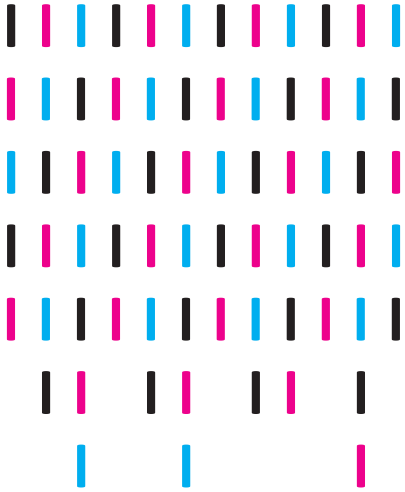

NEPS SURVEY PAPERS

Ann-Katrin van de Ham, Insa Schnittjer, and

Anna-Lena Gerken

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 3 IN GRADE 9

NEPS Survey Paper No. 38
Bamberg, June 2018

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

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

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

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

Ann-Katrin van den Ham, Insa Schnittjer, and Anna-Lena Gerken

IPN – Leibniz Institute for Science and Mathematics Education, Kiel

Email address of lead author:

vandenham@ipn.uni-kiel.de

Bibliographic Data:

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: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP38:1.0

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

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 on mathematical competence for starting cohort 3 - ninth 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 nine consists of 34 items which represent different content areas as well as different cognitive components and use different response formats. The test was administered to 4,890 participants in grade nine. A partial-credit model was used for scaling 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 shows a good reliability. As the correlations between the four content areas are very high in a multidimensional model, the assumption of unidimensionality seems adequate. Among the challenges of this test is the lack of very difficult items. Overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. This paper describes the data available in the Scientific Use File and provides ConQuest-Syntax for scaling the data – including the necessary item parameters.

Keywords

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

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span. 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 and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

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

In this paper, the results of these analyses are presented for mathematical competence in grade 9 (fifth wave) of starting cohort 3 (fifth grade). First, the main concepts of the mathematical competence test are introduced. Then, the mathematical competence data of the fifth 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.

The present report has been modeled on the previous reports (Pohl, Haberkorn, Hardt, & Wiegand, 2012; Haberkorn, Pohl, Hardt, & Wiegand, 2012). 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 cleansing 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 a single task related to it; sometimes there were two tasks. Each of the items belonged to one of the following content areas:

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

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.

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 several, usually four, available response options. In CMC items a number of subtasks with two response options were presented. SCR items required the test taker to write down an answer into an empty box.

3. Data

3.1 The Design of the Study

The study was conducted in 2014 and assessed different competence domains including scientific literacy, ICT literacy, mathematical competence, reading speed, and orthography. In order to control for the effects of test duration and test position, the tests were rotated. For this purpose, the sample was split into two groups that received the tests in different sequence. Assignment to the test sequence was random. About half of the sample completed the ICT test followed by the science test, while the other group completed the two tests in the opposite order (see Table 1). After a short break, all participants received the reading speed test, followed by the mathematics test and the orthography test. In order to measure participants' mathematics 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 three different versions of the mathematics competence test were developed that differed in their average difficulty (i.e., an easy, a medium, and a difficult test). Each test included 23 items that represented the four content areas (see Table 2) and the process-related components[1]. In order to evaluate the quality of these items, extensive preliminary analyses were conducted that identified an unsatisfactory model fit for item mag9v551_c (see 4.3.4 for an explanation). Therefore, this item was excluded from the final scaling procedure. Regarding the remaining items, there were seven common items in all three tests and at least 10 common items between two booklets.

Table 1: Design of the study

Rotation A		Rotation B
ICT		Science easy
Science		ICT
15 minute break		
Reading Speed		
Mathematics easy	Mathematics medium	Mathematics difficult
Orthography		

Overall, 34 different items with different response formats were used. Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test booklet. Participants were assigned either to the easy, medium, or the difficult test based on their estimated mathematics competence in the previous assessment in grade 7 (Schnittjer & Gerken, 2017).

[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 characteristics of the final set of 34 items are summarized in Tables 2 and 3. Table 2 shows the distribution of the four content areas, whereas Table 3 shows the distribution of response formats. One subtask of one CMC item (mag9d393_c) was excluded from the analyses due to an unsatisfactory item fit.

Table 2: Content Areas of Items in the Mathematics Test Grade 9

Content area	Frequency
Quantity	13
Space and shape	7
Change and relationships	7
Data and chance	7
Total number of items	34

Table 3: Response Formats of Items in the Mathematics Test Grade 9

Response format	Frequency
Simple Multiple-Choice	28
Complex Multiple-Choice	5
Short-constructed response	1
Total number of items	34

3.2 Sample

Overall, 4,890[2] persons from starting cohort 3 took the mathematics test in grade 9. 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 4,888 test takers. Of these, 1,382 participants received the easy test, 1,625 received the medium test, and 1,881 received the difficult test version. 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 response due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and finally, e) multiple kinds of missing responses 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

[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.

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 medium test and 12 items from the difficult test were missing by design. As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the 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 partial credit model (PCM; Masters, 1982). CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. Categories of polytomous variables with less than N = 200 responses were collapsed in the analyses in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items. For 4 of the 5 CMC items (mag9d05s_c, mag9d09s_c, mag9r10s_c, mag9r14s_c) categories were collapsed (see Appendix A). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). Mathematical competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 6.

3.5 Checking the Quality of the 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. All analyses were conducted for the whole test and for the different booklets, respectively.

Before aggregating the subtasks of CMC items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square error (WMNSQ), the respective t-value, pointbiserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC variables that were included in the final scaling model.

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

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (|*t*-value| > 6) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (|*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 (equal to the discrimination value as computed in ConQuest) 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 PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The 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). To estimate this multidimensional model, the Gauss-Hermite quadrature method in ConQuest was used (10 nodes per dimension). The

correlations between the subdimensions 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) Q_3. Because, in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality.

