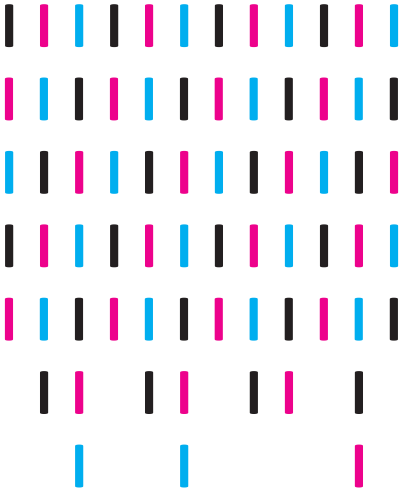

NEPS *SURVEY PAPERS*

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

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS FOR THE ADDITIONAL STUDY THURINGIA

NEPS *Survey Paper* No. 84
Bamberg, March 2021

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

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

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

NEPS Technical Report for Mathematics:

Scaling Results for the Additional Study Thuringia

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

Email address of the lead author:

alkock@leibniz-ipn.de

Bibliographic Data:

Kock, A.-L., Petersen, L. A., & Litteck, K. (2021). *NEPS Technical Report for Mathematics: Scaling Results for the Additional Study Thuringia* (NEPS *Survey Paper* No. 84). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP84:1.0

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

NEPS Technical Report for Mathematics: Scaling Results for the Additional Study Thuringia

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the whole life span and develops tests for assessing these competence domains in different age groups. To evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedure for the mathematics competence test administered in the additional study Thuringia. The test was designed to test the graduating classes of 2010 (the last year which was not affected by the reform of the "Leistungskurs-Grundkurs-System") and 2011 (the first year after the reform). In sum, 2,266 students participated in these two waves. The mathematics test consisted of 41 items (distributed among eight booklets), representing different content areas as well as different cognitive components. A Rasch model was used to scale the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test´s dimensionality were evaluated to ensure the quality of the test. These analyses showed that the test exhibited a good reliability and that the items showed a satisfactory model fit. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were some recognizable gaps at the upper end of the scale's item difficulties. Overall, the mathematics test had good psychometric properties that allowed for an estimation of a reliable mathematics competence score. Besides the scaling results, this paper also describes the data available in the Scientific Use File and provides the ConQuest-Syntax for scaling the data.

Keywords

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

Content

1. Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, and information and communication technologies (ICT) literacy. An overview of the competencies measured in NEPS is given by Weinert et al. (2011) as well as Fuß et al. (2019).

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

In this paper, the results of these analyses are presented for mathematical competence in the two waves of the additional study Thuringia. This study on the reform of the "Leistungskurs-Grundkurs-System" was conducted for the graduating classes of 2010 (last year which was not affected by the reform) and 2011 (first year after the reform). More detailed information about the aims of this study can be found on the NEPS website[1].

2. Testing Mathematical Competence

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

The items are not arranged in units. Thus, in the test, students usually faced a certain situation followed by one to three tasks related to it. Each of the items belonged to one of the following content areas:

- quantity,
- space and shape,
- change and relationships, or
- data and chance.

Each item was constructed in such a way to primarily address a specific content area (see Appendix A). The framework also describes, as a second and independent dimension, six cognitive components required for solving the tasks. These components were distributed across the items.

[1] https://www.neps-data.de/Data-Center/Data-and-Documentation/Additional-Study-Thuringia

3. Data

3.1 The Design of the Study

The study was conducted in 2010 (wave 1) and 2011 (wave 2) and assessed different competence domains including, among others, English reading competence, biological competence, physics competence, and mathematical competence. The mathematics test was administered as the second or sixth test in both waves (see Table 1).

Table 1

Design of the study

Position	Competence domain
1	Biology or Physics
2	Mathematics or English
	15 minute break
3	Cognitive ability test
4	Student Questionnaire
	20 minute break
5	Physics or Biology
6	English or Mathematics

Overall, 41 items were used in the mathematics test. All students received one out of eight different booklets (paper-pencil test). Each booklet included 20 or 21 items, which represented different content-related and process-related components and used different response formats. The characteristics of all 41 items are depicted in the following tables. Table 2 shows the distribution of the four content areas (see Appendix A for the assignment of the items to the content areas), whereas Table 3 shows the distribution of the response formats. The mathematics test included three types of response formats: simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR). In MC items, the test taker had to find the correct response option from four to six available response options. In CMC items, several subtasks with two response options were presented. SCR items required the test taker to write down an answer into an empty box.

Six items were excluded from the analyses due to severe misfit (e.g., low item-total correlations), resulting in a test of 35 items.

Table 2

Number of items by content areas

Content area	**Frequency**
Quantity	13 *(10)*
Space and shape	8 *(7)*
Change and relationships	11 *(10)*
Data and Chance	9 *(8)*
Total number of items	41 *(35)*

Note. The numbers shown in italics represent the frequency after the exclusion of the six misfitting items.

Table 3

Number of items by response formats

Content area	**Frequency**
Simple Multiple-Choice	35 *(29)*
Complex Multiple-Choice	1 *(1)*
Short-constructed response	5 *(5)*
Total number of items	41 *(35)*

Note. The numbers shown in italics represent the frequency after the exclusion of the six misfitting items.

3.2 Sample

Overall, 2,266 persons took the mathematics test: 1,368 in 2010 (Wave 1) and 896 in 2011 (Wave 2). For two of them, less than three valid responses were available. Because no reliable ability scores can be estimated based on such few valid 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 2,264 test-takers. A detailed description of the study design, the sample, and the administered instrument can be found on the NEPS website (https://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 in the booklet, and finally, e) multiple kinds of missing responses within CMC items that are not determined.

Invalid responses occurred, for example, when two response options were selected in simple MC or CMC items where only one was required. Omitted items occurred when test-takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response were coded as not-reached. Because of the eight different booklets, not all items were administered to every participant.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined to evaluate how well each of the items functioned.

