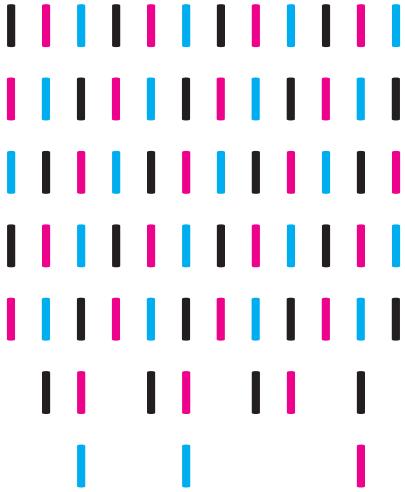

NEPS *SURVEY PAPERS*

Lara Aylin Petersen, Kristin Litteck, and Dunja Rohenroth

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 12

NEPS Survey Paper No. 75
Bamberg, September 2020 | updated April 2021

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

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

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

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 for Grade 12

Lara Aylin Petersen, Kristin Litteck, and Dunja Rohenroth

IPN – Leibniz Institute for Science and Mathematics Education, Kiel

Email address of lead author:

lpetersen@leibniz-ipn.de

Bibliographic Data:

Petersen, L. A., Litteck, K., & Rohenroth, D. (2020). NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 for Grade 12 (NEPS *Survey Paper* No. 75). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP75:2.0

Acknowledgment:

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

The present report has been modeled along previous reports published by the NEPS. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g. Kutscher & Scharl, 2020; Van den Ham, Schnittjer, & Gerken, 2018).

Note:

This manuscript has been modified in March 2021. A list of all modifications is given on the last page. The original version of the manuscript is available at https://doi.org/10.5157/NEPS:SP75:1.0.

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 for Grade 9

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and designs tests for assessing these different competence domains. 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 for the mathematical competence test in grade 12 for starting cohort 3 (starting fifth grade). The descriptive statistics for the data, the scaling model applied to estimate competence scores, and analyses performed to investigate the quality of the scale as well as the results of these analyses are explained. The mathematics test for grade 12 consists of 30 items (distributed among an easy and a difficult booklet) which represent different content areas as well as different cognitive components and use different response formats. The test was administered to 3,786 participants in grade twelve. A partial-credit 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. The results show that the items exhibited good item fit and measurement invariance across various subgroups. Moreover, the test showed a good reliability and a sufficiently broad range of item difficulties. As the correlations between the four content areas were very high in a multidimensional model and the model fit criteria favored the unidimensional model, unidimensionality of the test was assumed. Analyses of the missing responses showed that the test had too many items for the allocated test time. Overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. This paper describes the data available in the Scientific Use File and provides the R-Syntax for scaling the data.

Keywords

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

Content

1. Introduction 4
2. Testing Mathematical Competence 4
3. Data 5
3.1 The Design of the Study 5
3.2 Sample 6
3.3 Missing Responses 6
3.4 Scaling Model 7
3.5 Checking the Quality of the Test 7
3.6 Software 8
4. Results 9
4.1 Missing Responses 9
4.1.1 Missing responses per person 9
4.1.2 Missing responses per item and booklet 11
4.2 Parameter Estimates 12
4.2.1 Item parameters 12
4.2.2 Test targeting and reliability 13
4.3 Quality of the test 16
4.3.1 Distractor analyses 16
4.3.2 Item fit 17
4.3.3 Differential item functioning 17
4.3.4 Rasch-homogeneity 21
4.3.5 Unidimensionality 21
5. Discussion 22
6. Data in the Scientific Use File 23
6.1 Naming conventions 23
6.2 Linking of competence scores 23
6.2.1 Samples 23
6.2.2 Results 24
6.3 Mathematics competence scores 25
References 26
Appendix 30

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 grade 12 (wave 9) of starting cohort 3 (fifth grade). First, the main concepts of the mathematical competence test are introduced. Then, the mathematical competence data of the ninth wave of starting cohort 3 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 before public data release. Due to ongoing data protection and data cleaning issues, the data set in the SUF may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamental changes in the presented results.

2. Testing Mathematical Competence

The framework and test development for the mathematical competence test are described in 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 or two 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.

3. Data

3.1 The Design of the Study

The study was conducted in 2017 and assessed different competence domains including, among others, reading competence, information and communication technologies (ICT) literacy, and mathematical competence. The mathematics test was always administered as the third test (i.e., after the ICT literacy and reading test). All students received the test items in the same order (paper-pencil test).

In order to measure participants' mathematical competence with great accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies two different versions of the mathematic competence test were developed that differed in their average difficulty (i.e., an easy and a difficult booklet). Both test versions included 22 items that represented the four content areas (see Appendix A and Table 1) and the six process-related components[1]. 14 items were identical in both test versions, whereas 8 items were unique in each booklet. There were 30 items in total.

Table 1. Content Areas of Items in the Mathematics Test Grade 12

Content area	Frequency
Quantity	8
Space and shape	8
Change and relationships	7
Data and chance	7
Total number of items	30

The mathematics test included three types of response formats: simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR) (see Table 2). In MC items, the test taker had to find the correct response option from four or five available response options. In CMC items, a number of subtasks with two response options were presented. SCR items required the test taker to write down an answer into an empty box.

Table 2. Response Formats of Items in the Mathematics Test Grade 12

Response format	Frequency
Simple Multiple-Choice	27
Complex Multiple-Choice	1
Short-constructed response	2
Total number of items	30

[1] A more detailed description of the instruments used and, in particular, of the underlying framework of the mathematics competence test can be found on the NEPS website http://www.neps-data.de.

The panel study aimed at retesting all students that were initially included in the starting cohort 3 for fifth grade (see Van de Ham et al., 2018; Pohl et al., 2012). Because some students left their original schools during the course of the longitudinal study or left the school context altogether, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings: Students that remained at the same school as in the previous assessment were tested at school in a group setting; in contrast, students that left their original school were tracked and, subsequently, tested individually at home (for details regarding the data collection process, see the respective field report for wave 9; https://www.neps-data.de). Thus, the context of test administration differed between the two groups. Students who were still at school (usually in secondary school, in German "Gymnasium") always received the difficult mathematical competence test. The other participants who left the school context (e.g., students who graduated after grade 9) were assigned either the easy or the difficult booklet based on their estimated mathematical competence in the previous assessment (Van de Ham et al., 2018). Participants with an ability estimate below the sample's mean ability received the easy booklet, whereas participants with a mathematical competence above the sample's mean received the difficult booklet.

3.2 Sample

Overall, 3,786[2] persons from starting cohort 3 took the mathematics test in grade 12 (50.4% women). For one of them less than three valid responses were available. Because no reliable ability scores can be estimated based on such few valid responses, this case was excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 3,785 test takers. Of these, 1,635 participants received the easy booklet (setting: at home), and 2,150 received the difficult test version (setting: 389 at home; 1,761 at school). 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 given were coded as not-reached. Because of the branched testlet design, some items were not administered to all participants. For example, for respondents receiving the easy test 8 items from the difficult test were missing by design.

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

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

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 in order 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 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).

