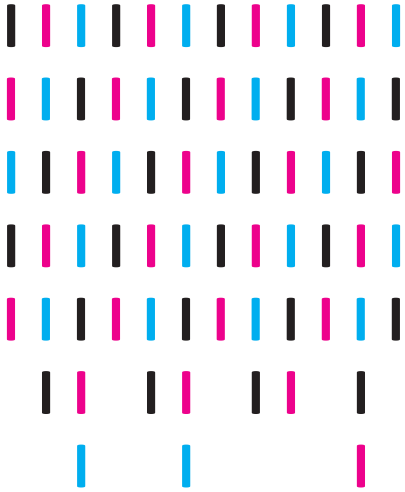

NEPS SURVEY PAPERS

Insa Schnittjer

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 1

NEPS Survey Paper No. 44
Bamberg, June 2018

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).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and 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 available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1

Insa Schnittjer[1, 2]

[1]*IPN – Leibniz Institute for Science and Mathematics Education at Kiel University*
[2]*University of Koblenz-Landau*

Email address of the lead author:

schnittjer@uni-landau.de

Bibliographic Data:

Schnittjer, I. (2018): *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1* (NEPS Survey Paper No. 44). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Acknowledgements:

I would like to thank Steffi Pohl and Kerstin Haberkorn for developing and providing standards for the technical reports, and Anna-Lena Gerken, Timo Gnambs, Anna Scharl, and Luise Fischer for assistance in scaling the data as well as 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 precious working papers (e.g., Schnittjer & Gerken, 2017).

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences 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) have been performed. This paper describes the data and scaling procedure for the mathematical competence test in grade 1 of starting cohort 2 (kindergarten). The mathematics test contained 22 items with different response formats representing different content areas as well as cognitive components while using different response formats. The test was administered to 6,510 children in first grade. Their responses were scaled using the partial credit 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 an acceptable reliability, good item fit and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. As the correlations between the five content areas were very 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. Besides the scaling results, this paper also describes the data available in the Scientific Use File and provides ConQuest-Syntax for scaling the data – including the necessary item parameters.

Keywords

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

Content

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competence domains measured in the NEPS is given by Weinert et al. (2011) and Fuß, Gnambs, Lockl, and Attig (2016).

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

In this paper, the results of these analyses are presented for mathematical competence in grade 1 of starting cohort 2 (kindergarten). First, the main concepts of the mathematical competence test are introduced. Subsequently, the mathematical competence data of the third wave of starting cohort 2 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the Scientific Use File (SUF) is presented.

Please note that the analyses in this report are based on the data 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 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 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 six cognitive components required for solving the tasks as a second and independent dimension. These are distributed across the items.

In the mathematics test there are two types of response formats. These are simple multiple-choice (MC) and complex multiple-choice (CMC). In MC items the test taker has to find the correct answer from several, usually four, response options, whereas in CMC items a number of subtasks with two response options are presented.

3 Data

3.1 The Design of the Study

The main study in 2013 assessed different competence domains including scientific literacy, mathematical competence as well as procedural metacognition (meta-p), receptive vocabulary (VOC), receptive grammatical competencies (GRA), and metacognition (MC). The competence tests for these domains took place on two testing days. On one testing day, the children's mathematical competence and scientific literacy were assessed; the other competence domains were assessed on the other testing day. In order to investigate the effects of test duration and to control possible effects of position and order, the two domains as well as the test days were rotated. For this purpose, the sample was split into four groups receiving the tests in different orders. Assignment to test booklets was random. Therefore, one testing group first completed the science test followed by the mathematics test (including procedural metacognition), while the other group completed the two tests in the opposite order. Moreover, one group started with these two tests on the first testing day, the other group started with receptive vocabulary, receptive grammatical competencies, and metacognition followed by either the science test and the mathematics test (including the procedural metacognition) or, in the opposite order, the mathematics test and the science test on the second testing day (see Table 1). Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test. All subjects received the same mathematics items in the same order. A special challenge of this test was to take into account that the reading competences of this age group are very heterogeneous. Regarding the status of the early readers, all items – including the response options – were read out to the children by a test instructor. There were up to 14 children bundled in one test session. As a consequence, it was up to the test instructors to keep the time limits for the whole group in mind.

Table 1: Design of the study.

Testing day	Rotation 1	Rotation 2	Rotation 3	Rotation 4
1st	Math (+meta-p) Science (+meta-p)	Science (+meta-p) Math (+meta-p)	VOC GRA MC	VOC GRA MC
2nd	VOC GRA MC	VOC GRA MC	Science (+meta-p) Math (+meta-p)	Math (+meta-p) Science(+meta-p)

Note. Math – mathematical competence, Science – Scientific literacy, meta-p – procedural metacognition for the respective competence, VOC – vocabulary, GRA – grammatical competencies, MC – metacognition.

The mathematics competence test for first grade students consisted of 22 items which represent different content-related and process-related components and used different response formats. One item was eliminated from further analysis because of insufficient item discrimination (see 4.3.4 for an explanation). The characteristics of the remaining 21 items are

summarized in the following tables. Table 2 shows the distribution of the five content areas, whereas Table 3 shows the distribution of response formats.

Three of the CMC items consisted of four subtasks. One subtask from item mag1v01s_c was excluded from analyses due to unsatisfactory item fit, resulting in three subtasks. The item mag1r19s_c consisted of five subtasks. Due to insufficient frequencies in categories (i.e., less than 200 test takers), categories had to be collapsed, and, therefore, this item was scored dichotomously.

Table 2: Number of Items by Content Areas

Content area	Frequency
Sets, numbers and operations	6
Units and measuring	3
Space and shape	3
Change and relationships	4
Data and chance	5
Total number of items	21

Table 3: Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	17
Complex Multiple-Choice	4
Total number of items	21

3.2 Sample

Overall, 6,510[1] students took the mathematics test. Twenty-two of them gave less than three valid responses. No reliable mathematical competence score can be estimated based on such few responses; these 22 cases were therefore excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 6,488 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) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and finally e) multiple kinds of missing responses that occur within one item and are not determined.

In this study, all respondents received the same set of items. As a consequence, there are no items that were not administered to a person. Invalid responses occurred, for example, when students selected two response options where only one was required. Omitted items occurred if test takers skipped some items. Regarding the fact that the items were read aloud to the children, missing responses due to items that have not been reached may have occurred to the whole group of up to 14 children. It was not possible to reconstruct whether some test

[1] Note that these numbers may differ from those found in the SUF. This is due to still ongoing data protection and data cleaning issues.

administrators could not finish the last item instruction due to time limits, or some children did not follow the instructions to the end of the test. All missing responses after the last valid response were coded as not reached, regardless of whether it was due to too slow instructions given from the test administrator, or due to individual reasons.

As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a non-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the subjects 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 partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

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 CMC item was scored as missing.

Categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category. This happened for three out of the four CMC items. For the item mag1z20s_c the three lowest categories were collapsed and for item mag1d09s_c the two lowest categories were collapsed. Finally, item mag1r19s_c was scored dichotomously, because the four lower categories had to be collapsed.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items 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 NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

Before aggregating the subtasks of CMC items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the single 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 error (WMNSQ), the respective *t*-value, point-biserial correlations of the responses with total correct score, and the item characteristic curve. 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 (incorrect response options). The quality of the distractors within MC items was evaluated using the point-biserial correlation between selecting an incorrect response and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.2 (*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) of greater than 0.3 were considered good, greater than 0.2 acceptable, and below 0.2 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 students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background, as well as test position (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 but 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 scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the mathematics test was evaluated by specifying a five-dimensional model based on the five content areas. Every item was assigned to one content area (between-item multidimensionality). To estimate this multidimensional model, the package TAM for the statistical software R 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 (10,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The IRT models were estimated in ConQuest version 4.5.2 (Adams, Wu, & Wilson, 2015). The 2PL model was estimated in MDLTM (Matthias von Davier, 2005). To check the multidimensionality, the IRT models were estimated in TAM version 2.4-9 (Kiefer, Robitzsch, & Wu, 2016) in R version 3.4.1 (R Core Tam, 2016) using the Quasi Monte Carlo integration with 10,000 nodes.

4 Responses

3.7 Missing Responses

4.1.1 Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was quite small. In fact, 83.4% of test takers gave no invalid response at all. Only 0.7% of the subjects had more than three invalid responses.

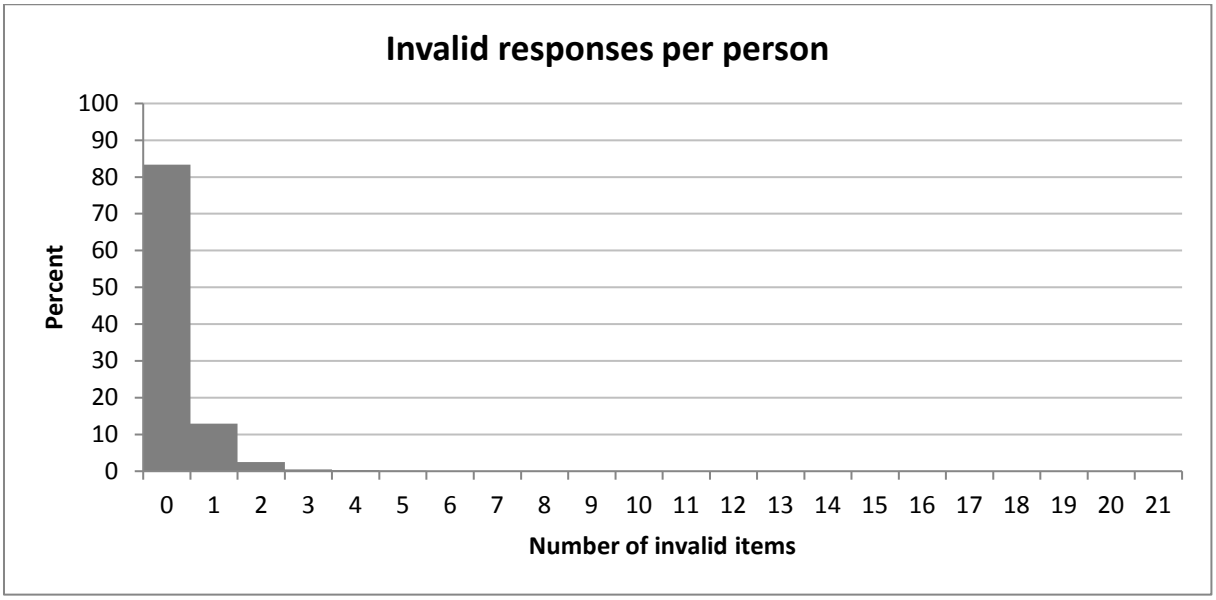

Figure 1: Number of invalid responses

Missing responses may also occur when test takers skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. It shows that 60.3% of the respondents omitted no item at all. 2.6% of the test takers omitted more than 5 items.

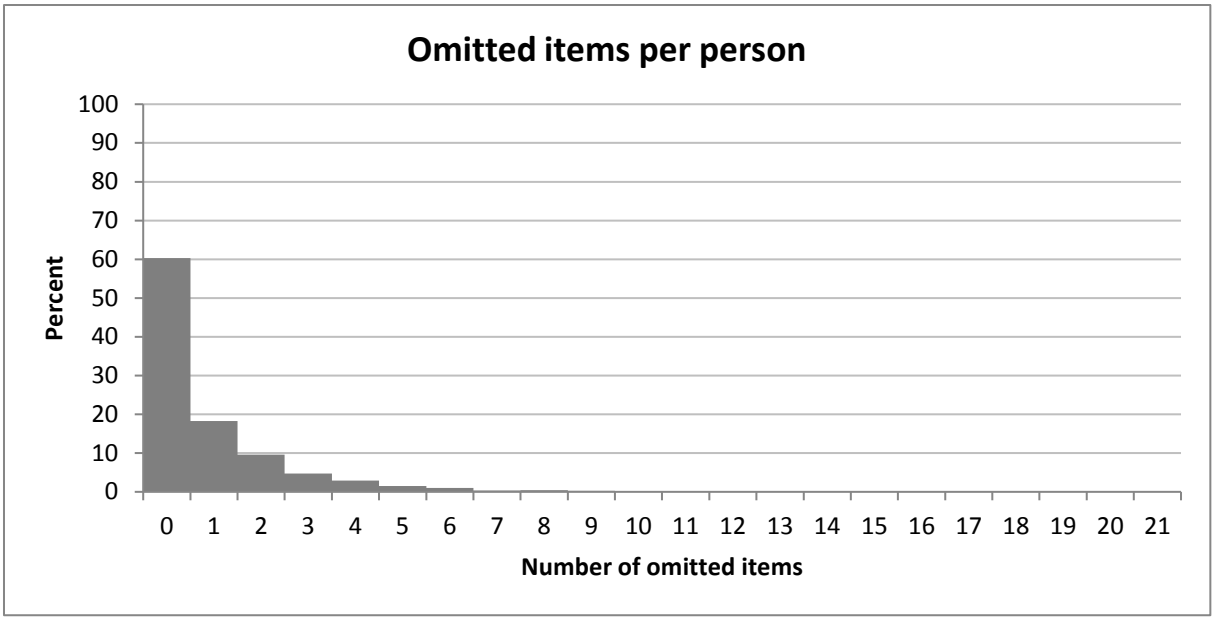

Figure 2: Number of omitted items

All missing responses after the last valid response are defined as not reached. Figure 3 shows the number of items that were not reached by a person, regardless of whether it was the individual test taker that did not complete the test or whether the test administrator did not keep a reasonable pace in order to finish it within the time limit. As can be seen, only 91.3% of the test takers reached the end of the test, 6.2% did not reach one to five items and only 2.5% of the children did not reach more than five items.