We ran all analyses separately for the three booklets and with the combined data. Because the analyses for each of the three booklets showed good fit, only the analyses of the combined data are presented here.

3.6 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

4. Results

4.1 Missing Responses

Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was small. In fact, 97.5% of the test takers gave no invalid response.

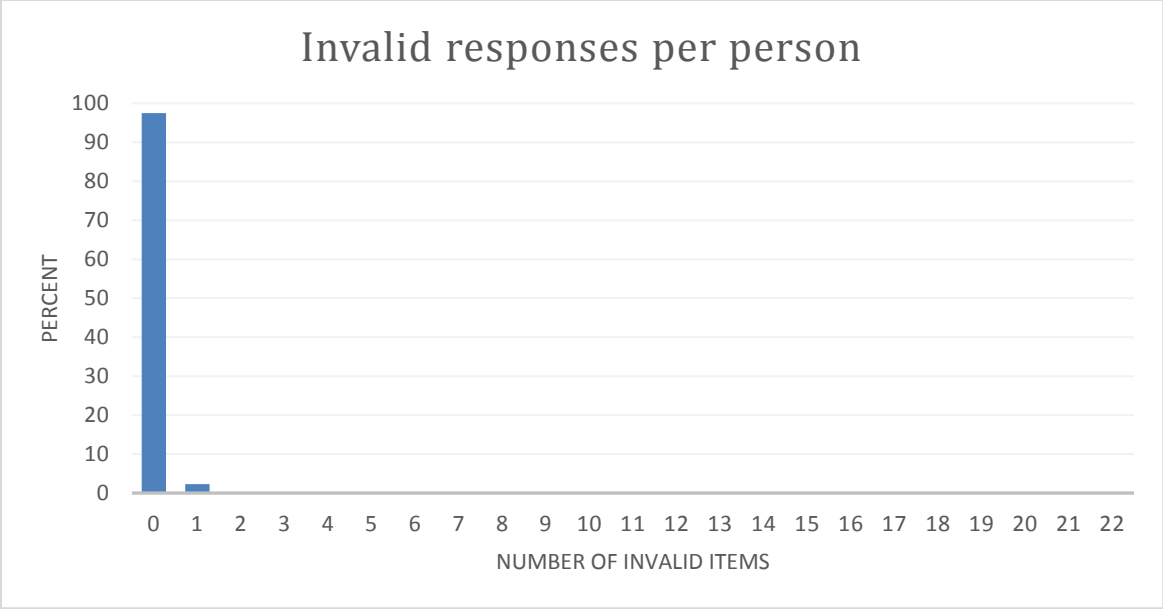

Figure 1: Number of invalid responses

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 67.7 % of the subjects omitted no item at all. 2.6 % of the subjects omitted more than 3 items.

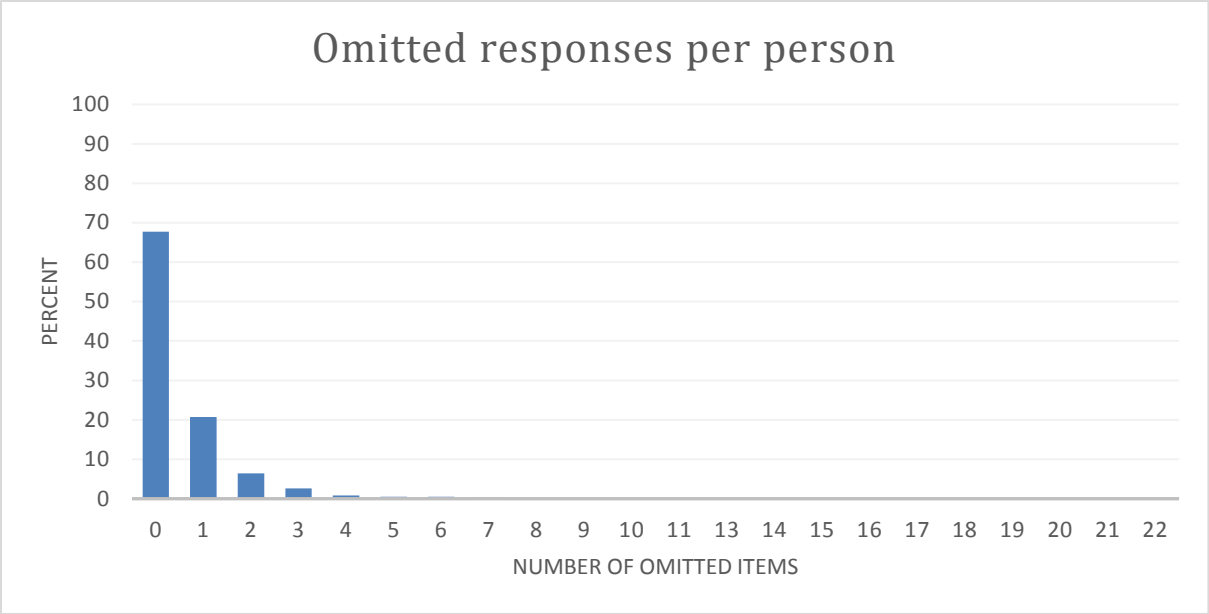

Figure 2: Number of omitted items

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, 94.1 % reached the end of the test. 5 % of the test takers had not reached one to five items. Only 0.9% of the participants had not reached more than five items.

