these calculations were performed in MDLTM (see Davier, 2005). Therefore, other results from other software programs for AIC and BIC using the Rasch model might differ from these results (see 4.3.4).

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 method implemented in TAM in R version 3.4.3 (R Core Team, 2017) was used. The number of nodes used in TAM was set to 15,000. The standard deviations and correlations of the five dimensions are shown in Table 8. Model fit between the unidimensional and the five-dimensional model is compared in Table 9.

Three of the five dimensions exhibited a relatively good variance. Dimension two "Units and measurement" had the smallest variance. The difficulties of the three items that could be classified to this dimension varied from -1.835 to 1.765, so the difficulties were balanced and could not explain the small variance.

In comparison, dimension one "Sets, numbers and operations" had the highest variance and its item difficulties varied from -2.434 to 2.279. However, it should also be mentioned that dimension one had more than twice as many items, so there was a better chance for dimension one to have a higher variance than a dimension with fewer items.

As expected, the correlations between the five dimensions were relatively high, but also somewhat heterogeneous, varying between 0.573 and 0.929. However, they deviated from a perfect correlation (i.e., they were lower than $r = .95$; see Carstensen, 2013).

Table 8: Results of Five-Dimensional Scaling

	Sets, numbers, and operations	**Units and measurement**	**Space and shape**	**Change and relation-ships**	**Data and chance**
Sets, numbers, and operations (8 items)	2.255				
Units and measuring (3 items)	0.744	0.219			
Space and shape (3 items)	0.782	0.829	0.767		
Change and relationships (3 items)	0.929	0.811	0.865	0.777	
Data and chance (3 items)	0.761	0.573	0.660	0.736	0.415

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

Still, according to model fit indices, the five-dimensional model fitted the data slightly better (AIC= 41997.21, BIC=42071.21, number of parameters = 37) than the unidimensional model (AIC = 42580.85, BIC= 42709.95, number of parameters = 23). These results indicate that the five content areas measure a common construct, although they are not completely unidimensional as the correlations were not satisfyingly high.

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

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	42534.85	23	42580.85	42709.95
Five-dimensional	41997.21	37	42071.21	42278.89

Note. Contrary to the calculations for the 1PL and 2PL models (see 4.3.3), results in this Table were achieved by using TAM in R.

5 Discussion

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

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for different response formats, as well as the aggregated polytomous SCR item, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the test's dimensionality as well as local item dependence.

However, all kinds of missing responses were negligible considering that the rather high omission rates could be explained by an age specific behavior. Furthermore, item as well as test quality were examined. As indicated by various fit criteria —e.g., WMNSQ, *t*-value of the WMNSQ and ICC— the items exhibited good fits, except item mak1d192_c which showed two of four fit parameters above a good valuation, but was included due to the content-related importance. Moreover, discrimination values of the items (either estimated in a 2PL model or as a correlation of the item score with the total score) were acceptable. The test also had an acceptable reliability (EAP/PV reliability = 0.70. WLE reliability = 0.67). It distinguished well between test takers. As a consequence, ability estimates should be relatively precise for all children.

Different variables were used for testing measurement invariance. No considerable DIF became evident for these variables, indicating that the test was fair for the examined subgroups. Fitting a five-dimensional model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model-fit than the unidimensional model. Nevertheless, due to the high correlations between the dimensions and variances the results indicate that the unidimensional model described the data reasonably well.

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

6 Data in the Scientific Use File: Naming Conventions

The SUF contains 20 items that are either scored as dichotomous variables (MC and SCR items) with 0 indicating an incorrect response and 1 indicating a correct response or scored as a polytomous variable (corresponding to the CMC items) indicating the number of correctly

answered subtasks. Scored items are marked with a '_c' at the end of the variable name. Manifest scale scores are provided in the form of WLE estimates (mak1_sc1) including the respective standard error (mak1_sc2). The ConQuest Syntax for estimating the WLE scores from the items are provided in the Appendix A. Test takers that did not take part in the test or that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE score for mathematical competence. Users interested in investigating 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 Contro, 19,* 716–722.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 397–479). Reading, MA: MIT Press.

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.

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. (2016). TAM: Test Analysis Modules. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 1.995-0).

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

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

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

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

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

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

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 A

ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort I – Four-year old children

```
Title SC I Four-year old children MATHEMATICS: Partial Credit Model;

/* load data */

data filename.dat;
format pid 4-10 responses 12-31;  /* insert number of columns with
data*/

labels << filename_with_labels.nam;
codes 0,1,2,3;

recode (0,1,2,3)(0,0,0,1)  !item (5);

/* scoring */

score (0,1,2,3)       (0,0.5,1,1.5)   !item (1);
score (0,1)           (0,1)           !item(2-20);

/* model specification and estimate model*/

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

/* save results to file */

show cases !estimates=wle >> filename.wle;
show  >> filename.shw;
itanal >> filename.ita;
```

Appendix B: Content Areas of Items in the Mathematics Test Four-year old children

Position	Item	Content area
1	mak1z17s_c	Sets, numbers, and operations
2	mak1z021_c	Sets, numbers, and operations
3	mak1v181_c	Change and relationships
4	mak1z161_c	Sets, numbers, and operations
5	mak1r14s_c	Space and shape
6	mak1d191_c	Data and chance
7	mak1z051_c	Sets, numbers, and operations
8	mak1g151_c	Units and measuring
9	mak1r131_c	Space and shape
10	mak1g111_c	Units and measuring
11	mak1z121_c	Sets, numbers, and operations
12	mak1v041_c	Change and relationships
13	mak1z081_c	Sets, numbers, and operations
14	mak1d091_c	Data and chance
15	mak1z201_c	Sets, numbers, and operations
16	mak1g101_c	Units and measuring
17	mak1z011_c	Sets, numbers, and operations
18	mak1r071_c	Space and shape
19	mak1d031_c	Data and chance
20	mak1v061_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 de Ham, 2016).

List of modifications as of September 2021 (V2.0)

	Date	Page	Modification
1.	September 2021	Page 17	Corrected sentence for Gender DIF interpretation. Female children exhibited a slightly higher mathematical competence than male children.
2.	September 2021	Page 19	Naming Conventions were corrected. The CMC Item was added in the first sentence (19 dichotomous items and one polytomous CMC item in the SUF).

The original paper (November 2018) can still be found using https://doi.org/10.5157/NEPS:SP45:1.0

List of modifications as of September 2021 (V3.0)

	Date	Page	Modification
1.	September 2021	Page 1	DOI has been corrected.

The previous version of this paper (September 2021) can still be found using https://doi.org/10.5157/LIfBi:SP45:2.0