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 an R 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. In order 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 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 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), school type, and booklet. Moreover, DIF was also examined for the test difficulty. In order to test for measurement invariance, differential item functioning was estimated 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 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 mathematics test was constructed to measure a unidimensional competence score. The assumption of unidimensionality 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). The correlations among the dimensions as well as differences in model fit between the unidimensional model and the respective multidimensional models were used to evaluate the unidimensionality of the test. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) $Q3$. Because in case of locally independent items, the $Q3$ statistic tends to be slightly negative, we report the adjusted $Q3$ that has an expected value of zero. Following prevalent rules-of-thumb (Yen, 1993) values of $Q3$ falling below .20 indicate essential unidimensionality.

All analyses were first conducted for the different booklets and settings (e.g., item fits and measurement invariance) to check whether the three data sets could be merged into one data set for final analyses. The CMC item mas1q02s_sc3g12_c showed unsatisfactory step parameters in the school setting (the difficulty decreased with increasing number of points). Therefore, this item was scored dichotomously in the final analyses and thus all items were scored dichotomously. Because the analyses for each booklet (and setting) showed good fit and measurement invariance, only the analyses of the combined data are presented here.

3.6 Software

The IRT models were estimated in R version 3.6.0 (R Core Team, 2019) using the TAM package version 3.2.24 (Robitzsch et al., 2019).

4. Results

4.1 Missing Responses

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, 98.76 % of the test takers did not have any invalid response.

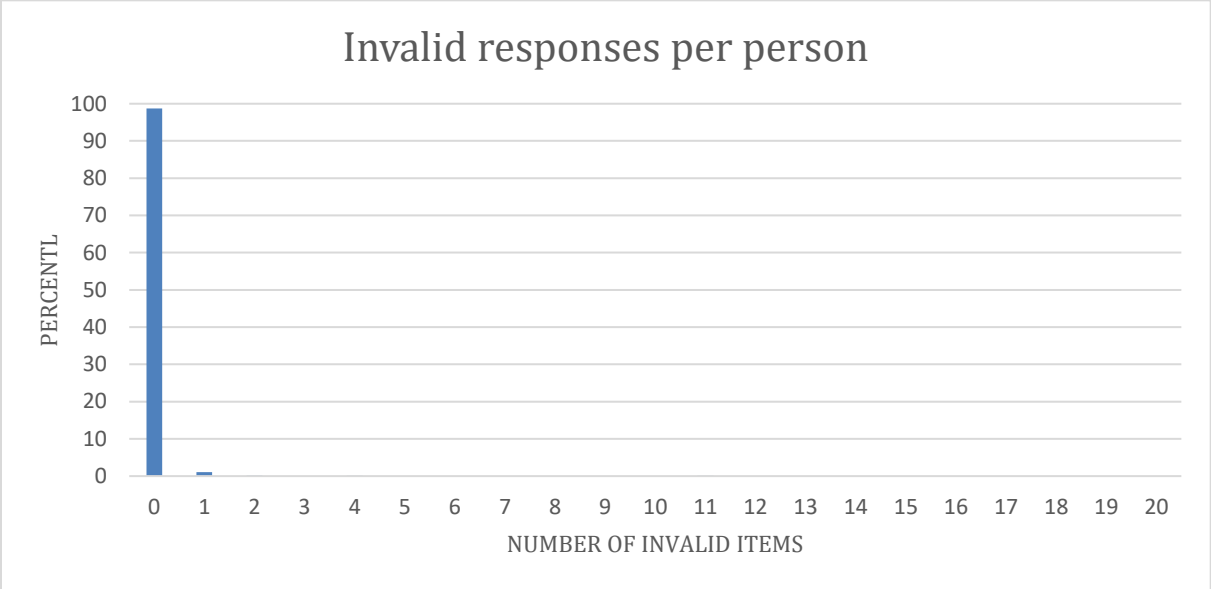

Figure 1. Number of invalid responses

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 65.5 % of the subjects omitted no item at all. Only 4.31 % of the subjects omitted more than 3 items.

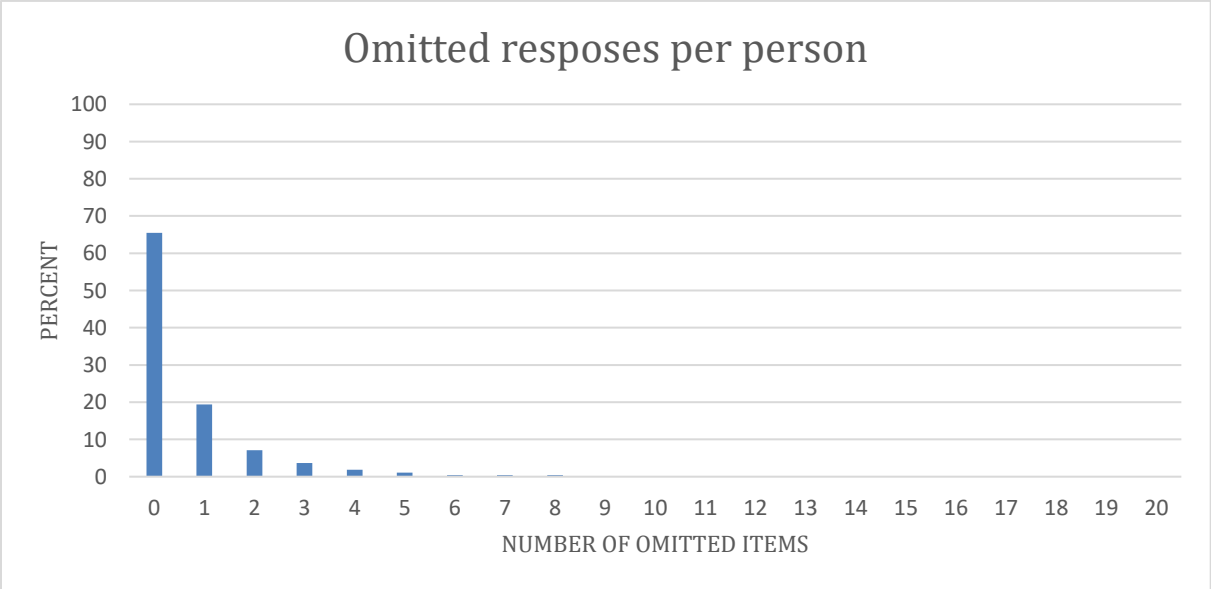

Figure 2. Number of omitted items

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, the number of not-reached items was rather high because many respondents were unable to finish the test within the allocated time limit. Only 61.35 % reached the end of the test. 21.9 % of the test takers did not reach one to five items. 4.36% of the participants did not reach more than 11 items (half of the test).