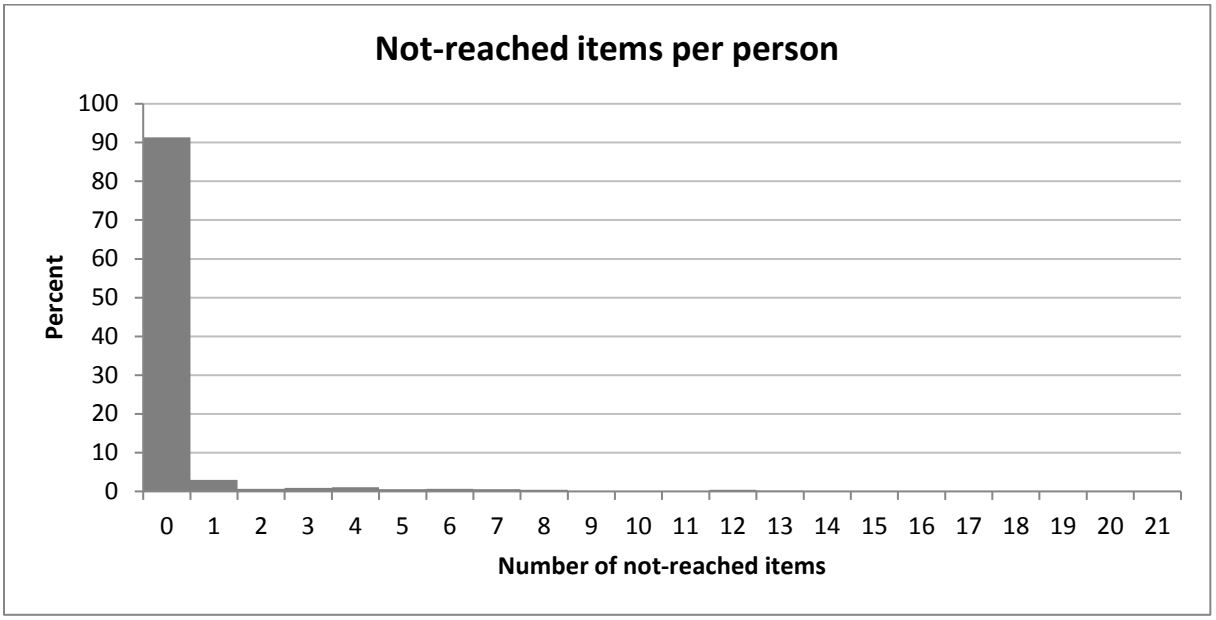

Figure 3: Number of not-reached items

Figure 4 shows the total number of missing responses per person which is the sum of invalid, omitted, not-reached, and not-determinable missing responses. In total, 42.4% of the test takers showed no missing response at all, whereas 7.64% showed more than five missing responses.

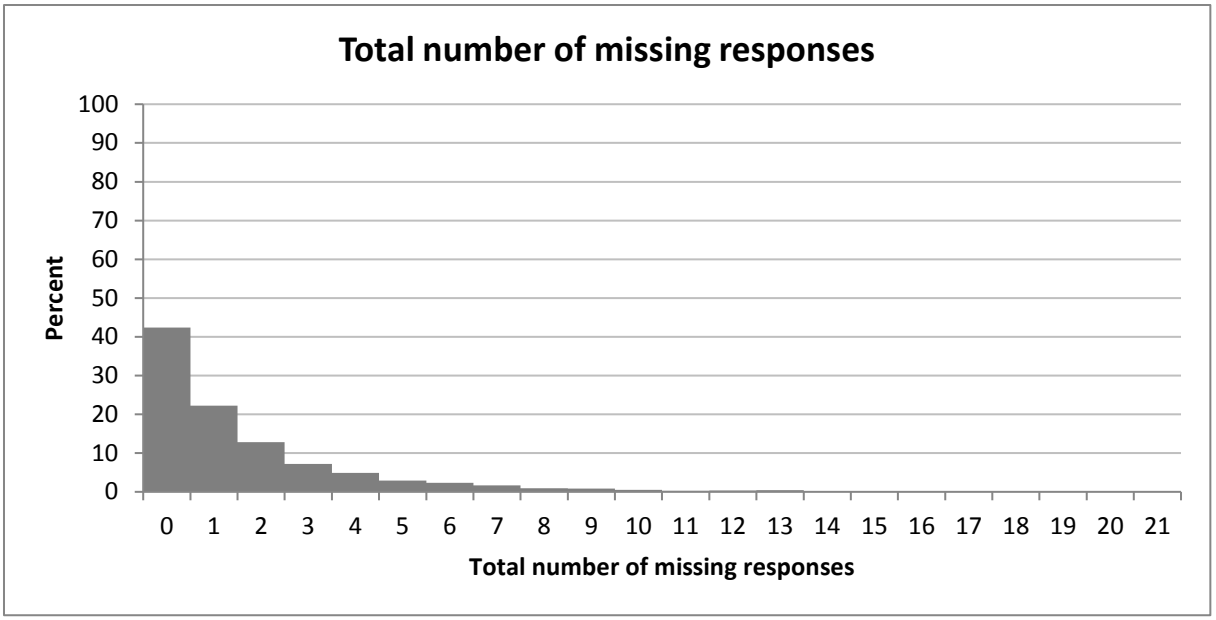

Figure 4: Total number of missing responses

Overall, there is a negligible number of not-reached and an insignificant number of invalid items. The number of omitted items is reasonable.

4.1.2 Missing responses per item

Table 3 shows the number of valid responses for each item as well as the percentage of missing responses. Overall, the number of invalid responses per item was very small. The omission rates were acceptable, varying between 1.08% (item mag1d081_c) and 8.34% (item mag1v021_c), except for one item that had an omission rate above 10% (item mag1d132_c). This highest omission rate (13.07%) appeared in the only item that was placed on the bottom of a page below another item, while all other items were placed on separate pages. The number of persons that did not reach an item increased with the position of the item in the test to 8.69%.

The total number of missing responses per item varied between 1.91% (item mag1g171_c) and 25.98% (item mag1r19s_c). Consider also that this last item was not only the last item of the test, but also a CMC item even though it was scored dichotomously for this analysis. Children might have been exhausted, or the whole test-group might not have reached this item in time and the instructions were more complex than for MC items.

Table 3: Missing values in the items

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items
mag1v051_c	1	6,101	0.83	5.13	0.00
mag1r141_c	2	6,341	0.51	1.76	0.00
mag1g171_c	3	6,364	0.60	1.31	0.00
mag1d131_c	4	5,826	2.42	7.77	0.02
mag1d132_c	5	5,432	3.19	13.07	0.02

mag1z061_c	6	6,066	0.96	5.53	0.02
mag1v01s_c	7	5,921	2.57	4.65	0.02
mag1z20s_c	9	6,218	1.90	2.16	0.02
mag1d09s_c	10	6,135	0.99	3.55	0.17
mag1z121_c	11	6,129	2.05	2.87	0.62
mag1g181_c	12	6,125	0.39	4.55	0.66
mag1d081_c	13	6,347	0.34	1.08	0.76
mag1r151_c	14	6,314	0.72	1.08	0.88
mag1z111_c	15	5,878	1.14	6.95	1.31
mag1v021_c	16	5,791	0.54	8.34	1.86
mag1z071_c	17	5,883	0.34	6.49	2.50
mag1d041_c	18	6,196	0.25	1.16	3.10
mag1g031_c	19	5,993	0.34	3.10	4.19
mag1z161_c	20	5,792	0.32	5.32	5.09
mag1v101_c	21	6,010	0.37	1.26	5.73
mag1r19s_c	22	4,802	2.25	1.66	8.69

Note. Item 8 was removed from further analyses due to unsatisfactory item fit.

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 was evaluated before performing any IRT analyses. Using each subtask of a CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 11.27% and 98.19% across all items. On average, the rate of correct responses was 65.97% (*SD* = 21.22%). One subtask of mag1v01s_c showed a *t*-value above |25|and therefore had to be removed from further analyses. From a descriptive point of view, the remaining items covered an acceptable wide range of difficulties with a tendency to being easy.

The estimated item difficulties (for dichotomous variables) are summarized in Table 4a. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are also given in Table 4a. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The step parameters of the polytomous items are depicted in Table 4b.