3.4 Scaling Model

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

The CMC item consisted of a set of subtasks that were aggregated to a polytomous variable, indicating the number of correctly responded subtasks within that item. Due to unsatisfactory step parameters (the difficulty decreased with an increasing number of points), the CMC item was scored dichotomously (all four subtasks with correct response = 1, three or fewer correct responses = 0). Simple MC and SCR items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). Therefore, all 35 items were scored dichotomously.

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 (for a Conquest syntax for scoring the CMC item, fitting the scaling model, and estimating WLEs, see Appendix B).

3.5 Checking the Quality of the Test

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

The MC items consisted of one correct response option and three to five distractors (i.e., incorrect response options). The quality of the distractors within MC items, that is whether they were chosen by students with a lower ability rather than by those with a higher ability, was evaluated using the point-biserial correlation between selecting an incorrect response option and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 indicate problematic distractors (Pohl & Carstensen, 2012).

The fit of the items to the Rasch model (Rasch, 1960) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (|*t*-value| > 6) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (|*t*-value| > 8) were judged as having a considerable item misfit and their performance was further investigated.

Correlations of the item score with the total score greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall, the 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 these 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), migration background (see Pohl & Carstensen, 2012, for a description of these variables), test position, and wave. Differential item functioning was estimated using a multi-group IRT model to test for measurement invariance, 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 that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

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

The dimensionality of the mathematics test was evaluated by specifying a four-dimensional model based on the four content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, TAM in R was used (Kiefer et al., 2017). 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 (Wu et al., 1997). The generalized partial credit model and the multi-dimensional model were estimated in R version 4.0.2 (R Core Team, 2020) using the TAM package version 3.5-19 (Robitzsch et al., 2020).

4. Results

4.1 Missing Responses

In this section, the scaling results of the two waves of the additional study Thuringia will be presented.

4.1.1 Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was small. In fact, 95.89 % of the test takers did not have any invalid response.

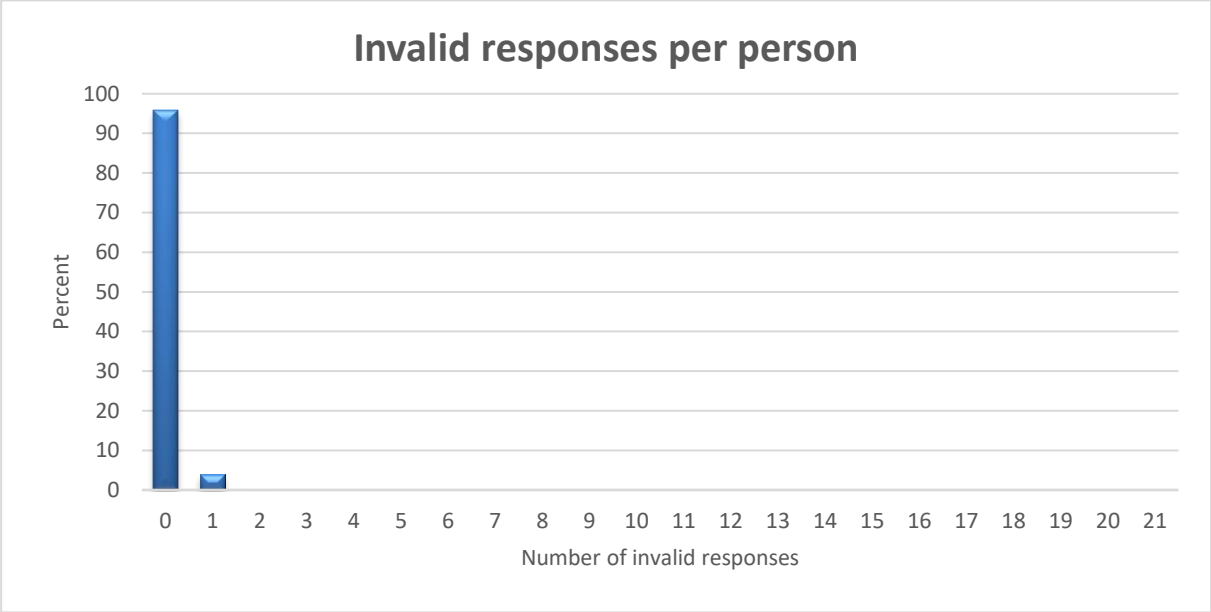

Figure 1. Number of invalid responses

Missing responses may also occur when persons skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. It shows that 72.84 % of the subjects omitted no item. Only 3.45 % of the subjects omitted more than 3 items.