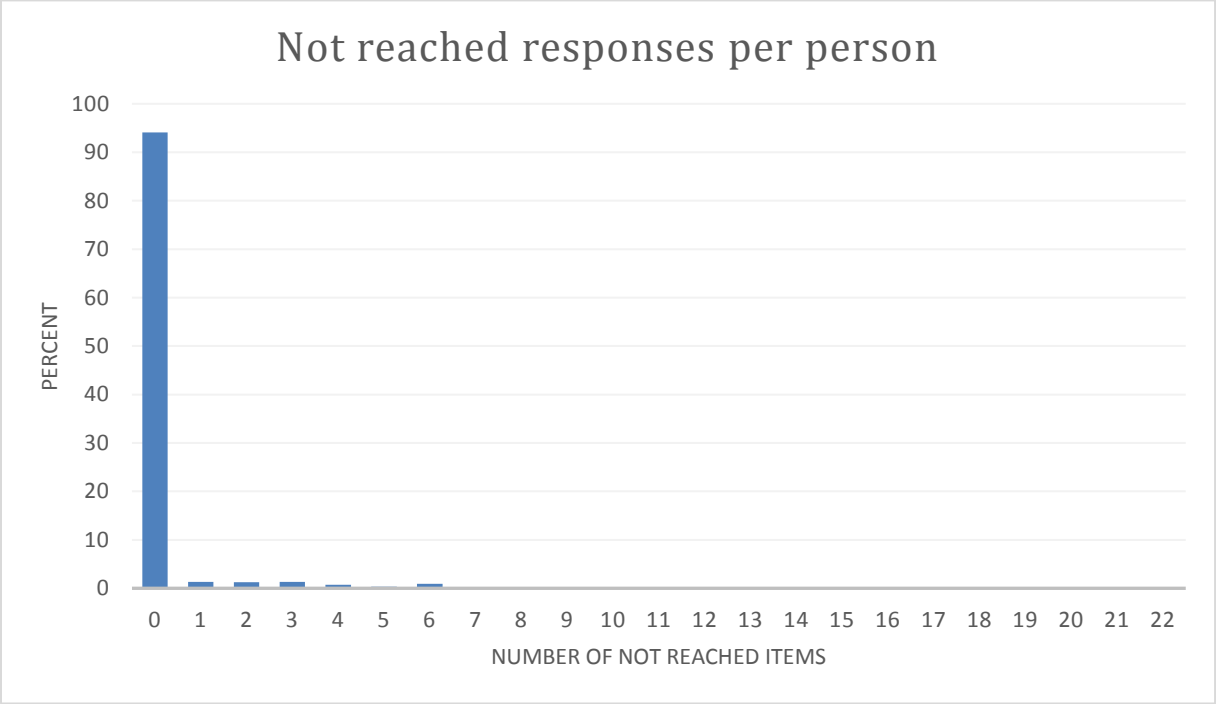

Figure 3: Number of not-reached items

Figure 4 shows the total number of missing responses per person, which is the sum of invalid, omitted, not-reached, and not-determinable missing responses. In total, 63.4 % of the subjects showed no missing response at all. 5.4 % showed more than three missing responses.

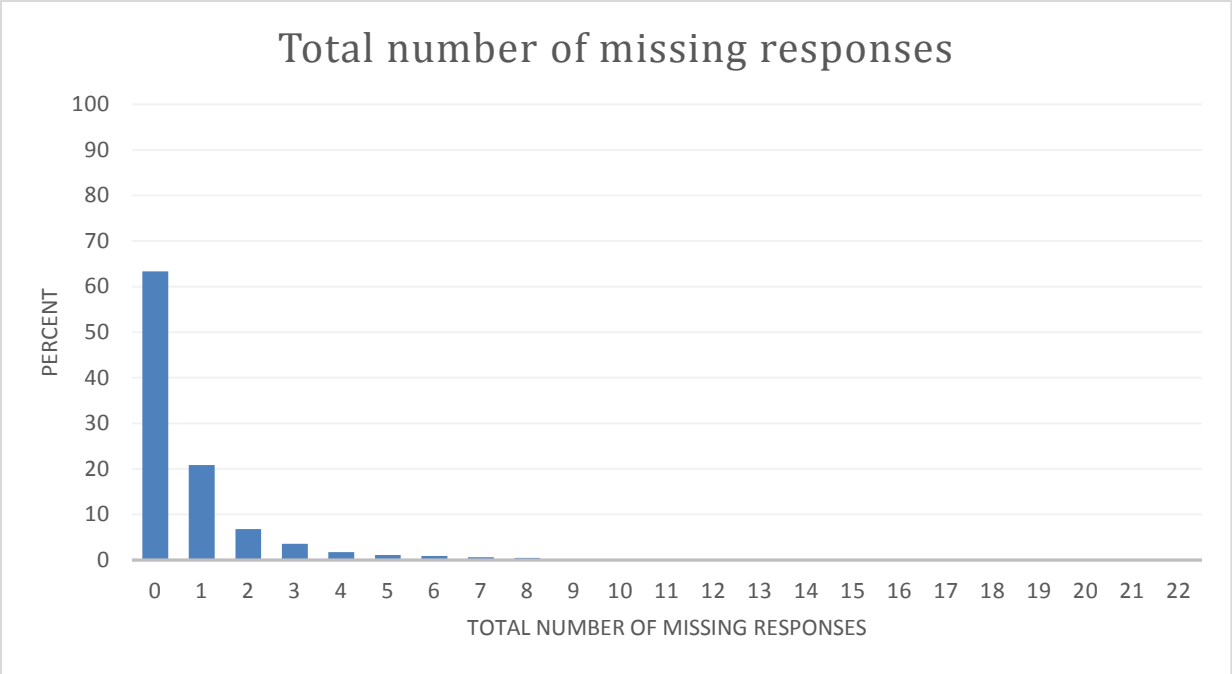

Figure 4: Total number of missing responses

Overall, there was a negligible number of not-reached and invalid responses. The number of omitted items was reasonable.

Missing responses per item

Tables 4 to 6 show the number of valid responses for each item in the three booklets, as well as the percentage of missing responses. Overall, the number of omitted responses per item was very small. The omission rates were acceptable, varying between 0.11% (item mag9v081_c, difficult booklet) and 11.06% (item mag9r061_sc3g9_c, difficult booklet), except for two items that had an omission rate of 22.21% and 19.75% (item mag9r061_sc3g9_c, easy booklet and mag9r061_sc3g9_c, medium booklet). The number of persons that did not reach an item increased with the position of the item in the test up to 8.72%. The number of invalid responses varied from 0.00% (various items in all three booklets) to 1.59% (mag9r061_sc3g9_c, easy booklet).