The estimated item difficulties varied between -1.81 (item mag1g171_c) and 2.42 (item mag1z121_c) with a mean of -0.241. Overall, the item difficulties are reasonably well distributed around the students with medium ability, yet some gaps appear, increasingly frequent towards the edges. However, there was only one item with a difficulty rating above 2 logits, and none rating in the lower section below -2 logits. Still, the test showed the

heterogeneity of the sample. Due to the large sample size, the standard error of the estimated item difficulties (column 4) was very small (*SE* ≤ 0.045).

Four items showed noticeable *t*-values between |6.0| and |8.2|. Another four items showed considerable *t*-values above |8.2|. Due to the large sample size and 0.89 ≤ WMNSQ ≤ 1.11 these items showed acceptable item fit in the test.

Table 4a: Item Parameters

Item	Position	Percentage correct	Difficulty	*SE*	WMNSQ	*t*-value of WMNSQ	r_{it}	Discr.
mag1v051_c	1	57.32	-0.344	0.031	0.97	-2.9	0.50	-0.01
mag1r141_c	2	25.22	1.301	0.034	1.03	1.7	0.38	1.00
mag1g171_c	3	82.32	-1.809	0.037	1.00	0.1	0.38	1.59
mag1d131_c	4	48.01	0.123	0.031	0.97	-2.8	0.50	1.05
mag1d132_c	5	58.62	-0.364	0.033	0.90	-9.2	0.58	1.49
mag1z061_c	6	54.80	-0.226	0.031	0.89	-10.5	0.59	1.72
mag1v01s_c	7	n.a.	-1.054	0.038	0.99	-1.0	0.41	2.51
mag1z20s_c	9	n.a.	-1.540	0.042	0.95	-3.6	0.45	2.33
mag1d09s_c	10	n.a.	-0.390	0.032	1.04	2.4	0.43	0.74
mag1z121_c	11	11.27	2.422	0.045	1.05	1.7	0.21	1.05
mag1g181_c	12	51.85	-0.089	0.031	1.09	8.4	0.37	0.59
mag1d081_c	13	62.49	-0.611	0.031	0.92	-6.7	0.54	0.70
mag1r151_c	14	59.09	-0.432	0.031	0.96	-3.9	0.51	0.88
mag1z111_c	15	71.11	-1.059	0.034	0.95	-3.7	0.50	2.04
mag1v021_c	16	38.08	0.589	0.032	1.05	4.4	0.38	1.60
mag1z071_c	17	35.07	0.7212	0.032	1.05	4.4	0.38	1.96
mag1d041_c	18	70.93	-1.069	0.033	1.11	7.2	0.32	0.98
mag1g031_c	19	50.06	-0.006	0.031	0.92	-8.1	0.55	0.95
mag1z161_c	20	46.29	0.155	0.031	1.11	10.3	0.34	0.75
mag1v101_c	21	74.16	-1.256	0.034	0.95	-3.4	0.48	1.99
mag1r19s_c	22	n.a.	-0.114	0.034	1.10	8.90	0.33	0.74

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 CMC item scores. These are denoted by n.a. Keep in mind that mag1r19s_c was scored dichotomously due to insufficient frequencies in categories. 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). Item 8 was removed from analyses.

Table 4b: Step Parameters of Polytomous Items

Item	Position in the test	Location parameter	step 1 (*SE*)	step 2 (*SE*)	Step 3
mag1v01s_c	7	-1.054	0.111 (0.030)	-0.111	
mag1z20s_c	9	-1.540	-0.180 (0.030)	0.180	
mag1d09s_c	10	-0.390	-0.553 (0.035)	-0.127 (0.036)	0.680

Note. The last subtask of mag1v01s_c had to be removed from analyses due to bad item fit. Mag1r19s_c was scored dichotomously and therefore cannot be found in table 4b.

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the participants' 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 variance was estimated to be 0.967, indicating that the test differentiated reasonably well between subjects. The reliability of the test (EAP/PV reliability =0.761, WLE reliability = 0.739) was good. Although the items covered a wide range of the ability distribution, the range of item difficulties showed some larger gaps on the upper and some smaller gaps on the lower end of the scale. As a consequence, person abilities in medium regions and mostly in lower regions were measured relatively precisely, whereas high and very low ability estimates had larger standard errors.

Scale in logits	Person ability	Item difficulty
3		
	X	10
	X	
	X	
2	X	
	XX	
	X	
	XX	
	XXX	
	XXXX	
	XXXXXX	2
	XXXXX	
1	XXXXX	
	XXXXXXXX	
	XXXXXXX	16
	XXXXXXXXX	15
	XXXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXXX	19
	XXXXXXXXX	4
0	XXXXXXXXX	11 18 21
	XXXXXXXXX	6
	XXXXXXXX	1 5 9
	XXXXXXX	13
	XXXXXXXXX	12
	XXXXXXXXX	
	XXXXXXXX	
-1	XXXXXXX	
	XXXXX	14 17
	XXXXX	20
	XXX	
	XXXX	8
	XXX	
	XX	3 7
	X	
-2	X	
	X	
-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 34.8 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4a).

4.3 Quality of the test

Since the items of the mathematical competence test referred to many different stimuli, the assumption of local item independence is plausible.

4.3.1 Fit of the subtasks of complex multiple-choice items

Before the responses to the subtasks of the CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks had been checked by analyzing the subtasks together with the simple multiple-choice items via a simple Rasch model. There were 33 variables altogether.

The rates of correct responses given to the subtasks of the CMC items varied from 49.77% to 98.19%. With one exception, the subtasks ranged between acceptable and very good item fit – WMNSQ ranging between 0.93 and 1.16 and the respective *t*-values between -3.20 and 11.50. The only subtask exhibiting unsatisfactory item fit – WMNSQ of 1.25 and a respective *t*-value of 26.3 – was excluded from further analysis. The good model fit of the other subtasks justified their aggregation to polytomous variables for each item (mag1v01s_c, mag1z20s_c and mag1d09s_c). As described in section 3.1, one item (mag1r19s_c) was scored dichotomously.

4.3.2 Distractor analyses

To investigate how well the distractors performed in the test, the point-biserial correlations between selecting each incorrect response (distractor) in MC items and the students´ total correct scores was evaluated. This distractor analysis was performed on the basis of preliminary analyses treating all subtasks of CMC items as single items.

Table 5 shows a summary of point biserial correlations between response and ability for correct and incorrect responses restricted to MC items (only the items where subjects were asked to choose between distractors).

Table 5: Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.34	-0.15
Minimum	0.11	-0.44
Maximum	0.50	0.02

4.3.3 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC and polytomous CMC items. After excluding one item as well as one subtask due to bad item fit, the final set of items fitted quite well. Therefore, altogether, item fit can be considered to be very good (see Table 4a). Values of the WMNSQ were close to 1 with the lowest value being 0.89 (item mag1z061_c) and the highest being 1.11 (item mag1z161_c). The item with the largest WMNSQ showed an acceptable, slightly flat item characteristic curve (ICC), and the item with the smallest WMNSQ showed an acceptable,

slightly steep item characteristic curve. Therefore, all ICC showed good or very good item fit. Overall, there was no indication of severe item over- or underfit in the final set of items. The correlations of the item score with the total score varied between 0.32 (item mag1d041_c) and 0.59 (item mag1z061_c), with one exception (item mag1z121_c) that showed a correlation of 0.21. However, this item showed good item fit with the WMNSQ being 1.05 and a weighted *t*-value of 1.7. An explanation for this small correlation could be that this item was the most difficult item of the test. Only 11.27% of the test takers were able to solve this item correctly. Taking these circumstances as well as the lack of other difficult items into account, the item discriminated very well and was therefore included in the analyses. Overall, the test showed an average correlation of 0.44.