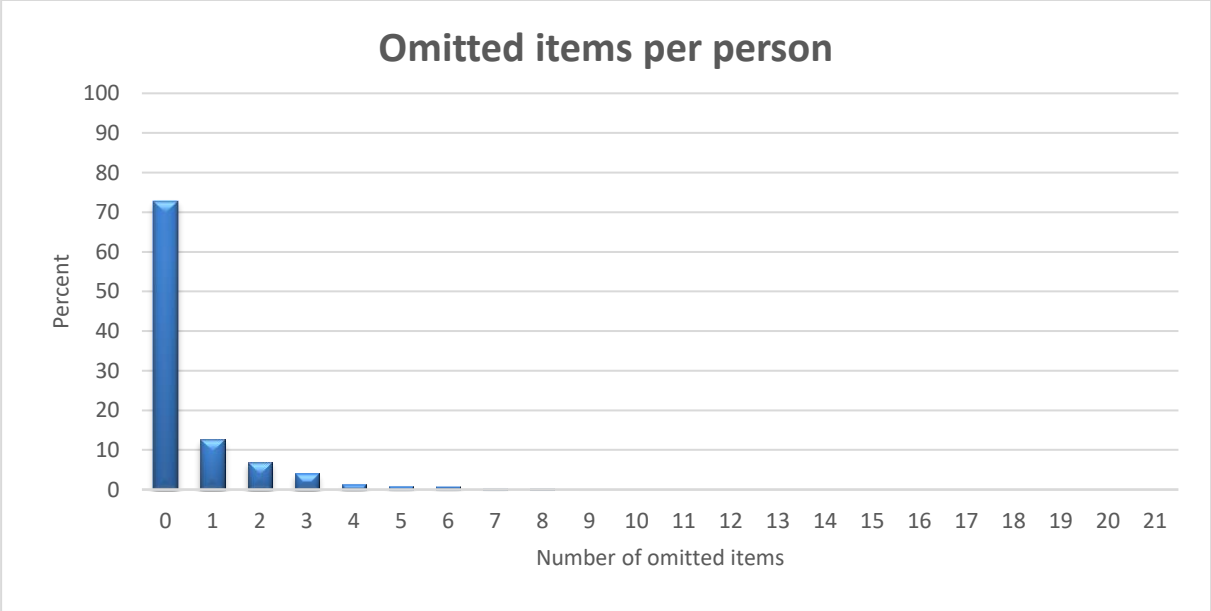

Figure 2. Number of omitted items

All missing responses after the last valid response were defined as not reached. Figure 3 shows the number of items that were not reached by a person. As can be seen, most participants (83.92 %) reached the end of the test within the allocated time limit. 15.15 % of the test takers

did not reach one to five items and only 0.93 % of the test takers did not reach more than 5 items.

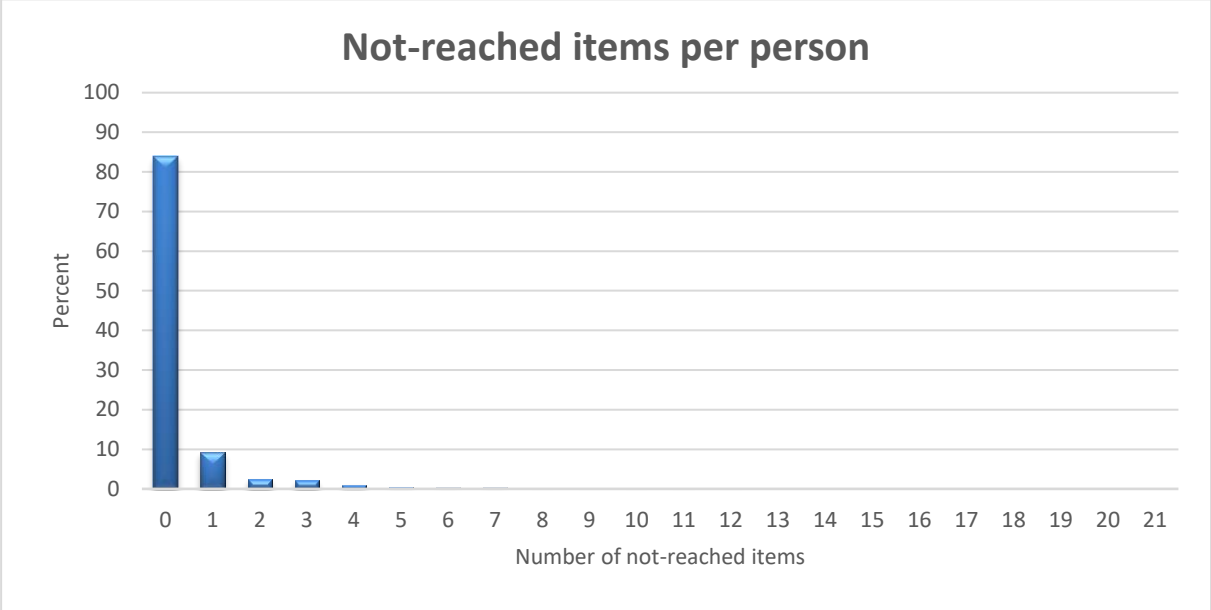

Figure 3. Number of not-reached items

Figure 4 shows the total number of total missing responses per person, which is the sum of invalid, omitted, and not-reached missing responses. In total, 60.73 % of the subjects showed no missing response. 3.30 % showed more than five missing responses.

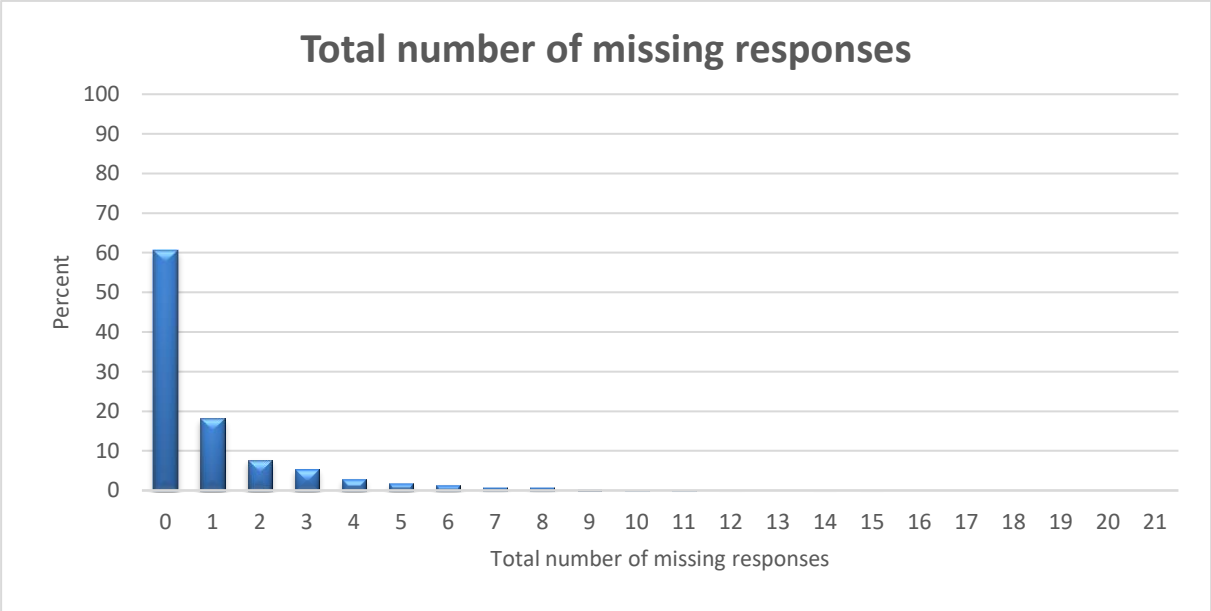

Figure 4. Total number of missing responses

4.1.2 Missing responses per item

Table 4 shows the number of valid responses for each item, as well as the percentage of missing responses.