Table 4: Percentage of Missing Values for the Easy Booklet

Item	***Position***	***N***	***NR***	***OM***	***NV***
mag9q071_sc3g9_c	1	1365	0.00	1.16	0.07
mag9v131_sc3g9_c	2	1365	0.00	1.09	0.14
mag9v13s_sc3g9_c	3	1255	0.00	8.97	0.22
mag9d05s_c	4	1352	0.00	2.10	0.07
mag9r111_sc3g9_c	5	1340	0.00	2.82	0.22
mag9q021_c	6	1366	0.00	0.94	0.22
mag9d151_sc3g9_c	7	1366	0.00	0.80	0.36

mag9q161_c	8	1374	0.00	0.43	0.14
mag9v011_sc3g9_c	9	1354	0.00	1.81	0.22
mag9v012_sc3g9_c	10	1342	0.00	2.89	0.00
mag9q011_c	11	1370	0.00	0.87	0.00
mag9d201_sc3g9_c	12	1367	0.00	1.01	0.07
mag9r191_sc3g9_c	13	1366	0.00	0.94	0.22
mag9q181_sc3g9_c	15	1371	0.00	0.58	0.22
mag9d061_c	16	1360	0.00	1.59	0.00
mag9r061_sc3g9_c	17	1053	0.00	22.21	1.59
mag9q151_c	18	1352	0.58	1.37	0.22
mag9q101_sc3g9_c	19	1308	0.72	4.41	0.22
mag9r14s_c	20	1313	1.01	3.26	0.72
mag9v091_sc3g9_c	21	1312	1.52	3.40	0.14
mag9d131_c	22	1322	2.10	2.17	0.07
mag9r10s_c	23	1310	2.89	1.16	1.16

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

The item on position 14 was excluded from the analyses due to an unsatisfactory item fit (see section 3.1).

Table 5: Percentage of Missing Values for the Medium Booklet

Item	***Position***	***N***	***NR***	***OM***	***NV***
mag9d111_c	1	1603	0.00	1.29	0.06
mag9v131_sc3g9_c	2	1609	0.00	0.86	0.12
mag9v13s_sc3g9_c	3	1512	0.00	6.95	0.00
mag9r261_sc3g9_c	4	1590	0.00	2.09	0.06
mag9r111_sc3g9_c	5	1589	0.00	2.22	0.00
mag9q021_c	6	1612	0.00	0.80	0.00
mag9d151_sc3g9_c	7	1606	0.00	0.80	0.37
mag9r051_sc3g9_c	8	1606	0.00	1.05	0.12
mag9v011_sc3g9_c	9	1613	0.00	0.74	0.00
mag9v012_sc3g9_c	10	1595	0.00	1.72	0.12
mag9q011_c	11	1594	0.00	1.91	0.00
mag9d201_sc3g9_c	12	1607	0.00	1.11	0.00
mag9q041_c	13	1617	0.00	0.49	0.00
mag9v121_sc3g9_c	14	1599	0.00	1.60	0.00

mag9q181_sc3g9_c	15	1615	0.00	0.62	0.00
mag9d09s_c	16	1547	0.00	4.74	0.06
mag9r061_sc3g9_c	17	1289	0.00	19.75	0.92
mag9q081_sc3g9_c	18	1568	0.92	2.46	0.12
mag9q101_sc3g9_c	19	1549	1.35	3.26	0.06
mag9q021_sc3g9_c	20	1528	1.78	4.06	0.12
mag9v091_sc3g9_c	21	1517	3.08	3.45	0.12
mag9d131_c	22	1524	4.12	2.03	0.06
mag9r10s_c	23	1514	5.17	1.23	0.43

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

Table 6: Percentage of Missing Values for the Difficult Booklet

Item	*Position*	*N*	*NR*	*OM*	*NV*
mag9d111_c	1	1869	0.00	0.58	0.05
mag9q031_c	2	1774	0.00	5.58	0.11
mag9d05s_c	3	1859	0.00	1.17	0.00
mag9r261_sc3g9_c	4	1858	0.00	1.22	0.00
mag9r111_sc3g9_c	5	1864	0.00	0.90	0.00
mag9q121_c	6	1875	0.00	0.32	0.00
mag9d151_sc3g9_c	7	1872	0.00	0.27	0.21
mag9r051_sc3g9_c	8	1849	0.00	1.70	0.00
mag9q161_c	9	1869	0.00	0.48	0.16
mag9v011_sc3g9_c	10	1867	0.00	0.74	0.00
mag9v012_sc3g9_c	11	1859	0.00	1.17	0.00
mag9d201_sc3g9_c	12	1868	0.00	0.64	0.05
mag9q041_c	13	1874	0.00	0.37	0.00
mag9v121_sc3g9_c	14	1867	0.00	0.74	0.00
mag9d09s_c	16	1766	0.00	6.01	0.11
mag9r061_sc3g9_c	17	1666	0.00	11.06	0.37
mag9q081_sc3g9_c	18	1827	1.12	1.75	0.00
mag9r14s_c	19	1838	1.44	0.64	0.21
mag9q021_sc3g9_c	20	1758	2.82	3.62	0.11
mag9v091_sc3g9_c	21	1743	4.84	2.50	0.00
mag9q211_sc3g9_c	22	1745	6.75	0.32	0.16
mag9v081_c	23	1715	8.72	0.11	0.00

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

The item on position 15 was excluded from the analyses due to an unsatisfactory item fit (see section 3.1).