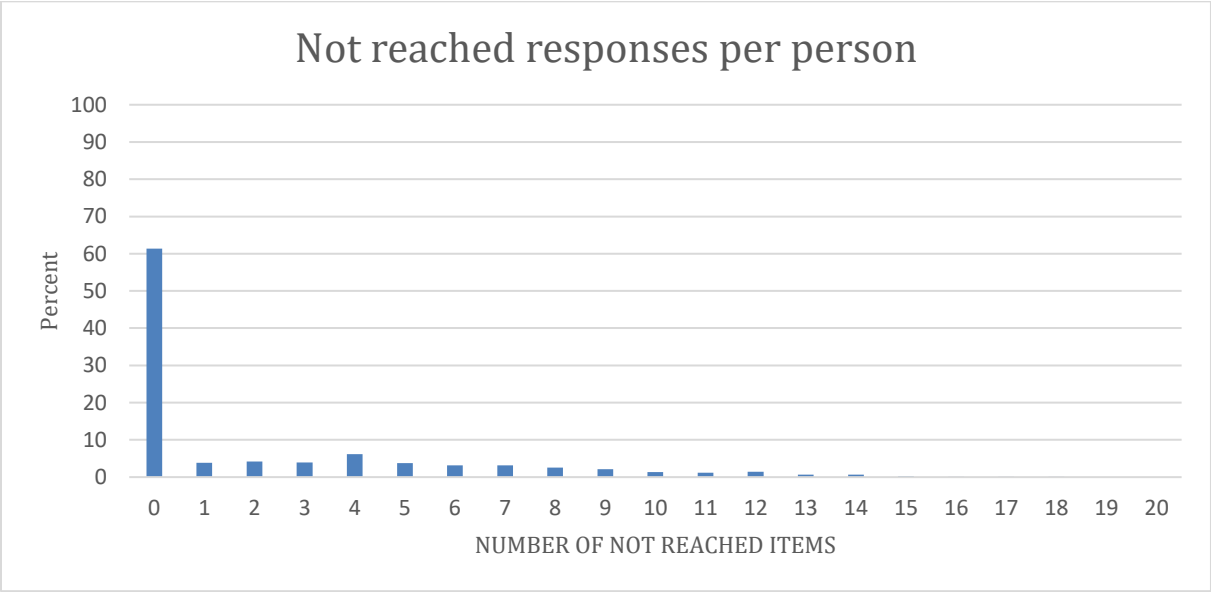

Figure 3. Number of not-reached items

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, 42.27 % of the subjects showed no missing response at all. 32.47 % showed more than three missing responses.

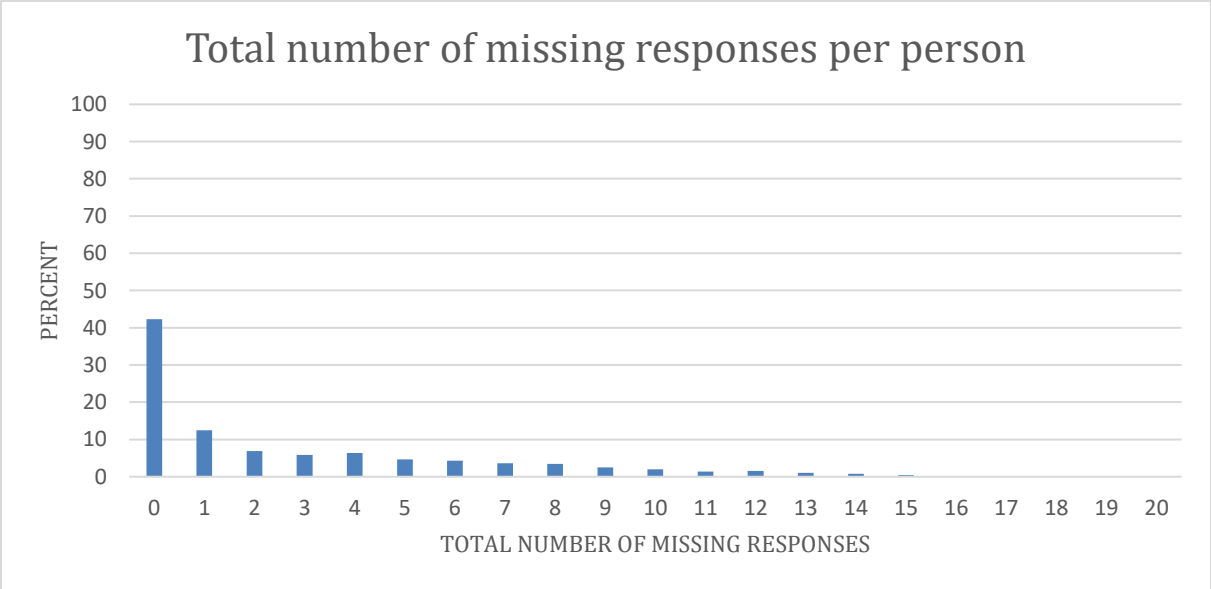

Figure 4. Total number of missing responses

In sum, the amount of invalid and omitted missing responses is small. The number of not reached items is, however, rather large and has the greatest impact on the total number of missing responses.

4.1.2 Missing responses per item and booklet

Tables 3 and 4 show the number of valid responses for each item in the two booklets, as well as the percentage of missing responses.

Table 3. Percentage of Missing Values for the Easy Booklet (setting: at home)

Item	***Position***	***N***	***NV***	***OM***	***NR***
maa3q071_sc3g12_c	1	1592	0.00	2.63	0.00
mag12v101_sc3g12_c	2	1561	0.06	4.46	0.00
mag12q121_sc3g12_c	3	1601	0.00	2.08	0.00
mag12v122_sc3g12_c	4	1568	0.00	4.10	0.00
maa3d131_sc3g12_c	5	1606	0.18	1.53	0.06
maa3d132_sc3g12_c	6	1578	0.12	3.24	0.12
mag12r091_sc3g12_c	7	1512	0.00	7.34	0.18
mag9r051_sc3g12_c	8	1604	0.18	1.41	0.31
mag9v011_sc3g12_c	9	1599	0.00	1.71	0.49
mag12d021_sc3g12_c	10	1600	0.00	1.22	0.92
mag12q051_sc3g12_c	11	1552	0.00	3.36	1.17
mag9d201_sc3g12_c	12	1577	0.00	0.92	2.63
mag9v121_sc3g12_c	13	1561	0.00	1.04	3.49
maa3r121_sc3g12_c	14	1568	0.00	0.31	3.79
mag12q111_sc3g12_c	15	1526	0.06	1.04	5.57
mag9r061_sc3g12_c	16	1164	0.55	19.76	8.50
maa3q101_sc3g12_c	17	1423	0.00	2.51	10.46
mag9q101_sc3g12_c	18	1349	0.00	1.71	14.56
mag12d071_sc3g12_c	19	1227	0.12	5.26	19.57
mag12r041_sc3g12_c	20	1243	0.37	1.10	22.51
mag12v131_sc3g12_c	21	1173	0.00	1.10	27.16
mag12v132_sc3g12_c	22	1154	0.06	0.00	29.36

Note. Position = Item position within test, *N* = Number of valid responses, NV = Percentage of respondents with an invalid response, OM = Percentage of respondents that omitted the item, NR = Percentage of respondents that did not reach item.