4.3.4 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, the number of books at home, the position of the test, 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. Female versus male, for example, indicates the difference in difficulty *ß*(female) – *ß*(male). A positive value indicates a higher difficulty for females, a negative value a lower difficulty for females compared to males.

Table 6.1: Differential Item Functioning (Absolute Differences Between Difficulties): Gender and Test Position

Item	Gender	Position					
	female vs. male	Day 1: Math/ Science vs. Day 1: Science/ Math	Day 1: Math/ Science vs. Day 2: Math/ Science	Day 1: Math/ Science vs. Day 2: Science/ Math	Day 1: Science/ Math vs. Day 2: Math/ Science	Day 1: Science/ Math vs. Day 2: Science/ Math	Day 2: Math/ Science vs. Day 2: Science/ Math
mag1v051_c	-0.206	-0.006	0.000	0.050	0.006	0.054	0.050
mag1r141_c	0.178	0.250	0.208	0.244	-0.042	-0.008	0.034
mag1g171_c	0.062	-0.056	-0.156	-0.130	-0.102	-0.074	0.026
mag1d131_c	-0.112	0.280	0.160	0.288	-0.118	0.008	0.128
mag1d132_c	-0.032	-0.048	-0.134	0.014	-0.084	0.062	0.146
mag1z061_c	0.088	0.060	0.154	-0.066	0.094	-0.128	-0.220
mag1v01s_c	-0.348	0.254	0.194	0.116	-0.060	-0.152	-0.088
mag1z20s_c	0.344	-0.308	-0.104	-0.380	0.198	-0.074	-0.268
mag1d09s_c	0.060	0.014	0.076	0.062	0.060	0.046	-0.012
mag1z121_c	0.032	0.214	0.244	0.484	0.028	0.270	0.240
mag1g181_c	-0.108	-0.098	-0.162	-0.112	-0.064	-0.016	0.048
mag1d081_c	0.074	-0.188	-0.142	-0.226	0.046	-0.040	-0.084
mag1r151_c	-0.024	0.110	0.182	0.168	0.072	0.058	-0.014

mag1z111_c	0.582	0.190	0.112	0.074	-0.078	-0.118	-0.038
mag1v021_c	-0.056	-0.108	-0.294	-0.192	-0.186	-0.086	0.100
mag1z071_c	0.138	0.002	-0.122	-0.064	-0.126	-0.068	0.058
mag1d041_c	0.068	-0.108	0.014	0.118	0.120	0.226	0.106
mag1g031_c	0.150	-0.216	0.004	-0.058	0.218	0.156	-0.062
mag1z161_c	-0.188	-0.082	-0.096	-0.114	-0.014	-0.032	-0.018
mag1v101_c	-0.128	-0.262	-0.196	-0.442	0.066	-0.180	-0.244
mag1r19s_c	-0.390	0.132	0.118	0.282	-0.014	0.150	0.164
Main effect (model with DIF)	**0.244**	**0.078**	**0.034**	**-0.012**	**-0.044**	**-0.088**	**-0.044**
Main effect (Model without DIF)	**0.244**	**0.078**	**0.034**	**-0.010**	**-0.044**	**-0.086**	**-0.044**

Table 6.2: Differential Item Functioning (Absolute Differences Between Difficulties): Migration Status and Number of Books

Item	**Migration status**			**Books**		
	Without vs. with	Without vs. missing	With vs. missing	<100 books vs. >100 books	<100 books vs. missing	>100 books vs. missing
mag1v051_c	-0.026	-0.040	-0.010	0.270	0.094	-0.186
mag1r141_c	-0.066	0.080	0.148	0.040	0.148	0.100
mag1g171_c	0.006	-0.014	-0.018	-0.068	-0.040	0.020
mag1d131_c	0.050	0.108	0.062	-0.264	-0.054	0.202
mag1d132_c	-0.104	-0.100	0.006	-0.032	-0.124	-0.100
mag1z061_c	-0.070	-0.184	-0.112	0.304	-0.056	-0.368
mag1v01s_c	-0.124	-0.186	-0.060	0.346	0.046	-0.308
mag1z20s_c	0.044	-0.164	-0.206	-0.036	-0.220	-0.176
mag1d09s_c	-0.102	-0.030	0.080	0.116	-0.078	-0.186
mag1z121_c	0.242	0.214	-0.026	-0.164	0.124	0.280
mag1g181_c	0.094	0.236	0.146	-0.248	0.084	0.322
mag1d081_c	-0.224	-0.316	-0.090	0.322	-0.142	-0.472
mag1r151_c	-0.324	-0.112	0.214	0.346	0.202	-0.154

mag1z111_c	0.082	-0.022	-0.102	-0.054	-0.128	-0.082
mag1v021_c	0.202	0.124	-0.074	-0.166	0.098	0.256
mag1z071_c	0.172	0.130	-0.040	-0.348	-0.062	0.278
mag1d041_c	0.136	0.252	0.118	-0.182	0.080	0.254
mag1g031_c	0.032	-0.088	-0.118	0.016	0.054	0.030
mag1z161_c	0.262	0.350	0.090	-0.320	0.180	0.492
mag1v101_c	-0.134	-0.308	-0.172	0.170	-0.184	-0.362
mag1r19s_c	-0.020	0.070	0.092	-0.172	-0.044	0.120
Main effect (model with DIF)	**-0.450**	**-0.496**	**-0.042**	**0.520**	**-0.202**	**-0.714**
Main effect (Model without DIF)	**-0.454**	**-0.494**	**-0.042**	**0.522**	**-0.204**	**-0.728**

Overall, 3,317 (51.1%) of the test takers were female and 3,171 (48.9%) were male. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.248 logits, Cohen's d = 0.254). There was no item with a considerable gender DIF. The only item for which the difference in item difficulties between the two groups exceeded 0.4 logits was item mag1z111_c (0.578 logits). However, this item showed good fit in the other categories and belongs to the category of sets, numbers and operations items with a focus on large numbers above the first graders' comfort zone. Therefore, this difference was not considered as severe.

The test takers received either the mathematics or the science test first and were also tested either on the first or the second testing day. A second DIF analysis was performed in order to determine whether there was a resulting position effect between the four groups. There were 1,583 (24.4%) subjects who took the mathematics test on the first testing day in first position, and 1,652 (25.5%) subjects took it on the first testing day following the science test, and therefore on second position. There were no considerable average differences between the two groups (main effect = 0.076 logits, Cohen's d = 0.076). There was no considerable DIF comparing participants with the different test positions on the first testing day.

On the second testing day, 1,653 (25.5%) children took the mathematics test in first position. There were no considerable average differences between this group and the group of children that took the test in the same position on the first testing day (main effect = 0.032, Cohen´s d = 0.032). Therefore, the test fairness could be confirmed for the two groups of participants taking the mathematics test in first position but on different testing days.

Furthermore, there were 1,598 (24.6%) test takers that took the mathematics test on the second testing day in second position. Comparing this group to the first group of children that took the mathematics test on the first testing day on first position (main effect = 0.014, Cohen´s d = 0.014) showed two items with small but not severe differences. Item mag1z121_c showed a difference between the two groups of 0.49 logits and mag1v101_c showed a difference of 0.44 logits. Therefore, both items slightly exceeded 0.4 logits. However, these

items showed good fit in the other categories and the differences were, thus, not considered severe.