Table 4

Percentage of Missing Values

Pos.	Item	Number of valid responses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items	Percentage of multiple missings
1	magcr511_c	1,065	0.00	0.18	2.69	0.00
2	magcq581_c	1,084	0.00	0.97	1.06	0.00
3	magcq583_c	1,057	0.04	1.72	1.46	0.00
4	maa2r081_c	1,079	0.00	2.03	0.22	0.00
5	maa2v082_c	1,065	0.00	2.30	0.57	0.00
6	mas2d071_c	1,088	0.00	1.55	0.31	0.00
7	magcq591_c	1,074	0.04	1.10	1.33	0.00
8	mas2q011_c	1,046	0.18	0.93	2.43	0.00
9	mas2d111_c	1,073	0.13	0.62	0.57	0.00
10	maa2d112_c	1,045	0.09	2.61	0.88	0.00
11	maa2d113_c	1,032	0.00	3.18	0.97	0.00
12	magcv501_c	1,087	0.00	1.24	0.49	0.00
13	magcv502_c	1,017	0.18	1.94	0.57	0.00
14	maa2r091_c	1,005	0.13	2.03	0.53	0.00
15	mas2r092_c	1,103	0.00	0.31	0.71	0.00
16	mas2v093_c	1,099	0.00	0.40	0.80	0.00
17	magcr561_c	1,036	1.50	0.18	2.30	0.00
18	maa2r011_c	1,073	0.00	0.44	1.90	0.00
19	mas2v031_c	1,089	0.00	0.84	0.80	0.00
20	mas2v032_c	1,067	0.04	1.59	0.97	0.00
21	maa2d131_c	1,101	0.00	0.71	0.40	0.00
22	maa2d132_c	1,073	0.00	1.77	0.57	0.00
23	mag2q02s_c	1,081	0.04	1.59	0.35	0.00
24	mas2q041_c	1,047	0.00	2.78	0.71	0.00
25	mas2v042_c	736	0.88	5.57	2.16	0.00
26	mag9r061_c	915	0.40	2.16	3.67	0.00
27	maa2q021_c	1,086	0.00	1.06	1.59	0.00
28	magcr532_c	1,124	0.04	0.44	0.49	0.00
29	mas2v061_c	1,049	0.00	4.06	0.22	0.00

30	mas2v062_c	1,017	0.40	5.08	0.22	0.00
31	mas2v063_c	1,099	0.18	1.63	0.27	0.00
32	magcd571_c	1,090	0.04	2.21	0.22	0.00
33	magcr551_c	1,134	0.00	0.27	0.27	0.00
34	magcd541_c	1,120	0.00	0.84	0.31	0.00
35	maa2q071_c	1,123	0.04	0.31	0.66	0.00

Overall, the number of not valid responses per item was very small. The omission rates were small, varying between 0.18 % and 5.57 % (item mas2v042_c). The number of persons that did not reach an item increased with the position of the items in the different booklets up to 3.67 %.

4.2 Parameter Estimates

4.2.1 Item parameters

The relative frequencies of the responses were evaluated before performing any IRT analyses to get a first descriptive measure of the item difficulties and to check for possible estimation problems. Using each subtask of the CMC items as single variables, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 16.86 % and 91.10 % across all items. On average, the rate of correct responses was 58.39 % (*SD* = 20.47 %).

Table 5

Item Parameters

Pos.	Item	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	aQ_3
1	magcr511_c	80.19	-1.65	0.09	0.95	-1.1	0.47	1.43	0.05
2	magcq581_c	90.41	-2.58	0.11	0.99	-0.1	0.30	1.08	0.05
3	magcq583_c	59.04	-0.46	0.07	0.99	-0.5	0.47	1.06	0.08
4	maa2r081_c	65.89	-0.81	0.07	0.99	-0.4	0.46	0.97	0.05
5	maa2v082_c	56.15	-0.33	0.07	1.06	2.4	0.40	0.62	0.05
6	mas2d071_c	44.49	0.22	0.07	1.09	3.8	0.34	0.45	0.06
7	magcq591_c	74.67	-1.30	0.08	1.05	1.4	0.34	0.66	0.05
8	mas2q011_c	68.74	-0.90	0.08	0.98	-0.5	0.46	1.13	0.05
9	mas2d111_c	52.84	-0.11	0.07	1.02	1.0	0.42	0.75	0.06
10	maa2d112_c	31.58	0.93	0.08	1.03	0.9	0.38	0.61	0.06
11	maa2d113_c	37.89	0.60	0.07	1.05	1.7	0.38	0.58	0.06
12	magcv501_c	69.64	-0.95	0.08	0.93	-2.2	0.51	2.29	0.09
13	magcv502_c	55.46	-0.23	0.07	0.92	-3.4	0.54	2.27	0.08
14	maa2r091_c	39.80	0.55	0.07	0.97	-1.1	0.47	1.03	0.04
15	mas2r092_c	29.10	1.07	0.08	1.00	0.1	0.41	0.91	0.04
16	mas2v093_c	72.34	-1.10	0.08	0.97	-0.9	0.45	1.11	0.05
17	magcr561_c	86.00	-2.07	0.10	1.04	0.7	0.27	0.58	0.05
18	maa2r011_c	76.51	-1.44	0.08	0.94	-1.6	0.49	1.54	0.06
19	mas2v031_c	84.57	-2.02	0.09	1.03	0.6	0.29	0.77	0.05
20	mas2v032_c	47.61	0.05	0.07	1.02	0.9	0.44	0.87	0.05
21	maa2d131_c	75.84	-1.39	0.08	1.00	0.0	0.40	1.02	0.04
22	maa2d132_c	68.13	0.09	0.07	0.91	-3.9	0.54	1.55	0.06
23	mag2q02s_c	n.a.	-0.46	0.07	0.97	-1.2	0.50	1.22	0.06
24	mas2q041_c	53.49	-0.23	0.07	1.04	1.7	0.41	0.78	0.04
25	mas2v042_c	33.97	0.79	0.09	0.89	-3.4	0.59	1.62	0.05
26	mag9r061_c	54.43	-0.22	0.08	1.02	0.7	0.45	0.86	0.05
27	maa2q021_c	39.32	0.52	0.07	0.97	-1.1	0.45	0.93	0.04
28	magcr532_c	53.91	-0.17	0.07	1.08	3.6	0.34	0.43	0.06
29	mas2v061_c	40.23	0.47	0.07	1.07	2.9	0.33	0.44	0.08