4.2 Parameter Estimates

Item parameters

In order to a) get a first rough descriptive measure of item difficulty and b) check for possible estimation problems, we evaluated the relative frequency of the responses given before performing IRT analyses. Regarding each subtask of a CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 11.55 % and 81.48 % across all items. On average, the rate of correct responses was 46.09 % (*SD* = 16.81 %). From a descriptive point of view, the items covered an acceptable wide range of difficulties.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are presented in Table 7. The step parameters for polytomous variables are summarized in Table 8. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.23 (item mag9q181_sc3g9_c) and 2.34 (item mag9r261_sc3g9_c) with a mean of -0.131. Overall, the item difficulties were reasonably well distributed around the students with medium ability. Yet, there were fewer items with very low of very high difficulty. However, there were only two items with a difficulty of 2 or above and two items with a difficulty in the lower section of -2 or below. Due to the large sample size, the standard errors of the estimated item difficulties were small (*SE*(ß) ≤ 0.07).

Table 7: Item Parameters

	Item	PC	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q3
1	mag9d151_sc3g9_c	76.51	-1.499	0.041	0.89	-5.8	0.48	2.15	0.03
2	mag9d201_sc3g9_c	48.04	0.068	0.036	1.02	1.5	0.44	1.09	0.03
3	mag9d05s_c	n.a.	-1.923	0.054	1.01	0.5	0.44	0.59	0.03
4	mag9d061_c	65.34	-1.679	0.065	0.98	-0.6	0.42	1.67	0.04
5	mag9d111_c	56.13	0.028	0.041	1.06	4.3	0.38	0.77	0.03
6	mag9d09s_c	n.a.	0.792	0.048	1.02	1.2	0.41	0.46	0.03
7	mag9d131_c	33.22	0.236	0.046	0.98	-1.4	0.44	1.40	0.03
8	mag9q021_sc3g9_c	51.37	0.113	0.042	0.99	-0.8	0.46	1.18	0.02
9	mag9q071_sc3g9_c	41.46	-0.537	0.062	1.06	2.9	0.35	0.92	0.03
10	mag9q081_sc3g9_c	46.81	0.432	0.041	1.04	3.2	0.40	0.87	0.03
11	mag9q101_sc3g9_c	45.33	-0.365	0.044	0.97	-2.4	0.47	1.52	0.03
12	mag9q181_sc3g9_c	81.48	-2.23	0.054	0.96	-1.5	0.39	1.89	0.04
13	mag9q211_sc3g9_c	59.65	0.036	0.057	1	0.1	0.42	0.98	0.03
14	mag9q121_c	53.38	0.577	0.054	1.1	5.7	0.32	0.54	0.03
15	mag9q151_c	31.33	-0.051	0.065	1.05	2	0.34	0.93	0.03
16	mag9q161_c	42.91	0.404	0.044	0.97	-1.6	0.49	1.23	0.03
17	mag9q021_c	72.46	-1.643	0.048	1.04	1.7	0.35	1.10	0.03
18	mag9q041_c	29.72	1.375	0.044	1.05	2.7	0.37	0.82	0.02
19	mag9q011_c	52.28	-0.619	0.044	1.01	0.9	0.42	1.25	0.03
20	mag9q031_c	23.71	2.015	0.062	0.99	-0.4	0.42	1.19	0.03
21	mag9r051_sc3g9_c	52.31	0.2	0.041	1.03	2.3	0.41	0.92	0.03
22	mag9r061_sc3g9_c	24.37	1.14	0.042	0.94	-3.4	0.5	1.41	0.03
23	mag9r111_sc3g9_c	58.92	-0.509	0.036	1.01	0.9	0.45	1.20	0.03
24	mag9r191_sc3g9_c	47.61	-0.82	0.062	1.12	5.7	0.29	0.58	0.03
25	mag9r261_sc3g9_c	15.40	2.342	0.053	0.94	-1.8	0.42	1.44	0.02
26	mag9r10s_c	n.a.	-1.096	0.056	1.02	0.9	0.31	0.55	0.03
27	mag9r14s_c	n.a.	-1.276	0.051	0.96	-1.8	0.49	0.70	0.03
28	mag9v011_sc3g9_c	69.11	-1.042	0.038	0.92	-5.1	0.49	1.74	0.03

29	mag9v012_sc3g9_c	54.68	-0.288	0.036	0.95	-4	0.49	1.43	0.03
30	mag9v091_sc3g9_c	43.62	0.142	0.037	0.94	-4.9	0.51	1.50	0.03
31	mag9v121_sc3g9_c	32.89	1.187	0.043	0.99	-0.8	0.44	1.13	0.03
32	mag9v131_sc3g9_c	39.31	0.018	0.044	1.1	6.2	0.32	0.69	0.04
33	mag9v13s_sc3g9_c	n.a.	-1.087	0.047	1.1	4.3	0.33	0.40	0.03
34	mag9v081_c	38.06	1.115	0.057	1.01	0.7	0.42	1.01	0.03

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 =Average absolute residual correlation for item (Yen, 1983).

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).

Table 8: Step Parameters (with Standard Errors) of Polytomous Items

Item	step 1	step 2	step 3
mag9d05s_c	-0.119 (0.067)	0.559 (0.071)	-0.439
mag9d09s_c	- 0.743 (0.044)	-0.345 (0.048)	1.088
mag9r10s_c	-0.435 (0.039)	0.435	
mag9r14s_c	-0.618 (0.057)	0.204 (0.056)	0.414
mag9v13s_sc3g9_c	0.026 (0.060)	-0.681 (0.061)	0.655

Note. The last step parameter is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In these analyses, the mean of ability was constrained to be zero. The variance was estimated to be 1.197, indicating that the test differentiated reasonably well between subjects. The reliability of the test (EAP/PV reliability =0.814, WLE reliability = 0.812) was good.

The extent to which the item difficulties and location parameters were targeted toward the test persons' abilities is shown in Figure 5. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The items covered a wide range of the ability distribution of test persons. However, there were no very difficult items. As a consequence, subjects with a low or medium ability will be measured relatively precisely, while subjects with a high mathematical competence will have a larger standard error.