Comparing the test takers that took the mathematics test on the first testing day following the science test to the group of test takers that took the test on the second testing day before the science test again revealed no considerable DIFs (main effect = 0.044, Cohen´s d = 0.045).

There was also no considerable DIF between the groups of test takers that took the mathematics test in second position, either on the first or the second testing day (main effect = 0.088, Cohen´s d = 0.090).

Furthermore, there was also no considerable DIF between the groups of test takers that took the mathematics test on the second day, either in first or in second position (main effect = 0.046, Cohen´s d = 0.048).

Overall, taking into account that all main effects were smaller than 0.1 logits, test fairness could be confirmed for the four different subgroups considering the test position.

There were 3,662 (56.4%) participants without migration background, 1,196 (18.4%) participants with migration background, and 1,630 (25.1%) participants without a valid response. All three groups were used for investigating DIF of migration. On average, participants without migration background performed considerably better in the mathematics test than those with migration background (main effect = 0.452 logits, Cohen's d = 0.481). Furthermore, subjects with missing values for migration differ from those without migration background (main effect = 0.496 logits, Cohen's d = 0.519). Here, too, participants without migration background showed a higher mathematical competence. Subjects with migration background performed slightly better compared to participants with missing values for migration (main effect = -0.042 logits, Cohen's d = -0.043). There was no considerable DIF comparing the three groups.

The number of books at home was used as a proxy for socioeconomic status. There were 1,920 (29.6%) test takers with 0 to 100 books at home, 3,517 (54.2%) test takers with more than 100 books at home, and 1051 (16.2%) test takers without any valid response. Group differences and DIF were investigated by using these three groups. There were considerable average differences between the three groups. Participants with 100 or fewer books at home performed on average 0.522 logits (Cohen's d = 0.564) lower in the mathematics test than participants with more than 100 books. Participants without a valid response in relation to the variable books at home performed 0.204 logits (Cohen's d = 0.213) or 0.726 logits (Cohen's d = 0.779) worse than participants with up to 100 and more than 100 books, respectively. There is no considerable DIF comparing the three groups. Differences in item difficulties exceeding 0.4 logits were observed in two items (mag1d081_c and mag1z161_c). However, the highest difference was 0.490 logits. The items showed good item fit in the other categories, therefore, the differences were small and not severe.

In Table 7, the models only including main effects are compared with those that additionally estimate DIF. Ignoring the subgroups with missing information, Akaike's (1974) information criterion (AIC) favored the models estimating DIF for the three DIF variables gender, books and migration background. Taking the subgroup of missing information into account, AIC favored the models estimating only the main effect for the variables less than 100 books versus missing, as well as with migration background versus missing. For the position variable AIC favored the models with DIF for all group comparisons but position 2 vs. 3 and 2 vs. 4 (for

group explanation see note in table 7). 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 for all groups of migration status and test position, respectively, were preferred over the more complex DIF models. However, for the variables gender and books, BIC favored the models estimating DIF. (Note that the analyses including gender contain fewer cases and, thus, the information criteria cannot be compared across analyses with different DIF variables.)

Table 7: Comparison of Models With and Without DIF

DIF variable		Model	Deviance	Number of parameters	AIC	BIC
Gender	Female vs. male	main effect	164,517.30	27	164,571.30	164,754.30
		DIF	164,283.37	48	164,379.37	164,704.70
Migration status	Without vs. with	main effect	122,532.26	27	122,586.30	122,761.44
		DIF	122,463.34	48	122,559.30	122,870.78
	Without vs. missing	main effect	133,899.24	27	133,953.20	134,130.73
		DIF	133,771.83	48	133,867.80	134,183.38
	With vs. missing	main effect	72,051.34	27	72,105.34	72,265.90
		DIF	72,020.66	48	72,116.66	72,402.10
Position	1-2	main effect	80,934.13	27	80,988.13	82,252.34
		DIF	80,860.87	48	80,956.87	81,248.79
	1-3	main effect	81,308.56	27	81,362.56	81,526.78
		DIF	81,244.95	48	81,340.95	81,632.89
	1-4	main effect	80,557.28	27	80,611.28	80,775.03
		DIF	80,449.19	48	80,545.19	80,836.31
	2-3	main effect	83,917.79	27	83,971.79	84,136.58
		DIF	83,887.54	48	83,983.51	84,276.49
	2-4	main effect	83,133.80	27	83,187.80	83,352.14
		DIF	83,096.77	48	83,192.77	83,484.92
	3-4	main effect	83,516.91	27	83,570.91	83,735.25
		DIF	83,472.89	48	83,568.89	83,861.05
Books	<100 books vs. >100 books	main effect	137,195.23	27	137,249.23	155,114.69
		DIF	136,965.82	48	137,061.82	137,378.67
	<100 books vs. missing	main effect	76,009.87	27	76,063.87	85,801.80
		DIF	75,976.11	48	76,072.11	76,359.95
	>100 books vs. missing	main effect	114,764.74	27	114,818.74	129,820.12
		DIF	114,567.30	48	114,663.30	114,971.79

Note. Position 1-2 means Day 1 Math/ Science vs. Day 1 Science/ Math; position 1-3 means Day 1 Math/ Science vs. Day 2 Math/ Science; position 1-4 means Day 1 Math/ Science vs. Day 2 Science/ Math; position 2-3 means Day 1 Science/ Math vs. Day 2 Math/ Science; position 2-4 means Day 1 Science/ Math vs. Day 2 Science/ Math; position 3-4 means Day 2 Math/ Science vs. Day 2 Science/ Math.

4.3.5 Rasch-homogeneity

In order to test the assumption of Rasch-homogeneity, we also fitted a generalized partial credit model (2PL) to the data. The estimated discrimination parameters are given in Table 4a. They range from 0.48 (item mag1d09s_c) to 1.99 (items mag1g171_c).

The 2PL model (AIC = 168,856.69214, BIC = 169,334.68723, number of parameters = 72) fitted the data better than the partial credit model (1PL) (AIC = 173,169.62655, BIC = 173,447.51265, number of parameters = 41). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework (see Pohl & Carstensen, 2012, 2013 for a discussion of this issue), and thus, the partial credit model was used to model the data and to estimate competence scores.

4.3.6 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 method implemented in TAM in R version 3.4.1 (R Core Team, 2017) was used. Due to convergence problems even with 25 nodes per dimension, model parameters could not be estimated in ConQuest using the Gauss-Hermite quadrature method. This might be caused by very high correlations between the five dimensions. The number of nodes used in TAM was set to 10,000.

The variances and correlations of the five dimensions are shown in Table 8. All five dimensions exhibit a substantial variance. The correlations between the five dimensions vary between 0.632 and 0.953, while the lowest correlation occurs between dimensions 2 and 3. This could be explained by the test design since there are only 3 items in both categories, while the other categories contain 4 to 6 items. Model fit between the unidimensional model and the five-dimensional model is compared in Table 9.