Table 4. Percentage of Missing Values for the Difficult Booklet (setting: at home and at school)

Item	*Position*	*N*	*NV*	*OM*	*NR*
maa3q071_sc3g12	1	2115	0.05	1.58	0.00
mag12v101_sc3g12	2	2092	0.09	2.60	0.00
mag12q121_sc3g12	3	2108	0.14	1.81	0.00
mag12v122_sc3g12	4	2034	0.14	5.26	0.00
mag12r011_sc3g12	5	2102	0.05	2.14	0.05
mag12v061_sc3g12	6	2085	0.09	2.88	0.05
mag12r091_sc3g12	7	1979	0.05	7.81	0.09
mag9r051_sc3g12	8	2111	0.05	1.35	0.42
mag12q081_sc3g12	9	2022	0.05	4.51	1.40
mag12d021_sc3g12	10	2092	0.05	0.37	2.28
mag12q051_sc3g12	11	1983	0.05	3.44	4.28
mag9d201_sc3g12	12	2015	0.05	0.56	5.67
mag9v121_sc3g12	13	1978	0.14	0.51	7.35
mas1q02s_sc3g12_c	14	1758	0.00	6.70	10.93
mas1d081_sc3g12	15	1782	0.19	2.84	14.09
maa3d112_sc3g12	16	1670	0.00	4.93	17.40
mag9r061_sc3g12	17	1473	0.14	9.81	21.53
maa3r011_sc3g12	18	1587	0.00	1.21	24.98
mag12d071_sc3g12	19	1357	0.28	4.65	31.95
mag12r041_sc3g12	20	1327	0.09	1.53	36.65
mag12v131_sc3g12	21	1253	0.00	1.16	40.56
mag12d031_sc3g12	22	1167	0.00	0.00	45.72

Note. Position = Item position within test, *N* = Number of valid responses, NV = Percentage of respondents with an invalid response, OM = Percentage of respondents that omitted the item, NR = Percentage of respondents that did not reach item.

Overall, the number of not valid responses per item was very small. The omission rates were acceptable, varying between 0.00% and 9.81/19.76% (item mag9r061_sc3g9_c, easy/difficult booklet). The number of persons that did not reach an item increased with the position of the item in the test up to 29.36% (easy booklet) and 45.72% (difficult booklet).

4.2 Parameter Estimates

4.2.1 Item parameters

We calculated the descriptive item parameters to check for possible estimation problems (Table 5). We further evaluated the relative frequency of the responses given before performing IRT analyses. The percentage of persons correctly responding to an item (relative

to all valid responses) varied between 20.13 % and 76.66 % across all items. On average, the rate of correct responses was 49.85 % (*SD* = 14.64 %). From a descriptive point of view, the items covered an acceptable wide range of correct responses.

The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -1.74 (mag12v132_sc3g12_c) and 1.86 (mag12q081_sc3g12_c) with a mean of -0.07. Overall, the item difficulties were reasonably well distributed around zero. Due to the large sample size, the standard errors of the estimated item difficulties were small (*SE*(ß) ≤ 0.07).

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 0.97, indicating that the test differentiated reasonably well between subjects. The reliability of the test (EAP/PV reliability = 0.77, WLE reliability = 0.74) was good.

The items covered a wide range of the ability distribution, although an additional very easy and very difficult item would have captured the extreme person abilities even better. Nevertheless, person abilities will be measured relative precisely on the whole ability spectrum.

Table 5. Item Parameters

Nr	Item	Pos1	Pos2	PC	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	aQ3
1	maa3q071_sc3g12_c	1	1	57.19	-0.34	0.04	1.07	1.92	0.43	0.86	0.03
2	mag12v101_sc3g12_c	2	2	53.46	-0.16	0.04	0.96	-2.96	0.50	1.21	0.03
3	mag12q121_sc3g12_c	3	3	32.30	0.88	0.04	1.07	3.99	0.37	0.67	0.03
4	mag12v122_sc3g12_c	4	4	48.61	0.06	0.04	1.06	4.08	0.41	0.74	0.03
5	mag12r011_sc3g12_c		5	45.58	0.46	0.05	0.97	-2.10	0.49	1.22	0.03
6	mag12v061_sc3g12_c		6	35.11	0.99	0.05	0.99	-0.51	0.46	1.05	0.02
7	mag12r091_sc3g12_c	7	7	34.26	0.76	0.04	1.12	7.24	0.32	0.49	0.03
8	mag9r051_sc3g12_c	8	8	60.62	-0.52	0.04	0.93	-4.95	0.52	1.41	0.03
9	mag12q081_sc3g12_c		9	20.13	1.85	0.06	0.94	-1.94	0.47	1.40	0.04
10	mag12d021_sc3g12_c	10	10	57.18	-0.36	0.04	1.04	3.02	0.41	0.79	0.02
11	mag12q051_sc3g12_c	11	11	27.89	1.11	0.04	1.06	2.91	0.36	0.71	0.02
12	mag9d201_sc3g12_c	12	12	67.37	-0.89	0.04	0.92	-4.91	0.52	1.50	0.03
13	mag9v121_sc3g12_c	13	13	44.48	0.24	0.04	0.93	-5.28	0.53	1.38	0.03
14	mas1q02s_sc3g12_c		14	51.54	0.33	0.05	0.93	-4.18	0.53	1.44	0.03
15	mas1d081_sc3g12_c		15	76.66	-1.17	0.06	0.98	-0.64	0.41	1.14	0.03
16	maa3d112_sc3g12_c		16	34.31	0.99	0.06	1.02	0.70	0.41	0.90	0.03
17	mag9r061_sc3g12_c	16	17	44.41	0.28	0.04	0.92	-5.52	0.54	1.48	0.04
18	maa3r011_sc3g12_c		18	54.95	-0.06	0.06	0.92	-4.50	0.53	1.59	0.04
19	mag12d071_sc3g12_c	19	19	40.44	0.33	0.04	1.15	8.72	0.29	0.42	0.04
20	mag12r041_sc3g12_c	20	20	44.63	-0.51	0.04	1.04	2.68	0.40	0.79	0.02
21	mag12v131_sc3g12_c	21	21	50.41	-0.20	0.04	1.11	6.51	0.34	0.56	0.03
22	mag12d031_sc3g12_c		22	57.24	-0.26	0.06	0.96	-1.90	0.48	1.20	0.03
23	maa3d131_sc3g12_c	5		56.72	-0.67	0.03	0.93	-3.52	0.52	1.45	0.04
24	maa3d132_sc3g12_c	6		26.62	0.85	0.06	0.89	-3.88	0.54	1.83	0.04
25	mag9v011_sc3g12_c	9		68.04	-1.24	0.06	0.99	-0.57	0.44	1.06	0.03

26	maa3r121_sc3g12_c	14	73.60	-1.57	0.06	0.98	-0.54	0.42	1.08	0.03
27	mag12q111_sc3g12_c	15	47.51	-0.25	0.06	1.01	0.37	0.44	0.96	0.03
28	maa3q101_sc3g12_c	17	41.25	0.03	0.06	1.00	0.04	0.44	0.99	0.02
29	mag9q101_sc3g12_c	18	68.59	-1.33	0.06	0.91	-3.26	0.52	1.68	0.04
30	mag12v132_sc3g12_c	22	74.44	-1.74	0.07	0.99	-0.20	0.41	1.10	0.03

Note. Pos.1 and Pos.2 = item position within the easy and difficult test versions, PC = Percentage correct, Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = adjusted average absolute residual correlation for item (Yen, 1993). For the dichotomous items. The item-total correlation corresponds to the point-biserial correlation between the correct response and the total score.