30	mas2v062_c	26.06	1.21	0.08	0.95	-1.2	0.44	0.97	0.04
31	mas2v063_c	30.39	0.98	0.08	0.98	-0.6	0.43	0.90	0.04
32	magcd571_c	55.78	-0.27	0.07	1.09	4.0	0.33	0.41	0.06
33	magcr551_c	84.22	-1.90	0.09	0.99	-0.1	0.35	0.99	0.05
34	magcd541_c	37.05	0.63	0.07	1.01	0.4	0.41	0.70	0.05
35	maa2q071_c	72.84	-1.13	0.08	1.00	-0.1	0.43	1.11	0.05

Note. Pos. = Item position in the test. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a two-parametric logistic (2PL) model, *aQ3* = adjusted average absolute residual correlation for item (Yen, 1993).
Percent correct scores are not informative for polytomous CMC item scores. These are denoted by n.a.
For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items, it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

From a descriptive point of view, the items covered a wide range of difficulties, although some additional very difficult items would have completed the upper end of the scale. The estimated item difficulties (all items were scored dichotomously) are depicted in Table 5. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.58 (magcq581_c) and 1.21 (mas2v062_c) with a mean of -0.40. Overall, the item difficulties were acceptably well distributed around zero. Due to the large sample size, the standard errors of the estimated item difficulties were small (*SE*(ß) ≤ 0.11).

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 5, the item difficulties of the items and the ability of the respondents are plotted on the same scale. The distribution of the estimated respondents' 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 0.84, indicating that the test differentiated reasonably well between subjects. The reliability of the test (EAP/PV reliability = 0.69, WLE reliability = 0.66) was acceptable.

The items covered a wide range of the ability distribution, although an additional very difficult item would have captured the very high person abilities at the upper end of the scale even better. Therefore, person abilities in medium and in lower ability levels were measured relative precisely, whereas very high ability estimates had larger standard errors.

Scale in logits	Person ability	Item difficulty
	X	
	X	
2	X	
	X	
	XX	
	X	
	XXX	
	XX	
	XXXX	
	XXXXX	30
1	XXXXX	15 31
	XXXXX	10
	XXXXXXX	25
	XXXXXX	34
	XXXXXXXXX	11 14 27
	XXXXXXXXX	29
	XXXXXXXXXX	
	XXXXXXXXX	6
	XXXXXXXXX	20 22
0	XXXXXXXXXX	9
	XXXXXXXXX	13 24 26 28
	XXXXXXXX	5 32
	XXXXXXXX	3 23
	XXXXXXXXXX	
	XXXXXXXX	
	XXXXXXXX	4
	XXXXXX	8 12
-1	XXXXX	
	XXXXX	16 35
	XXX	7
	XXXX	18 21
	XX	
	XX	1
	X	
		33
-2	X	17 19
		2

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 12.4 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 5).

4.3 Quality of the test

4.3.1 Distractor analyses

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

Table 6 shows a summary of the point-biserial correlations between the responses and person's abilities for correct and incorrect responses restricted to MC items (only the items where subjects were asked to choose between distractors). The results indicate that the distractors functioned well.

Table 6

Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.28	-0.12
Minimum	0.17	-0.39
Maximum	0.44	0.08

4.3.2 Item fit

The evaluation of the item fit was performed based on the final scaling model, the Rasch model (see Table 5). Overall, the item fit was good. Values of WMNSQ were close to 1 with the lowest value being 0.89 (item mas2v042_c) and the highest being 1.09 (items mas2d071_c and magcd571_c). No item exhibited a *t*-value of the WMNSQ greater than |6|. Thus, there was no indication of a severe item over- or underfit. All item characteristic curves showed a good fit for the items. The correlation of the item score with the total score varied between .27 (item magcr561_c) and .59 (items mas2v042_c), averaging at .42.

4.3.3 Differential item functioning

We examined test fairness for different groups (i.e., measurement invariance) by estimating the severity of differential item functioning (DIF). Differential item functioning was investigated for the variables gender, migration background, the number of books at home (as a proxy for socioeconomic status), and the position of the mathematics test on test day (see Pohl & Carstensen, 2012, for a description of these variables). Also, the effect of the two waves was analyzed, comparing the two assessment settings in 2010 and 2011. Table 7 shows the difference between the estimated difficulties of the items in different subgroups. For example, the column "Male versus female" indicates the difference in difficulty *ß*(male) – *ß*(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females.