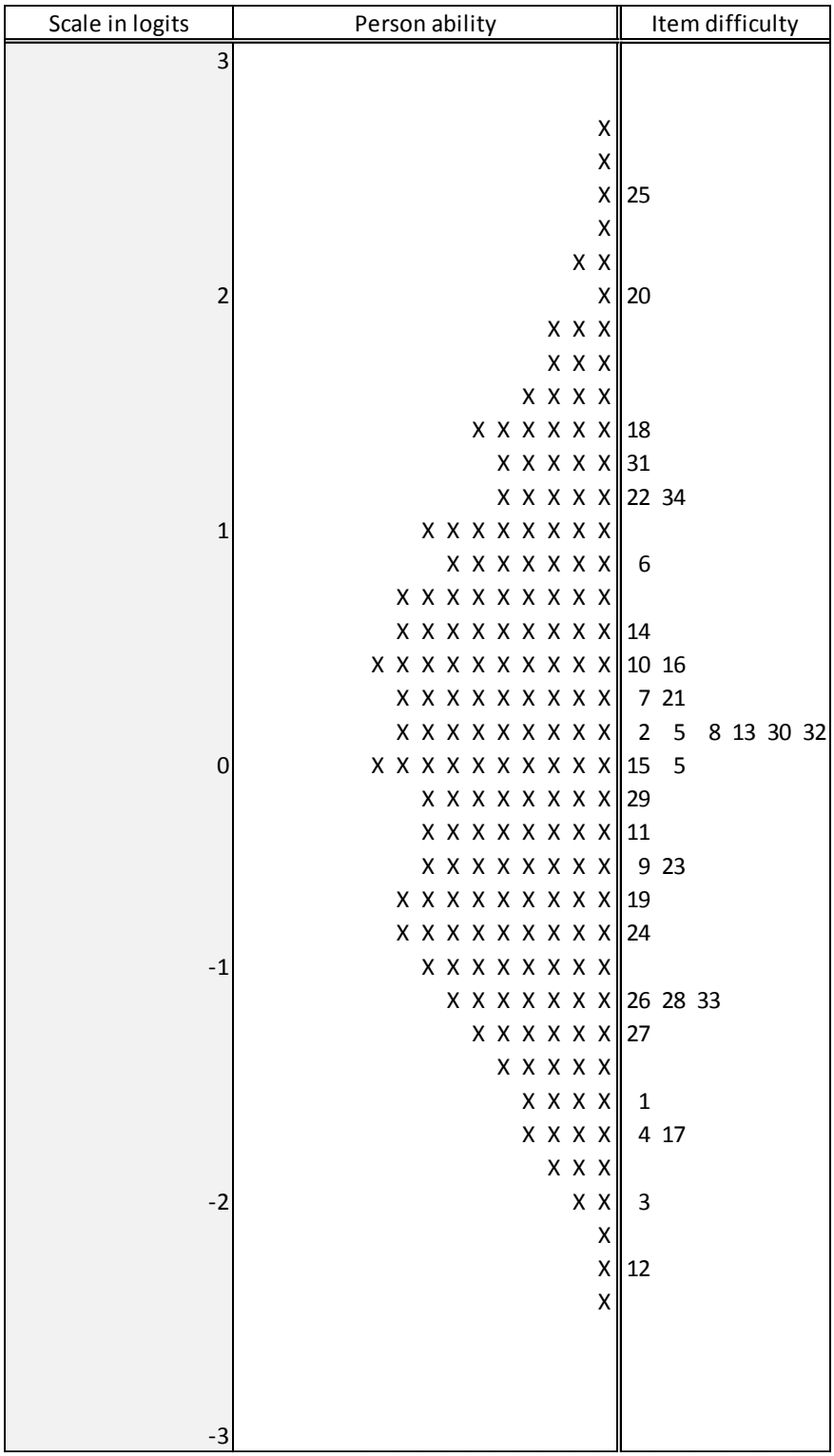

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph, with each 'X' representing 28 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 7).

4.3 Quality of the test

Fit of the subtasks of complex multiple-choice items

Before the responses to the subtasks of the CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the subtasks together with the simple multiple-choice items in a Rasch model. Counting the subtasks of CMC items separately, there were 64 items. The rates of correct responses given to the subtasks of the CMC items ranged from 30.71% to 93.07%. With one exception, the subtasks showed a good item fit with the WMNSQ ranging between 0.89 and 1.15 and the respective *t*-values between -3.20 and 10.7. The only subtask exhibiting unsatisfactory item fit (WMNSQ of 1.25, *t*-value of 15.3 and a respective item discrimination of -0.03) was excluded from further analysis. The good model fit of the other subtasks justified their aggregation to polytomous variables for each item.

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 treating all subtasks of the CMC items as single items.

Table 9 shows a summary of point-biserial correlations between response and ability for correct and incorrect responses restricted to MC items (only the items where subjects were asked to choose between distractors).

Table 9: Point-Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.40	-0.17
Minimum	0.26	-0.49
Maximum	0.51	-0.05

Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC and polytomous CMC items. Overall, the item fit was good. Values of WMNSQ were close to 1 with the lowest value being 0.89 (item mag9d151_sc3g9_c) and the highest being 1.12 (item mag9r191_sc3g9_c). Only one item exhibited a *t*-value of the WMNSQ greater than |6| and none exceeded a value of |8|. 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 (mag9r191_sc3g9_c) and .51 (mag9v091_sc3g9_c). Overall, the test showed an average correlation of .42.

Differential item functioning

We examined test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables gender, the number of books at home, the school type, and migration background (see Pohl & Carstensen, 2012, for a description of these variables), as well as for the difficulty of the booklet. Table 10 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.