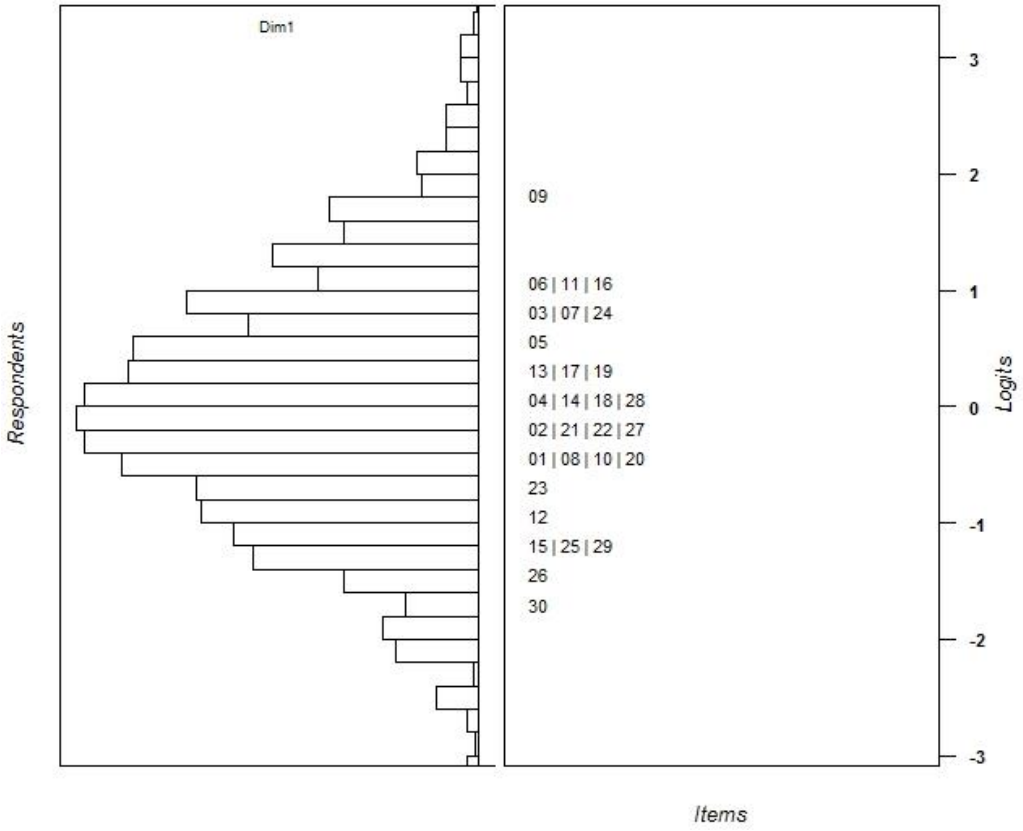

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 5).

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 on the basis of preliminary analyses.

Table 6 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). 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	0.27	-0.18
Minimum	0.13	-0.45
Maximum	0.45	-0.04

4.3.2 Item fit

The evaluation of the item fit was performed on the basis of 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 maa3d132_sc3g12_c) and the highest being 1.15 (item mag12d071_sc3g12_c). Only two items exhibited a *t*-value of the WMNSQ greater than |6| (items mag12v131_sc3g12_c and mag12d071_sc3g12_c). Thus, there was no indication of severe item over- or underfit. All item characteristic curves showed a good fit of the items. The correlation of the item score with the total score varied between .29 (item mag12d071_sc3g12_c) and .54 (items mag9r061_sc3g12_c and maa3d132_sc3g12_c), averaging at .45.

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, see Table 7). 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 school type (see Pohl & Carstensen, 2012, for a description of these variables). In addition, the effect of the two settings and the two booklets was also analyzed. Thus, we compared the two assessment settings (at school or at home) and booklets (difficult vs. easy) for the common items that were administered to all participants. Table 8 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 β(male) – β(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males compared to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 7, Cohen's *d*).

Gender: Overall, 1,867 (49.46 %) of the test takers were male and 1,908 (50.54 %) were female. On average, female test takers exhibited a lower mathematical competence than male test takers (main effect = -0.54 logits, Cohen's *d* = -0.56). An overall test for DIF (see Table 8) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) and Bayesian information criterion (BIC; Schwarz, 1978; takes the number of estimated parameters into account) slightly favored the model estimating DIF. But since there was no item with a considerable gender DIF above 0.6 logits (only four items above 0.4 logits), we assume no considerable overall gender effect.

Migration: There were 2,717 (71.78 %) participants without migration background, 690 (18.23 %) participants with migration background, and 378 (9.99 %) participants without a valid response. Only the first two 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 = -0.51 logits, Cohen's *d* = -0.53). There was no considerable DIF above 0.6 logits (item mag12v061_sc3g12_c had the highest DIF = 0.43). Moreover, the overall test for DIF using the BIC favored the main effects model (Table 8).

Table 7. Differential Item Functioning

Item	Gender		Migration		Books		School		Setting		Test difficulty	
	male vs. female		without vs. with		< 100 vs. > 100		No sec. vs. sec.		school vs. home		difficult vs. easy	
	DIF	Cohen's *d*	DIF	Cohen's *d*	DIF	Cohen's *d*	DIF	Cohen's *d*	DIF	Cohen's *d*	DIF	Cohen's *d*
maa3q071_sc3g12_c	0.08	0.09	0.09	0.09	0.06	0.07	0.06	0.07	-0.02	-0.02	-0.09	-0.11
mag12v101_sc3g12_c	-0.19	-0.20	0.07	0.07	-0.08	-0.09	-0.06	-0.07	0.06	0.07	0.09	0.12
mag12q121_sc3g12_c	0.29	0.31	-0.01	-0.01	0.19	0.20	0.10	0.11	-0.09	-0.11	-0.03	-0.04
mag12v122_sc3g12_c	0.40	0.42	0.18	0.19	0.15	0.16	0.25	0.28	-0.09	-0.12	-0.07	-0.09
mag12r011_sc3g12_c	-0.08	-0.08	-0.18	-0.18	0.02	0.02	-0.19	-0.21				
mag12v061_sc3g12_c	-0.55	-0.57	0.43	0.45	-0.01	-0.01	0.13	0.14				
mag12r091_sc3g12_c	0.25	0.26	0.04	0.04	0.33	0.36	0.58	0.65	-0.36	-0.45	-0.49	-0.62
mag9r051_sc3g12_c	0.23	0.24	-0.30	-0.32	-0.24	-0.26	-0.28	-0.32	0.30	0.37	0.37	0.46
mag12q081_sc3g12_c	0.06	0.06	-0.17	-0.18	-0.12	-0.13	-0.12	-0.13				
mag12d021_sc3g12_c	-0.33	-0.35	0.15	0.16	0.18	0.19	0.15	0.17	-0.17	-0.21	-0.23	-0.28
mag12q051_sc3g12_c	0.11	0.11	0.24	0.25	0.22	0.24	0.06	0.07	0.02	0.03	-0.02	-0.03
mag9d201_sc3g12_c	-0.09	-0.09	-0.03	-0.03	-0.13	-0.14	-0.57	-0.63	0.50	0.62	0.57	0.71
mag9v121_sc3g12_c	-0.12	-0.13	0.00	0.00	-0.14	-0.15	-0.16	-0.18	0.21	0.26	0.17	0.22
mas1q02s_sc3g12_c	-0.25	-0.26	-0.01	-0.01	0.07	0.07	0.10	0.11				
mas1d081_sc3g12_c	-0.06	-0.07	0.06	0.07	-0.04	-0.04	-0.15	-0.17				
maa3d112_sc3g12_c	-0.09	-0.10	0.25	0.26	-0.01	-0.02	0.05	0.06				
mag9r061_sc3g12_c	0.01	0.01	-0.16	-0.17	-0.18	-0.19	-0.25	-0.28	0.29	0.36	0.28	0.35
maa3r011_sc3g12_c	-0.23	-0.25	0.31	0.33	0.10	0.11	-0.02	-0.03				
mag12d071_sc3g12_c	0.53	0.56	0.14	0.15	0.13	0.14	0.47	0.52	-0.45	-0.55	-0.44	-0.54
mag12r041_sc3g12_c	0.50	0.52	0.24	0.25	0.04	0.04	0.10	0.12	0.06	0.07	0.06	0.07
mag12v131_sc3g12_c	0.37	0.39	0.30	0.31	0.27	0.29	0.19	0.21	-0.27	-0.33	-0.18	-0.22
mag12d031_sc3g12_c	-0.10	-0.10	-0.16	-0.16	-0.08	-0.09	0.05	0.06				
maa3d131_sc3g12_c	0.09	0.10	-0.27	-0.28	-0.24	-0.25	-0.56	-0.63				