Gender: Overall, 1,040 (45.90 %) of the test takers were male, 1,225 (54.06 %) were female, and one student (0.04 %) without valid gender information. On average, female students exhibited a lower mathematical competence than male students did (main effect = -0.59 logits, Cohen's *d* = 0.55). There was no item with a considerable gender DIF above 0.6 logits. DIF exceeding 0.4 logits occurred for the items magcq581_c, maa2r081_c, mas2q011_c, maa2d113_c, magcv501_c, mas2r092_c, maa2q021_c, magcr532_c, magcd571_c, but were considered not to be severe.

Migration status: There were 1,844 (81.38 %) participants without a migration background, 103 (4.55 %) participants with a migration background, and 319 (14.08 %) participants without a valid response. Due to the very low number of students with a migration background, no reliable DIF could be calculated. For reasons of completeness, the results are nevertheless included in Table 7. Only the first two groups were used for investigating DIF of migration. On average, participants with migration background performed slightly better than those without migration background (main effect = 0.18 logits, Cohen's *d* = 0.20). Two items showed a very strong DIF above 1 logit (magcv501_c, magcv502_c), and one item showed a considerable DIF above 0.6 logits (maa2d112_c). DIF exceeding 0.4 logits occurred for eight items (magcq591_c, mas2d111_c, maa2d113_c, mas2v093_c, mas2v031_c, mas2v042_c, mag9r061_c, magcr532_c). However, as none of these results are reliable due to the small number of persons with a migration background, we did not exclude any items due to the migration DIF.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 325 (14.34 %) test takers with 0 to 100 books at home, 1,557 (68.71 %) test takers with more than 100 books at home, and 384 (16.95 %) test-takers without a valid response. Group differences and DIF were investigated for all three groups (≤100 books vs. >100 books, ≤100 books vs. missings, and >100 books vs. missings). Participants with 100 or fewer books at home performed, on average, 0.28 logits (Cohen's *d* = 0.31) lower in mathematics than participants with more than 100 books. There was no item with a considerable DIF above 0.4 logits. Participants with 100 or fewer books at home also performed 0.36 logits (Cohen's *d* = 0.42) lower than participants without any valid information. Four items exceeded 0.4 logits (mas2d111_c, mas2r092_c, mag9r061_c, mas2v062_c), but were considered not to be severe. Participants with more than 100 books at home performed 0.08 logits (Cohen's *d* = 0.09) lower than persons without any valid information. Three items exceeded 0.4 logits (mas2r092_c, mag9r061_c, magcr532_c), but were also considered not to be severe.

Position: The mathematics test was administered to the students together with other competence tests (see section 3.1 for the design of the study). The order of the different tests was rotated, resulting in two different positions, in which the mathematics test was administered. 1,138 test-takers (50.22 %) received the mathematics test in the second position on the testing day, whereas 1,128 (49.78 %) received the mathematics test in the sixth position. Test takers, who received the mathematics test in the second position, exhibited a higher average mathematics competence (-0.18 logits, Cohen's *d* = 0.20) than test-takers, who received the mathematics test in the sixth position. There were two items slightly exceeding 0.4 logits (magcv501_c, magcd541_c). Both were considered not to be severe.

Table 7

Differential Item Functioning

Nr.	Item	Gender	Migration status	Books			Position	Wave
		male vs. female	without vs. with	≤100 vs. >100	≤100 vs. missing	>100 vs. missing	2 vs. 6	1 vs. 2
1	magcr511_c	-0.17	0.34	-0.39	-0.25	0.13	0.27	0.00
2	magcq581_c	0.48	-0.29	0.01	-0.32	-0.34	0.05	-0.04
3	magcq583_c	0.25	-0.38	0.02	0.04	0.02	-0.17	-0.01
4	maa2r081_c	-0.49	-0.14	0.03	0.09	0.05	0.12	-0.08
5	maa2v082_c	-0.03	0.09	-0.34	-0.03	0.30	0.30	0.11
6	mas2d071_c	0.06	-0.20	0.05	0.03	-0.03	0.11	0.12
7	magcq591_c	0.33	0.59	-0.11	0.16	0.27	0.05	-0.12
8	mas2q011_c	-0.59	0.03	-0.05	-0.16	-0.10	-0.11	0.17
9	mas2d111_c	0.16	-0.43	0.26	0.45	0.20	0.08	-0.12
10	maa2d112_c	0.36	-0.93	0.06	0.10	0.04	0.21	-0.15
11	maa2d113_c	0.54	-0.53	-0.07	-0.10	-0.03	0.08	-0.06
12	magcv501_c	-0.56	1.11	0.30	0.17	-0.12	-0.46	-0.13
13	magcv502_c	-0.37	1.26	-0.03	-0.06	-0.03	-0.25	-0.01
14	maa2r091_c	0.03	0.35	0.06	0.16	0.10	0.02	-0.10
15	mas2r092_c	-0.49	0.00	-0.03	0.42	0.46	-0.09	-0.23
16	mas2v093_c	-0.08	-0.49	0.08	-0.28	-0.35	0.10	0.11
17	magcr561_c	0.18	-0.05	0.16	0.07	-0.08	0.16	0.24