Table 8: Results of Five-Dimensional Scaling

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Change and relationships (4 items)	1.077				
Space and shape (3 items)	0.830	1.232			
Units and measuring (3 items)	0.887	0.632	1.208		
Data and chance (5 items)	0.911	0.793	0.777	1.348	
Sets, numbers and operations (6 items)	0.912	0.677	0.953	0.822	1.103

Note. Variances of the dimensions are depicted in the diagonal; 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	165,975.4	27	166,029.4	166,212.40
Five-dimensional	165,641.5	41	165,723.5	166,001.39

Note. Contrary to the calculations for the 1PL and 2PL models, results in this table were achieved by using TAM in R 3.4.1 (Quasi Monte Carlo estimation).

The comparison shows that using either AIC or BIC, the five-dimensional model describes the data better than the unidimensional model. However, the rather high correlations implicate that a substantial common construct of mathematics competence is measured. Therefore, the unidimensional model still seems reasonable.

5 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 2 and at describing how the mathematics competence score had been estimated.

The number of different kinds of missing responses was evaluated and the number of all kinds of missing responses was rather low. Furthermore, item as well as test quality was examined. As indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, ICC – the items exhibited good item fit. Moreover, discrimination values of the items (either estimated in a 2PL model or as correlations of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test was fair for the examined subgroups.

Overall, there was a negligible number of invalid and not reached items, as well as a reasonable number of omitted items.

The test had good reliability (EAP/PV-reliability = .761, WLE reliability = .739) and, taking into account the strong heterogeneity of the test group and the lack of some very difficult items into account, the test distinguished reasonably well between test takers, as indicated by the test's variance (=0.967). However, the item distribution along the ability scale is acceptable, that is, the test distinguished relatively precisely for lower and well for medium abilities, but showed a lack of difficult items.

Fitting a five-dimensional partial credit model (between-item multidimensionality, the dimensions being the content areas) yielded a better model-fit than the unidimensional partial credit model. However, high correlations between the five dimensions of 0.819 on average indicated that the unidimensional model described the data reasonably well.

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

6.1 Naming conventions

The data in the SUF contain 21 items, 18 of which were scored as dichotomous variables (17 MC items, as well as one CMC item due to the collapse of categories) with 0 indicating an

incorrect response and 1 indicating a correct response. Four items were scored as polytomous variables (CMC items) indicating the number of correctly answered subtasks. MC variables are marked with '_c' at the end of variable names; CMC variables end in 's_c'. In the scaling model, polytomous variables are scored in steps of 0.5 – 0 for the lowest category and 1.5 for the highest.

6.2 Linking the data of Kindergarten and 1st Grade

In starting cohort 2, the mathematics competence tests administered in kindergarten (see Schnittjer, 2018) and first grade consist of different items that were constructed to allow an accurate measurement of mathematical competence within each age group. As a consequence, the competence scores derived in the different grades cannot be compared directly. Differences in observed scores would reflect differences in competencies as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the comparison of competencies across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). The process of linking puts adjacent measurement points on the same scale. As such, the scale of the first measurement of each competence within a cohort is used as a reference scale that all subsequent measurements will refer to. The linking of mathematical competence between the kindergarteners and the first graders is achieved using an anchor-group design because, as described above, there were no common items for those two tests. Therefore, common information on both tests was created by using a wave-independent link sample. This independent link sample was drawn from first graders, that is, the same age group as starting cohort two at the current measurement point. The independent test takers received both tests at a single measurement occasion.

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 present test. For more information on the selection of link samples and the method for linking the tests of mathematical competence in starting cohort 2 (Kindergarten and first grade) see Fischer et al. (2016).

6.2.1 Samples

A total of 528 children of the main study participated at both measurement occasions (kindergarten and first grade). These respondents and the independent link sample of $N = 438$ children (49.3% girls) were used to link the two tests across both grades (see Fischer et al., 2016).

6.2.2 The design of the link study

The test administered in kindergarten consisted of 26 items (see Schnittjer, 2018), whereas the test administered in grade 1 consisted of 21 items that were used for the final analyses (see above). Furthermore, the grade 1 test was administered either at the first or second position of the test battery while the kindergarten test was administered only at the first position. The test positions in the link sample were also rotated randomly. The items themselves were administered in the same order for all participants.

6.2.3 Correcting and change in study design

As noted above, the test rotation was changed between studies. Thus, all WLEs of the current study had to be corrected if the respective child received the test at second position. To

achieve this correction, half of the estimated position effect of the longitudinal subsample was added to the link constant (see below). Additionally, the full position effect (0.014 logits) was subtracted from the WLE of those participants who received the mathematics test at second position. This was necessary as participants who worked on the mathematics test first were worse by 0.014 logits. Thus, these corrected WLEs reflect comparable competence scores for an artificial test design without any rotations in test positions.

6.2.4 Results

To examine whether the two tests administered in the link sample measured a common scale, a one-dimensional model that specified a single latent factor for all items was compared to a two-dimensional model that specified separate latent factors for the two tests. The information criteria are inconclusive. Akaike's information criterion (AIC = 18,981.0) favored the two-dimensional model over the one-dimensional model (AIC = 18,983.9), whereas the more conservative Bayesian information criterion (BIC) favored the one-dimensional model (BIC = 19,170.3) over the two-dimensional model (BIC = 19,175.5). However, an additional examination of the residual correlations for the one-dimensional model using the corrected Q3 statistic (Yen, 1984) indicated a largely unidimensional scale—the average absolute residual correlation was *M* = -.02 (*SD* = .06, *Max* = .37, *Min* = -.20). While the range of Q3 also indicates slight problems with the assumption of undimensionality, the proportion of residual correlations per item that exceeded an absolute value of .20 was never higher than 5% (items mak2g211_c, mak2z141_c and mag1z121_c). We therefore concluded that the mathematics competence tests administered in kindergarten and first grade were essentially unidimensional.

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 to the longitudinal subsample of the starting cohort. The differences in item difficulties between the link sample and starting cohort 2 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 10.

Table 10: Differential Item Functioning Analysis between the Starting Cohort and the Link Sample.

	Kindergarten				Grade 1			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1	mak2z221_c				mag1v051_c	-0,193	0,164	1,38
2	mak2z231_c	0,422	0,192	4,82	mag1r141_c	-0,256	0,177	2,08
3	mak2z101_c	0,464	0,199	5,44	mag1g171_c	0,114	0,193	0,35
4	mak2r111_c	-0,334	0,163	4,21	mag1d131_c	-0,592	0,171	12,00
5	mak2g041_c	0,643	0,172	14,01	mag1d132_c	-0,146	0,174	0,71
6	mak2g051_c	1,776	0,177	100,93	mag1z061_c	-0,668	0,165	16,36

	Kindergarten				Grade 1			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
7	mak2v001_c	-0,032	0,174	0,03	mag1v01s_c	-0,269	0,202	1,78
8	mak2r151_c	-0,496	0,164	9,17	mag1z20s_c	-0,325	0,209	2,42
9	mak2z031_c	0,736	0,176	17,58	mag1d09s_c	-0,128	0,172	0,56
10	mak2d062_c	-0,366	0,161	5,18	mag1z121_c			
11	mak2z161_c	-0,557	0,163	11,68	mag1g181_c	-0,234	0,162	2,09
12	mak2z171_c	0,046	0,204	0,05	mag1d081_c	-0,135	0,162	0,70
13	mak2g211_c	-0,244	0,160	2,32	mag1r151_c	-0,574	0,162	12,56
14	mak2r131_c	0,010	0,167	0,00	mag1z111_c	-0,141	0,173	0,67
15	mak2z091_c	-0,501	0,158	10,03	mag1v021_c			
16	mak2v081_c	-0,136	0,167	0,66	mag1z071_c	-0,328	0,177	3,43
17	mak2z201_c	1,118	0,205	29,87	mag1d041_c	-0,021	0,171	0,02
18	mak2d011_c	-0,801	0,174	21,24	mag1g031_c	-0,118	0,165	0,51
19	mak2z241_c	1,085	0,171	40,47	mag1z161_c			
20	mak2z121_c	0,881	0,213	17,11	mag1v101_c	0,061	0,187	0,11
21	mak2v071_c	-0,294	0,174	2,86	mag1r19s_c			
22	mak2g021_c	0,794	0,284	7,82				
23	mak2z251_c	0,108	0,161	0,45				
24	mak2r191_c							
25	mak2v181_c	-0,309	0,167	3,42				
26	mak2z141_c	0,051	0,187	0,07				

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in kindergarten or grade 1 and the link sample (positive values indicate easier items in the link sample); *SE*$_{Δσ}$ = 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 F_{0154} (2, 953) = 30.43. A non-significant test indicates measurement invariance.