	Gender		Migration		Books		School		Setting		Test difficulty	
maa3d132_sc3g12_c	-0.15	-0.16	-0.35	-0.36	-0.22	-0.23	-0.14	-0.15				
mag9v011_sc3g12_c	-0.16	-0.17	-0.35	-0.36	-0.09	-0.10	-0.08	-0.09				
maa3r121_sc3g12_c	-0.53	-0.56	-0.13	-0.14	-0.05	-0.05	0.10	0.11				
mag12q111_sc3g12_c	0.25	0.26	-0.09	-0.10	0.12	0.13	0.53	0.59				
maa3q101_sc3g12_c	-0.07	-0.07	0.09	0.10	0.06	0.07	-0.15	-0.16				
mag9q101_sc3g12_c	0.03	0.03	-0.34	-0.36	-0.05	-0.05	-0.06	-0.06				
mag12v132_sc3g12_c	-0.19	-0.20	-0.07	-0.07	-0.27	-0.29	-0.16	-0.18				
Main effect (with DIF)	-0.54	-0.56	-0.51	-0.53	0.67	0.71	0.93	1.03	-0.76	-0.95	-0.82	-1.03
Main effect (without DIF)	-0.50	-0.52	-0.49	-0.51	0.65	0.69	0.89	0.99	-0.77	-0.95	-0.82	-1.02

Note. Raw differences (DIF) between item difficulties with standardized differences (Cohen's *d*). Sec. = Secondary school (German: „Gymnasium"). No differences in item difficulty parameters are larger than 0.60 logits. All absolute standardized differences are not significantly greater than 0.4 (α = 5%; see Fischer et al., 2016).

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,372 (36.25 %) test takers with 0 to 100 books at home, 2,381 (62.90 %) test takers with more than 100 books at home, and 32 (0.01 %) test takers without any valid response. Group differences and DIF were investigated by using the first two groups. Participants with 100 or fewer books at home performed, on average, 0.67 logits (Cohen's *d* = 0.71) lower in mathematics than participants with more than 100 books. There was no item with a considerable DIF above 0.4 logits (item mag12r091_sc3g12_c reached the highest DIF = |0.33|). Moreover, the overall test for DIF using the BIC favored the main effects model without DIF (Table 8).

School: 1,628 test takers (43.01 %) were in secondary school (German: "Gymnasium") whereas 2,157 (56.99 %) were enrolled in other school types or dropped out of school. Test takers in secondary schools exhibited a higher average mathematics competence (0.93 logits, Cohen's *d* = 1.03) than subjects in other school types/ out of school. There was no considerable item DIF; no item exhibited DIF greater than 0.6 logits (five items exhibited DIF greater than 0.4 logits). Moreover, the overall test for DIF using the BIC favored the main effects model (Table 8).

Setting: The mathematical competence test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 1,761 person (46.53 %) received the test in small groups at school, whereas 2,024 (53.47 %) participants took the test individually at their private homes. Subjects who received the mathematical competence test at school performed on average -0.76 logits (Cohen's *d* = -0.95) worse than test takers at their private homes (regarding the common items). The overall test for DIF (Table 8) using the AIC und BIC slightly favored the DIF model. However, this difference must not be interpreted as a causal effect of the administration setting because respondents were not randomly assigned to the different settings. Rather, it is likely that the group testing in school may have led students to feel less motivated to try their hardest compared to the individual setting at home where the subjects were supervised much more closely by a test administrator. More importantly, all differences in item difficulties were smaller than 0.6 logits (only two of 14 items exhibited DIF greater than 0.4 logits).

Test difficulty: To estimate the participants' proficiency with great accuracy the participants received different tests that either included a larger number of easy or difficult items (see section 3.1 for the design of the study). A subset of 14 items were included in both tests and administered to all participants. For these common items we examined potential DIF across the two test versions (difficult versus easy). A subsample of 2,150 (56.80 %) persons received the difficult test and 1635 (43.20 %) persons received the easy test. As expected, subjects who were administered the difficult test scored on average -0.82 logits (Cohen's *d* = -1.03) lower than subjects who received the easy test. There was no DIF for the common items with regard to the test version. The largest difference in difficulties between the two groups was 0.57 logits (item mag9d201_sc3g12_c). The overall test for DIF (Table 8) using the AIC und BIC favored the DIF model.

Table 8. Comparison of Models With and Without DIF

DIF variable	Model	N	Deviance	Number of parameters	AIC	BIC
Gender	main effect	3,775	87,876.12	32	87,940.12	88,139.68
	DIF	3,775	87,628.64	61	*87,750.64*	*88,131.05*
Migration	main effect	3,407	79,451.46	32	79,515.46	*79,711.74*
	DIF	3,407	79,378.02	61	*79,500.02*	79,874.17
Books	main effect	3,785	87327.11	32	87,391,11	*87,590,75*
	DIF	3,785	87243.61	61	*87,365,61*	87,746,18
School	main effect	3,785	87,729.41	32	87,793.41	*87,993.05*
	DIF	3,785	87,525.41	61	*87,647.41*	88,027.98
Setting	main effect	3,785	58,483.47	16	58,515.47	58,615.29
	DIF	3,785	58,340.23	29	*58,398.23*	*58,579.16*
Test difficulty	main effect	3,785	58,410.06	16	58,442.06	58,541.88
	DIF	3,785	58,232.61	29	*58,290.61*	*58,471.54*

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. In order 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 0.42 (item mag12d071_sc3g12_c) and 1.83 (item maa3d132_sc3g12_c). The 2PL model (AIC = 87,592.94; BIC = 87,967.27; number of parameters = 60) fitted the data better than the Rasch model (AIC = 88,385.75; BIC = 88,579.15; number of parameters = 31). 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 Gauss-Hermite quadrature method. The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained (snodes=10000).

The variances, correlations and EAP Reliability of the four dimensions are shown in Table 9. All four dimensions exhibited a substantial variance. The correlations among the four dimensions were rather high and varied between .85 and .95. However, all correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95. see Carstensen, 2013). Moreover, the AIC and BIC favored the unidimensional Model (Table 10). 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 = .03, SD = .01) — the largest individual residual correlation was .04 — 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 9. Results of Four-Dimensional Scaling.