18	maa2r011_c	-0.27	0.27	0.22	0.11	-0.12	-0.18	0.02
19	mas2v031_c	0.37	-0.50	-0.08	-0.21	-0.14	-0.24	0.16
20	mas2v032_c	0.04	-0.29	-0.14	-0.29	-0.15	-0.04	0.04
21	maa2d131_c	-0.22	0.22	-0.06	0.04	0.10	-0.21	-0.08
22	maa2d132_c	-0.04	0.08	-0.23	-0.21	0.02	-0.17	0.08
23	mag2q02s_c	0.00	-0.23	-0.04	-0.28	-0.25	-0.18	0.28
24	mas2q041_c	0.06	-0.13	0.08	0.12	0.03	0.09	0.05
25	mas2v042_c	-0.12	0.55	-0.24	-0.20	0.03	-0.09	-0.19
26	mag9r061_c	0.06	0.57	0.06	0.55	0.49	0.10	0.13
27	maa2q021_c	-0.44	0.30	-0.28	0.03	0.31	0.12	-0.01
28	magcr532_c	0.58	-0.57	0.38	-0.09	-0.47	0.09	0.15
29	mas2v061_c	0.33	-0.36	0.06	0.13	0.07	-0.05	0.01
30	mas2v062_c	-0.06	0.22	0.34	0.51	0.16	-0.38	-0.09
31	mas2v063_c	-0.06	0.14	0.08	-0.06	-0.14	0.01	-0.08
32	magcd571_c	0.45	-0.12	0.18	-0.03	-0.22	0.04	-0.15
33	magcr551_c	-0.03	0.19	-0.37	-0.30	0.07	0.11	0.07
34	magcd541_c	-0.15	-0.12	0.05	-0.33	-0.38	0.43	-0.07
35	maa2q071_c	0.01	-0.26	-0.23	-0.17	0.06	-0.01	0.01
Main effect (DIF model)		**-0.59**	**0.18**	**0.28**	**0.36**	**0.08**	**-0.18**	**0.00**
Main effect (Main effect model)		**-0.59**	**0.21**	**0.29**	**0.37**	**0.08**	**-0.18**	**0.00**

Wave: The mathematical competence test was administered in the first wave in 2010 (the last year that was not affected by the reform of the "Leistungskurs-Grundkurs-System") and again in the second wave in 2011 (the first year after the reform). 1,369 students (60.41 %) participated in 2010, and 897 (39.59 %) students participated in 2011. Both groups performed, on average, comparably (0.00 logits, Cohen's *d* = 0.00). No item exhibited DIF greater than 0.4 logits.

Overall, test fairness could be confirmed for all tested subgroups. In Table 8, we compared the models that only included the main effects to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) only favored the model estimating DIF for the variable gender. For the other DIF variables, the models estimating the main effect were favored. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents an overparameterization of models. Using BIC, the more parsimonious models including only the main effects of all seven variables were preferred over the more complex DIF models.

Table 8
Comparison of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	42,737.69	37	42,811.69	*43,023.53*
	DIF	42,581.74	72	*42,725.74*	43,137.96
Migration status	Main effect	37,045.66	37	*37,119.66*	*37,331.51*
	DIF	36,996.53	72	37,140.53	37,552.78
Books ≤100 vs. >100	Main effect	35,878.89	37	*35,952.89*	*36,157.88*
	DIF	35,851.87	72	35,995.87	36,395.76
Books ≤100 vs. mis.	Main effect	13,398.24	37	*13,472.24*	*13,641.10*
	DIF	13,371.23	72	13,515.23	13,843.83
Books >100 vs. mis.	Main effect	36,501.02	37	*36,575.02*	*36,781.15*
	DIF	36,458.15	72	36,602.15	37,003.26
Position	Main effect	42,917.53	37	*42,991.53*	*43,203.39*
	DIF	42,861.05	72	43,005.05	43,417.31
Wave	Main effect	42,933.23	37	*43,007.23*	*43,219.08*
	DIF	42,911.64	72	43,055.64	43,467.89

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

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. To test for this assumption of Rasch-homogeneity, we also fitted a two-parametric logistic model (2PL; Birnbaum, 1968) to the data. The estimated discrimination parameters are depicted in Table 5 ("Discr."). They ranged between 0.41 (item magcd571_c) and 2.29 (item magcv501_c). The 2PL model (AIC = 42,717.38; BIC = 43,118.12; number of parameters = 70) fitted the data better than the Rasch model (AIC = 43,000.00; BIC = 43,206.09; number of parameters = 36). Nevertheless, the Rasch model more adequately matches the theoretical conceptions underlying the test construction (for a discussion of this issue, see Pohl & Carstensen, 2012; 2013), and, thus, the Rasch model was used to model the data and to estimate competence scores.

4.3.5 Unidimensionality

The unidimensionality of the test was investigated by specifying a four-dimensional model based on the four different content areas. Each item was assigned to one content area (between-item-multidimensionality). Estimation of the models was carried out in R using the Gauss-Hermite quadrature method. The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained (snodes = 15000).

Table 9

Results of Four-Dimensional Scaling

	Quantity	**Space and shape**	**Change and Relationship**	**Data and chance**
Quantity (10 items)	(1.106)			
Space and shape (7 items)	0.906	(0.945)		
Change and relationships (10 items)	0.870	0.775	(1.218)	
Data and chance (8 items)	0.850	0.896	0.805	(0.729)
EAP Reliability	0.675	0.634	0.645	0.625

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

The variances, correlations, and EAP Reliability of the four dimensions are shown in Table 9. Three of the four dimensions exhibited a substantial variance. In dimension four (data and chance), six of the seven items showed difficulties ranging from -0.27 to 0.93, so the difficulties were relatively homogenous in this dimension. This might explain the rather small variance of 0.729 in dimension four.

The correlations between the four dimensions were rather high and varied between .78 and .91. However, all correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95. see Carstensen, 2013). According to the model fit indices, the four-dimensional model fitted the data slightly better than the unidimensional model (see Table

10). These results indicate that the four content areas measure a common construct, although it is not completely unidimensional. Additionally, for the unidimensional model the average absolute residual correlations as indicated by the adjusted Q_3 statistic (see Table 5) were quite low (M = .05, SD = .01) — the largest individual residual correlation was .09 — and, thus, indicated an essentially unidimensional test. Because the mathematics test was constructed to measure a single dimension, a unidimensional mathematics competence score was estimated.

Table 10

Comparison of the Unidimensional and the Four-Dimensional Model

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	42,927.92	36	42,999.92	43,206.02
Four-dimensional	42,855.73	45	*42,945.73*	*43,203.35*

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

5. Discussion

The analyses in the previous sections aimed to provide information on the quality of the mathematics test in the additional study Thuringia and at describing how the mathematics competence scores were estimated.