The items mak2z221_c, mak2g051_c, mak2z241_c, and mak2r191_c from the kindergarten test and the items mag1z121_c, mag1v021_c, mag1z161_c, and mag1r19s_c from the grade 1 test were excluded from the computation of the link constant because of ceiling effects (only 5 children did not solve mak2z221_c correctly) or because they either failed to be measurement invariant over time (mak2g051_c, mak2z241_c) or exhibited DIF. All other items of the mathematics tests were, therefore, used for the following calculation using the mean/mean method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as c = 1.3452 + 0.007 = 1.3522. This correction term was subsequently added to each difficulty parameter estimated in grade 1 (see Table 4a) to derive the linked item parameters (see Appendix D). The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.1248 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.

6.3 Mathematical competence scores

In the SUF, manifest mathematical competence scale scores are provided in the form of two different WLEs, mag1_sc1 and mag1_sc1u, including their respective standard errors, mag1_sc2 and mag1_sc2u. Both WLE scores are linked to the underlying reference scale of kindergarten. If the focus of research lies on longitudinal issues such as competence development, mag1_sc1u should be used as it is corrected so that, even though the test position has been changed between tests, the WLE does not reflect that change, but can be used as though no design changes had taken place. Therefore, resulting differences in WLE scores can be interpreted as competence development across measurement points. Consequently, mag1_sc1 that corrected for the position of the math test within the booklet can be used if the research interest is based on cross-sectional issues. The ConQuest syntax for estimating the WLE scores is provided in Appendix A, the cross-sectional item parameters used for the above psychometric test analyses are provided in Appendix B, and the linked item parameters used for WLE estimation for the SUF in Appendix D. 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.

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*.*

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.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: 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. (2016). TAM: Test Analysis Modules. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 1.995-0).

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).

Schnittjer, I. (2018). *NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 2 in Kindergarten* (NEPS Survey Paper No. 42). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Appendix

Appendix A: ConQuest-Syntax for Estimating Cross-sectional WLE Estimates in Starting Cohort II - First Grade

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

data filename.dat;
format pid 1-10 responses 14-34;
labels << labels.nam;

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

recode (0,1,2,3,4) (0,0,0,1,2)!items(8);  /* the former item 8 was
taken out - the consecutive item numbering was adjusted */

recode (0,1,2,3,4) (0,0,1,2,3)!items(9);
recode (0,1,2,3,4,5) (0,0,0,0,0,1)!items(21);

score (0,1) (0,1)!items(1-6,10-21);
score (0,1,2,3) (0,0.5,1,1.5)!items(7,9);
score (0,1,2) (0,0.5,1) !items(8);

model item + item*step – rotation;
set constraint=cases;
import anchor_parameters << linked_prm.prm;

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

Appendix B: Cross-sectional Item Parameters for Grade 1

```
 1      -0.344    /* item mag1v051_c */
 2       1.300    /* item mag1r141_c */
 3      -1.808    /* item mag1g171_c */
 4       0.122    /* item mag1d131_c */
 5      -0.364    /* item mag1d132_c */
 6      -0.226    /* item mag1z061_c */
 7      -1.768    /* item mag1v01s_c */
 8      -1.539    /* item mag1z20s_c *//* the former item 8 was removed
                  and the consecutive item numbering was adjusted */
 9      -0.390    /* item mag1d09s_c*/
10       2.420    /* item mag1z121_c */
11      -0.390    /* item mag1g181_c */
12      -0.611    /* item mag1d081_c */
13      -0.432    /* item mag1r151_c */
14      -1.058    /* item mag1z111_c */
15       0.588    /* item mag1v021_c */
16       0.721    /* item mag1z071_c */
17      -1.068    /* item mag1d041_c */
18      -0.006    /* item mag1g031_c */
19       0.155    /* item mag1z161_c */
20      -1.256    /* item mag1v101_c */
21      -0.114    /* item mag1r19s_c */
```

Appendix C: Content Areas of Items in the Mathematics Test for Grade 1

Position	Item	Content area
1	mag1v051_c	Change and relationships
2	mag1r141_c	Space and shape
3	mag1g171_c	Units and measuring
4	mag1d131_c	Data and chance
5	mag1d132_c	Data and chance
6	mag1z061_c	Sets, numbers, and operations
7	mag1v01s_c	Change and relationships
8	mag1z20s_c	Sets, numbers, and operations
9	mag1d09s_c	Data and chance
10	mag1z121_c	Sets, numbers, and operations
11	mag1g181_c	Units and measuring
12	mag1d081_c	Data and chance
13	mag1r151_c	Space and shape
14	mag1z111_c	Sets, numbers, and operations
15	mag1v021_c	Change and relationships
16	mag1z071_c	Sets, numbers, and operations
17	mag1d041_c	Data and chance
18	mag1g031_c	Units and measuring
19	mag1z161_c	Sets, numbers, and operations
20	mag1v101_c	Change and relationships
21	mag1r19s_c	Space and shape

Note. Up to now, the internal validity of the individual dimensions of mathematical competence as dependent measures has not yet been confirmed (van den Ham, 2016).

Appendix D: Import-file of the anchor parameters for linking the grade 1 test to kindergarten scale

```
1    1.00751123 /* item mag1v051_c */
2    2.65332123 /* item mag1r141_c */
3   -0.45746877 /* item mag1g171_c */
4    1.47463123 /* item mag1d131_c */
5    0.98823123 /* item mag1d132_c */
6    1.12572123 /* item mag1z061_c */
7    0.29804123 /* item mag1v01s_c */
8   -0.18821877 /* item mag1z20s_c */
9    0.96204123 /* item mag1d09s_c */
10   3.77405123 /* item mag1z121_c */
11   1.26320123 /* item mag1g181_c */
12   0.74074123 /* item mag1d081_c */
13   0.91944123 /* item mag1r151_c */
14   0.29311123 /* item mag1z111_c */
15   1.94120123 /* item mag1v021_c */
16   2.07441123 /* item mag1z071_c */
17   0.28300123 /* item mag1d041_c */
18   1.34586123 /* item mag1g031_c */
19   1.50739123 /* item mag1z161_c */
20   0.09519123 /* item mag1v101_c */
21   1.23826123 /* item mag1r19s_c */
22   0.11101    /* item mag1v01s_c step 1 */
23  -0.18040    /* item mag1z20s_c step 1 */
24  -0.55269    /* item mag1d09s_c step 1 */
25  -0.12726    /* item mag1d09s_c step 2 */
```