Table 10: Differential Item Functioning (Absolute Differences Between Difficulties)

Item	**Sex**	**Migration**	**School**	**Books**		**Booklet**	
	male vs. female	without vs. with	no sec. vs. sec.	<100 books vs.>100 books	easy vs. medium	medium vs. difficult	easy vs. difficult
mag9d151_sc3g9_c	0.19	-0.27	0.42	0.48	-0.59	-0.21	-0.69
mag9d201_sc3g9_c	-0.06	0.26	0.10	-0.06	0.58	-0.40	0.28
mag9d05s_c	0.24	-0.09	-0.06	0.13			0.47
mag9d061_c	0.03	0.07	-0.19	-0.11			
mag9d111_c	0.46	0.09	0.04	-0.08		0.28	
mag9d09s_c	0.06	-0.11	0.09	0.13		0.15	
mag9d131_c	0.21	-0.41	0.25	-0.02	-0.07		
mag9q021_sc3g9_c	0.17	0.03	0.26	0.07		-0.15	
mag9q071_sc3g9_c	-0.12	0.09	-0.36	0.10			
mag9q081_sc3g9_c	-0.54	-0.09	-0.22	-0.25		0.33	
mag9q101_sc3g9_c	0.27	0.07	0.06	-0.03	0.00		
mag9q181_sc3g9_c	0.41	-0.05	0.13	0.07	-0.26		
mag9q211_sc3g9_c	-0.03	0.20	0.08	-0.11			
mag9q121_c	0.14	0.46	-0.24	-0.14			
mag9q151_c	0.26	0.29	-0.13	-0.29			
mag9q161_c	-0.11	0.13	-0.01	0.04			0.12
mag9q021_c	-0.19	0.08	-0.50	-0.18	0.20		
mag9q041_c	0.20	-0.15	-0.17	-0.11		0.12	
mag9q011_c	0.20	0.10	0.04	-0.07	0.08		
mag9q031_c	-0.47	0.08	-0.44	-0.11			
mag9r051_sc3g9_c	-0.03	0.02	-0.22	0.04		0.01	
mag9r061_sc3g9_c	0.00	-0.03	0.08	0.27	-0.02	-0.16	-0.09
mag9r111_sc3g9_c	0.35	-0.02	-0.01	-0.04	0.03	0.25	0.39
mag9r191_sc3g9_c	0.02	0.26	-0.50	-0.14			
mag9r261_sc3g9_c	-0.18	-0.24	0.03	0.06		-0.11	

mag9r10s_c	0.12	-0.29	0.02	0.24	-0.09		
mag9r14s_c	0.16	-0.11	0.16	0.10			-0.09
mag9v011_sc3g9_c	-0.29	-0.33	0.12	0.05	-0.28	-0.20	-0.38
mag9v012_sc3g9_c	-0.21	0.08	0.09	-0.01	-0.30	0.06	-0.14
mag9v091_sc3g9_c	-0.41	0.02	0.19	-0.04	-0.12	-0.09	-0.11
mag9v121_sc3g9_c	-0.37	-0.18	-0.18	-0.11		-0.10	
mag9v131_sc3g9_c	-0.03	0.07	-0.13	-0.19	0.44		
mag9v13s_sc3g9_c	0.21	0.37	0.09	-0.05	0.23		
mag9v081_c	-0.57	0.12	-0.26	0.03			
Main effect (model with DIF)	-0.280	-0.482	1.258	0.878	-1.048	-1.124	-2.302
Main effect (model without DIF)	-0.280	-0.484	1.242	0.876	-1.046	-1.120	-2.294

Gender: Overall, 2,464 (50.4 %) of the test takers were female and 2,424 (49.59 %) were male. On average, female students exhibited a lower mathematical competence than male students (main effect = 0.28 logits, Cohen's *d* = 0.258). There was no item with a considerable gender DIF. The only items for which the difference in item difficulties between the two groups exceeded 0.4 logits were item mag9d291_c (0.464 logits) and item mag9q181_sc3g9_c (0.412 logits).

Migration: There were 3,384 (69.23 %) participants without migration background, 1,204 (24.63 %) participants with migration background, and 300 (6.14 %) 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 = 0.48 logits, Cohen's *d* = 0.449). There was no considerable DIF comparing the two groups. One item showed a small DIF between the two groups with the highest difference in item difficulties being 0.46 logits (mag9d321_c).

School: Overall, 2,569 subjects (52.56%) who took the mathematics test attended secondary school (German: "Gymnasium") whereas 2,319 (47.44%) were enrolled in other school types. Subjects in secondary schools showed a higher mathematics competence on average (1.24 logits, Cohen's *d* = 1.393) than subjects in other school types. There was no noteworthy item DIF; no item exhibited DIF greater than 0.6 logits. The only item for which the difference in item difficulties between the two groups exceeded 0.4 logits was item mag9d151_sc3g9_c (0.420 logits).

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,172 (23.98 %) test takers with 0 to 100 books at home, 2,976 (60.88 %) test takers with more than 100 books at home, and 740 (15.14 %) test takers without any valid response. Group differences and DIF were investigated by using the first two groups. Participants with 100 or less books at home performed, on average, 0.88 logits (Cohen's *d* = 0.869) lower in mathematics than participants with more than 100 books. The only item for

which the difference in item difficulties between the two groups exceeded 0.4 logits was item mag9d151_sc3g9_c (0.478 logits).