	Dim 1	Dim 2	Dim 3	Dim 4
Dim 1: Quantity (8 items)	(.97)			
Dim 2: Data and chance (8 items)	.95	(1.03)		
Dim 3: Space and shape (7 items)	.91	.93	(1.09)	
Dim 4: Change and relationships (7 items)	.94	.85	.94	(1.01)
EAP Reliability	.74	.75	.73	.75

Note. Variances of the dimensions are given in the diagonal and correlations are presented in the off-diagonal; Dim = Dimension.

Table 10. Comparison of the Unidimensional and the Four-Dimensional Model.

Model	N	Deviance	Number of parameters	AIC	BIC
Unidimensional	3785	88,323.72	31	*88,385.72*	*88,579.12*
Four-dimensional	3785	88,309.05	40	88,389.05	88,638.60

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

5. Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in grade 12 of starting cohort 3 and at describing how the mathematics competence scores were estimated. We investigated different kinds of missing responses and examined the item and test parameters to check the quality of the test. Further quality inspections were conducted by examining differential item functioning, testing Rasch homogeneity and investigating the tests' dimensionality.

Various criteria (WMNSQ, *t*-value of the WMNSQ, item characteristic curves) indicated a good fit of the items and measurement invariance across various subgroups (no item exceeded a DIF of 0.6 logits; indicating test fairness for the considered subgroups).

However, the amount of not-reached items was rather high (only 61.35 % reached the end of the test), indicating that the test had too much items for the allocated testing time. Other types of missing responses were reasonably small. The EAP reliability (r = 0.77) and the item distribution along the ability scale was good, indicating that the test distinguished relatively precisely for lower to higher abilities. Further, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with the total score) were acceptable. The high correlations between the four dimensions as well as a lower AIC and BIC indicated 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

6.1 Naming conventions

There are 30 items in the data set that are scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. The polytomous variable (mas1q02s_sc3g12_c) was also scored dichotomous for estimation of the mathematics competence score and scaling model (see section 5; all four subtasks correct = 1, three or fewer correct responses = 0). The dichotomous variables are marked with a 'sc3g12_c' behind their variable names; the polytomous variable is marked with a 's_sc3g12_c' behind its variable name.

6.2 Linking of competence scores

In starting cohort 3, the mathematics competence tests administered in grade 5, grade 7, grade 9, and grade 12 included different items that were constructed in such a way as to allow for an accurate measurement of mathematical competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared. Differences in observed scores would reflect differences in competencies as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competencies across grades, we adopted the linking procedure described in Fischer et al. (2016). The process of linking combines adjacent measurement points on the same scale. Therefore, the first wave of each competence scale within a cohort is used as a reference scale that all subsequent measurement waves will refer to. For the domain of mathematical competence, linking is achieved using overlapping items (also known as common items). For the linking procedure of the mathematical competences across grade 5 and 7 see Fischer et al. (2016), and across grade 7 and 9 see Van den Ham et al. (2018).

To link the tests of mathematics competence conducted in grade 9 and grade 12, six items which already were administered in grade 9 were, again, administered in grade 12. An empirical study that evaluated different link methods with regard to the appropriateness of linking NEPS data (Fischer et al., 2016) showed that the method of mean/mean linking (see Kolen & Brennan, 2004) is appropriate for the NEPS tests. Five of the six common items that were administered in grade 9 and grade 12 were found to be measurement invariant across the two measurement points. Therefore, they served as link items and the anchor-items design as described in Fischer et al. (2016) was used. For more information on the selection of link items and the method for linking the tests of mathematical competence see Fischer et al. (2016).

6.2.1 Samples

In starting cohort 3, a longitudinal subsample of 2,862 students participated at both measurement occasions (in grade 9 and also in grade 12). Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016.).

6.2.2 Results

To examine whether the two tests administered in the longitudinal sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model. For the two-dimensional model, the common items load on the first dimension and the unique items (i.e., the items included in only one test) load on the second dimension. In both grades, the information criteria slightly favored the two-dimensional model (AIC = 82,209.89/ BIC = 82,484.03 for grade 9, and AIC = 66,704.02/ BIC = 66,900.68 for grade 12), over the one-dimensional model (AIC =82,302.17/ BIC =82,564.39 for grade 9, and AIC = 66,907.66/ BIC = 67,092.40 for grade 12). We also examined the residual correlations for the one-dimensional models. The corrected Q_3 statistics indicated largely unidimensional scales in grade 9 ($M(aQ_3) = 0.00$, $SD(aQ_3) = 0.01$), and grade 12 ($M(aQ_3) = 0.03$, $SD(aQ_3) = 0.02$). This indicates that unidimensional scales can be assumed for the mathematics tests in grades 9 and 12, although the model test slightly favored the two-dimensional model.

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 longitudinal subsample showed a non-negligible shift in item difficulties comparing grade 9 and grade 12. The differences in item difficulties between the link subsample grade 9 and link subsample grade 12 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 11.

Table 11. DIF Analyses for the common items in the tests for mathematical competence in grades 9 *and* 12

Grade 9	Grade 12	Δσ	$SE_{\Delta\sigma}$	*t*	*F*
mag9d201_sc3g9_c	mag9d201_sc3g12_c	0.23	0.06	3.80	14.42
mag9q101_sc3g9_c	mag9q101_sc3g12_c	0.23	0.10	2.32	5.36
mag9r051_sc3g9_c	mag9r051_sc3g12_c	-0.03	0.06	-0.51	0.26
mag9r061_sc3g9_c	mag9r061_sc3g12_c	0.03	0.07	0.40	0.16
mag9v011_sc3g9_c	mag9v011_sc3g12_c	*-0.61*	0.09	-6.85	46.88
mag9v121_sc3g9_c	mag9v121_sc3g12_c	0.15	0.06	2.41	5.83

Note. Δσ = Difference in item difficulty parameters between grades 9 and 12 (positive values indicate easier items in grade 9); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis; F_{crit} = Critical value for the minimum effects hypothesis test for an α of .05; the degrees of freedom (df_1, df_2) are based on the number of measurement points (df_1 = *k*-1) and the number of test takers taking both tests (df_2 = n-1). The critical *F* (1, 2862) = 69.66. A non-significant test indicates measurement invariance. The differences in item difficulty parameters larger than 0.40 logits are indicated in italics.

The analyses of differential item functioning (Δσ) identified one item with considerable DIF greater than 0.6 logits (mag9v011_sc3g9_c/ mag9v011_sc3g12). Therefore, this item was excluded as a common item from the final linking procedure.