We investigated different kinds of missing responses and their amount was reasonably small. Furthermore, item as well as test quality were examined. The items exhibited a good item fit as indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, ICC. The item distribution along the ability scale was good, except for some gaps at the upper end of the scale. As a consequence, person abilities in medium and in lower ability levels were measured relatively precisely, whereas very high ability estimates had larger standard errors. Nevertheless, the test had a good reliability and distinguished well between test-takers, as indicated by the test's variance. Moreover, the discrimination values of the items (either estimated in a 2PL or as a correlation of the item score with the total score) were good. Different variables were used for testing measurement invariance. As there were only 103 students with a migration background, no reliable DIF could be calculated for the variable migration status. For the other variables, no item of the test exceeded a considerable DIF of 0.6 logits; indicating test fairness for the considered subgroups. Fitting a four-dimensional Rasch model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model than the unidimensional model. Nevertheless, high correlations between the four dimensions indicate that the unidimensional model described the data reasonably well.

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

6. Data in the Scientific Use File

The data in the Scientific Use File contains 35 items that were all scored dichotomously with 0 indicating an incorrect response and 1 indicating a correct response. The polytomous variable mag2q02s_c was also scored dichotomously for the estimation of the mathematics competence score and scaling model. The dichotomous variables are marked with a '_c' at the end of their variable names; the polytomous variable is marked with a 's_c' behind its variable name. Appendix B provides the syntax that was used to generate person estimates using the ConQuest software (Wu, Adams, & Wilson, 1997).

In the SUF, manifest mathematics competence scores are provided in the form of WLEs ("mas2_sc1"), including their respective standard error ("mas2_sc2"). As described in section 4, these person estimates are from the joint scaling of both waves of the study. For persons who did not give enough valid responses, no WLE was estimated. The value on the WLE and the respective standard error for these persons are shown as non-determinable missing values in the SUF.

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

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-722. http://doi.org/10.1007/978-1-4612-1694-0_16

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.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (eds.), *Research Outcomes of the PISA Research Conference 2009* (pp. 199-214). New York, NY: Springer. http://doi.org/10.1007/978-94-007-4458-5_12

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

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

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

Kiefer, T., Robitzsch, A., & Wu, M. (2017). *TAM: Test Analysis Modules*. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 3.5-19).

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

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

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

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

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges*. Journal for Educational Research Online, 5*, 189-216. https://www.pedocs.de/volltexte/2013/8430/pdf/JERO_2013_2_Pohl_Carstensen_Scaling_of_competence_tests.pdf

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

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. https://doi/10.1214/aos/1176344136

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

https://doi.org/10.1007/BF02294627

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

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

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

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

Appendix

Appendix A.

Overview of the items in the mathematical competence test Thuringia

Nr.	Item	Content area	Booklet	Response format
1	magcr511_c	space and shape	1, 4, 5, 7	MC
2	magcq581_c	quantity	1, 4, 5, 7	MC
3	magcq583_c	quantity	1, 4, 5, 7	MC
4	maa2r081_c	space and shape	1, 4, 5, 7	MC
5	maa2v082_c	change and relationship	1, 4, 5, 7	MC
6	mas2d071_c	data and chance	1, 4, 5, 7	MC
7	magcq591_c	quantity	1, 4, 5, 7	MC
8	mas2q011_c	quantity	1, 2, 6, 8	MC
9	maa2d111_c	data and chance	1, 2, 6, 8	SCR
10	maa2d112_c	data and chance	1, 2, 6, 8	MC
11	maa2d113_c	data and chance	1, 2, 6, 8	MC
12	magcv501_c	change and relationship	1, 2, 6, 8	MC
13	magcv502_c	change and relationship	1, 2, 6, 8	SCR
14	maa2r091_c	space and shape	1, 2, 6, 8	SCR
15	mas2r092_c	quantity	1, 2, 6, 8	MC
16	mas2v093_c	change and relationship	1, 2, 6, 8	MC
17	magcr561_c	space and shape	1, 2, 6, 8	MC
18	maa2r011_c	quantity	2, 3, 5, 7	MC
19	mas2v031_c	change and relationship	2, 3, 5, 7	MC
20	mas2v032_c	change and relationship	2, 3, 5, 7	MC
21	maa2d131_c	data and chance	2, 3, 5, 7	MC
22	maa2d132_c	data and chance	2, 3, 5, 7	MC
23	mas2q02s_c	quantity	2, 3, 5, 7	CMC
24	mas2q041_c	quantity	2, 3, 5, 7	MC
25	mas2v042_c	change and relationship	2, 3, 5, 7	SCR
26	mag9r061_c	space and shape	2, 3, 5, 7	SCR
27	maa2q021_c	quantity	2, 3, 5, 7	MC
28	magcr532_c	space and shape	3, 4, 6, 8	MC
29	mas2v061_c	change and relationship	3, 4, 6, 8	MC
30	mas2v062_c	change and relationship	3, 4, 6, 8	MC

31	mas2v063_c	change and relationship	3, 4, 6, 8	MC
32	magcd571_c	data and chance	3, 4, 6, 8	MC
33	magcr551_c	space and shape	3, 4, 6, 8	MC
34	magcd541_c	data and chance	3, 4, 6, 8	MC
35	maa2q071_c	quantity	3, 4, 6, 8	MC

Appendix B.

ConQuest Syntax for Estimating WLE Estimates in the Additional Study Thuringia

```
Title Additional Study Thuringia, MATHEMATICS: Rasch Model;


data filename.dat;
format pid 1-10 responses 12-46; /* insert number of columns with data*/
labels << labels.nam;


codes 0,1,2,3,4;


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

score (0,1)            (0,1)           !item (1-35);


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

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