In the longitudinal subsample, the mean of the item difficulty parameters for the five common items was 0.132 in grade 9 and -0.464 in grade 12. Mean/mean linking (Loyd &

Hoover, 1980) resulted in a correction term of $c_{9\text{-}12}$ = 0.133 – (-0.464) = 0.596. The correction term for linking Grade 5 to Grade 7 was $c_{5\text{-}7}$ = 0.726 (Fischer et al., 2016) and for linking grade 7 to 9 was $c_{7\text{-}9}$ = 0.310 (this value does not correspond to the linking constant published in Van de Ham et al. (2016), because an error occurred in the calculation of the linking constant. Here the corrected value is shown. This has also been corrected in the SUF of the SC3). The sum of the correction terms $c_{5\text{-}7}$+ $c_{7\text{-}9}$ + $c_{9\text{-}12}$ = 1.632 was added to each item difficulty parameter derived in grade 12. The linked item parameters can be seen in Appendix C. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.048 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 Mathematics competence scores

In the SUF, manifest mathematics competence scores are provided in the form of two different WLEs (“ma12_sc1” and “ma12_sc1u”), including their respective standard error (“ma9_sc2” and “ma12_sc2u”). For “ma12_sc1u”, person abilities were estimated using the linked item difficulty parameters. As a result the WLE scores provided in “ma_sc1u” can be used for longitudinal comparisons between grades 5, 7, 9 and 12. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “ma12_sc1” are not linked to the underlying reference scale of grade 5. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The R Syntax for estimating the WLE is provided in Appendix B. For persons who either did not take part in the mathematics test or 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 denoted as not-determinable missing values. 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* (S. 313-327). Münster, Germany: Waxmann.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg, Germany: Leibniz Institute for Educational Trajectories. National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_I.pdf

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4939-0317-7

Kutscher, T., & Scharl, A. (2020). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 12* (NEPS Survey Paper No. 67). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_LXVII.pdf

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196. http://doi.org/10.1007/BF02294457

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

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 (JERO), 5*(2), 80-109. https://www.pedocs.de/volltexte/2013/8426/pdf/JERO_2013_2_Neumann_et_al_Modeling_and_assessing_mathematical_competencies.pdf

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. http://doi.org/10.1111/jedm.12028

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

K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel https://www.neps-data.de/Portals/0/Working%20Papers/WP_XV.pdf

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

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules. R package version 3.3-10.* https://CRAN.R-project.org/package=TAM

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

Van den Ham, A.-K., Schnittjer, I., & Gerken, A.-L. (2018). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 for Grade 9* (NEPS Survey Paper No. 38). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XXXVIII.pdf

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. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. http://doi.org/10.1007/s11618-011-0182-7

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*, 125-145. http://doi.org/10.1177/014662168400800201

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 tests SC3 grade 12

Item	Content area	Booklet or common	Setting	Response format
maa3q071_sc3g12_c	quantity	common	At home and at school	MC
mag12v101_sc3g12_c	change and relationship	common	At home and at school	MC
mag12q121_sc3g12_c	quantity	common	At home and at school	MC
mag12v122_sc3g12_c	change and relationship	common	At home and at school	MC
mag12r011_sc3g12_c	space and shape	difficult	At home and at school	MC
mag12v061_sc3g12_c	change and relationship	difficult	At home and at school	MC
mag12r091_sc3g12_c	space and shape	common	At home and at school	MC
mag9r051_sc3g12_c	space and shape	common	At home and at school	MC
mag12q081_sc3g12_c	quantity	difficult	At home and at school	MC
mag12d021_sc3g12_c	data and chance	common	At home and at school	MC
mag12q051_sc3g12_c	quantity	common	At home and at school	MC
mag9d201_sc3g12_c	data and chance	common	At home and at school	MC
mag9v121_sc3g12_c	change and relationship	common	At home and at school	MC
mas1q021s_sc3g12_c	quantity	difficult	At home and at school	CMC
mas1d081_sc3g12_c	data and chance	difficult	At home and at school	SCR
maa3d112_sc3g12_c	data and chance	difficult	At home and at school	MC
mag9r061_sc3g12_c	space and shape	common	At home and at school	SCR
maa3r011_sc3g12_c	space and shape	difficult	At home and at school	MC
mag12d071_sc3g12_c	data and chance	common	At home and at school	MC
mag12r041_sc3g12_c	space and shape	common	At home and at school	MC
mag12v131_sc3g12_c	change and relationship	common	At home and at school	MC
mag12d031_sc3g12_c	data and chance	difficult	At home and at school	MC
maa3d131_sc3g12_c	data and chance	easy	At home	MC
maa3d132_sc3g12_c	data and chance	easy	At home	MC
mag9v011_sc3g12_c	change and relationship	easy	At home	MC
maa3r121_sc3g12_c	space and shape	easy	At home	MC
mag12q111_sc3g12_c	quantity	easy	At home	MC
maa3q101_sc3g12_c	quantity	easy	At home	MC
mag9q101_sc3g12_c	quantity	easy	At home	MC
mag12v132_sc3g12_c	change and relationship	easy	At home	MC

Appendix B. R Syntax for fitting the partial credit model in starting cohort 3 grade 12

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

### load data
dat <- read_sav(file = "SC3_xTargetCompetencies_D_9-0-0.sav")
items <- c( [add the names of the items provided in Appendix A] )

### Score the CMC Item mas1q02s_sc3g12_c dichotomous
dat$mas1q02s_sc3g12_c <- recodeVar(dat$mas1q02s_sc3g12_c, 0:4, c(0, 0, 0, 0,1))

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

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

Appendix C. Original and linked item difficulties for the mathematics test in Grade 12.

	item	Common item	Original item difficulties	Linked item difficulties
1	maa3q071_sc3g12_c	no	-0.34	1.29
2	mag12v101_sc3g12_c	no	-0.16	1.47
3	mag12q121_sc3g12_c	no	0.88	2.52
4	mag12v122_sc3g12_c	no	0.06	1.69
5	mag12r011_sc3g12_c	no	0.46	2.09
6	mag12v061_sc3g12_c	no	0.99	2.62
7	mag12r091_sc3g12_c	no	0.76	2.39
8	mag9r051_sc3g12_c	yes	-0.52	1.11
9	mag12q081_sc3g12_c	no	1.85	3.49
10	mag12d021_sc3g12_c	no	-0.36	1.28
11	mag12q051_sc3g12_c	no	1.11	2.74
12	mag9d201_sc3g12_c	yes	-0.89	0.74
13	mag9v121_sc3g12_c	yes	0.24	1.88
14	mas1q02s_sc3g12_c	no	0.17	1.80
15	mas1d081_sc3g12_c	no	-1.17	0.46
16	maa3d112_sc3g12_c	no	0.99	2.62
17	mag9r061_sc3g12_c	yes	0.28	1.91
18	maa3r011_sc3g12_c	no	-0.06	1.57
19	mag12d071_sc3g12_c	no	0.33	1.97
20	mag12r041_sc3g12_c	no	-0.51	1.12
21	mag12v131_sc3g12_c	no	-0.20	1.44
22	mag12d031_sc3g12_c	no	-0.26	1.37
23	maa3d131_sc3g12_c	no	-0.67	0.97
24	maa3d132_sc3g12_c	no	0.85	2.49
25	mag9v011_sc3g12_c	yes	-1.24	0.39
26	maa3r121_sc3g12_c	no	-1.57	0.06
27	mag12q111_sc3g12_c	no	-0.25	1.38
28	maa3q101_sc3g12_c	no	0.03	1.67
29	mag9q101_sc3g12_c	yes	-1.33	0.30
30	mag12v132_sc3g12_c	no	-0.74	0.90

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Table 6). Linked item difficulty parameters were derived by adding $c_{5\text{-}12}$ to the original item parameters.

List of modifications

	Date	Page	Modification
1.	March 2021	Page 17	Corrected main effects for gender and migration
2.	March 2021	Page 19 (Table 7)	Corrected main effects
3.	March 2021	Page 20	Corrected main effects for books, school, setting, and test difficulty