Booklet: To estimate the participants' proficiency with great accuracy, the participants received different tests with low, medium or high difficulty (see section 3.1 for the design of the study). The booklet with low difficulty and the booklet with medium difficulty shared a subset of 15 items in common. The easy and the difficult booklet shared 10 common items. The easy and the difficult booklet shared 10 common items. For these common items we examined potential DIF across the respective versions (easy versus difficult, easy versus medium and medium versus difficult). A subsample of 1,382 (28.27%) persons received the easy test, 1,625 (33.24%) received the medium test and 1,881 (38.48%) persons received the difficult test. As expected, subjects who were administered the easy test scored on average -2.30 logits (Cohen's d = 2.12) lower on the common items than subjects who received the difficult test. There was no noticeable DIF for the common items with regard to the test version. The largest difference in difficulties between the two groups was 0.466 logits (item mag9d27s_c). Subjects who were administered the medium test scored on average 1.048 logits (Cohen's d = 0.965) higher on the common items than subjects who received the easy test. There was no noteworthy item DIF. The only items for which the difference in item difficulties between the two groups exceeded 0.4 logits were items mag9d201_sc3g9_c (0.582 logits) and item mag9v131_sc3g9_c (0.442 logits). Participants who took the medium test performed on average -1.124 logits (Cohen's d = 1.113) lower on the common items than participants who took the difficult test. There was no DIF for the common items.

In Table 7, the models including only main effects are compared with those that additionally estimated DIF. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for all DIF variables. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents from overparametrization of models. Using BIC, the more parsimonious models including only the main effects of migration status, gender, school type, books and booklet (medium vs. easy), respectively, were preferred over the more complex DIF models. However, BIC preferred the models including both main effect and DIF effect of gender and booklet (medium vs. difficult and easy vs. difficult), respectively, to the models including only the respective main effect.

Table 11: Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	main effect	141,851.55	45	141,941.55	142,233.80
	DIF	141,522.48	79	141,680.48	142,193.55
Migration	main effect	129,573.07	45	129,663.07	129,951.32
	DIF	129,462.07	79	129,620.07	130,126.11
School	main effect	140,381.81	45	140,471.81	140,764.06
	DIF	140,231.09	79	140,389.09	140,902.15
Books	main effect	137,252.08	45	137,342.08	137,633.03
	DIF	137,156.86	79	137,314.86	137,825.63
Booklet easy vs. medium	main effect	57,255.97	20	57,295.97	57,416.15
	DIF	57,082.87	35	57,152.87	57,363.18

Booklet medium vs. difficult	main effect	623,18.58	19	623,56.58	624,73.66
	DIF	622,22.20	34	622,90.20	624,99.72
Booklet easy vs. difficult	main effect	401,31.64	16	401,63.64	402,61.09
	DIF	400,08.91	26	400,60.91	402,19.26

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 generalized partial credit model (GPCM) to the data. The estimated discrimination parameters are depicted in Table 7. They ranged between 0.40 (item mag9v13s_sc3g9_c) and 2.15 (item mag9d151_sc3g9_c). The GPCM (AIC = 142,155.55, BIC = 142,817.99, number of parameters = 102) fitted the data better than the PCM (AIC = 143,771.15, BIC = 144,212.78, number of parameters = 68). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework (see Pohl & Carstensen, 2012, 2013 for a discussion of this issue), and, thus, the PCM was used to model the data and to estimate competence scores.

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).

To estimate this multidimensional model, the Gauss-Hermite quadrature estimation implemented in ConQuest was used. The number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained (nodes=10). The variances and correlations of the four dimensions are shown in Table 12. All four dimensions exhibited a substantial variance. The correlations among the three dimensions were rather high and varied between .89 and .97. However, five of the six correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013).

Model fit between the unidimensional model and the four-dimensional model is compared in Table13. The comparison showed that, according to the fit-indices, the four-dimensional model described the data slightly better than the unidimensional model. Additionally, for the unidimensional model the average absolute residual correlations as indicated by the Q_3 statistic (see Table 7) were quite low (M = .02, SD = .03)—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 12: Results of Four-Dimensional Scaling. Variance of the Dimensions are Depicted in the Diagonal, Correlations are given in the Off-Diagonal.

	Dim 1	Dim 2	Dim 3	Dim 4
Data and chance (7 items)	(1.355)			
Quantity (13 items)	.966	(1.069)		
Space and shape (7 items)	.885	.901	(1.501)	
Change and relationships (7 items)	.915	.934	.938	(1.448)

Table 13: Comparison of the Unidimensional and the Four-Dimensional Model.

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	141,932.73	44	142,020.73	142,306.49
Four-dimensional	141,850.80	53	141,956.80	142,301.01

5. Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in grade 9 of starting cohort 3 and at describing how the mathematics competence scores were estimated.

The amount of different kinds of missing responses was evaluated. Different kinds of missing responses were rather low. Furthermore, item as well as test quality were examined. Indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, item characteristic curves – the items exhibited a good item fit. Also, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test is fair to the considered subgroups. The test had a good reliability and the item distribution along the ability scale was acceptable, that is the test distinguished relatively precisely for lower or medium abilities, but showing a lack of difficult items. The high correlations between the four dimensions as well as a lower 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 34 items in the data set that are either scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (corresponding to the CMC items) indicating the number of correctly answered subtasks. The dichotomous variables are marked with a '_c' behind their variable names; the polytomous variable is marked with a 's_c' behind its variable name. In the scaling model the polytomous variables are scored as 0.5 for each category.

6.2 Linking of competence scores

In starting cohort 3, the mathematics competence tests administered in grade 5, grade 7 and grade 9 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, Rohm, Gnambs, and Carstensen (2016). The process of linking combines adjacent measurement points on the same scale. As such, the first wave of each competence scale within a cohort is used as a reference scale that all subsequent measurement waves will refer to. For the domain of mathematical competence, linking 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). In order to link the tests of mathematics competence conducted in grade 7 and grade 9, six items which already were administered in grade 7 were, again, administered in grade 9. 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 7 and grade 9 were found to be measurement invariant across the two measurement points. As such, they served as link items. Therefore 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 subsample of 4,720 students participated at both measurement occasions, in grade 7 and also in grade 9. 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 favored the two-

dimensional model, AIC = 118,978 and BIC = 119,152 for Grade 7, and AIC = 137,005 and BIC = 137,302 for Grade 9, over the one-dimensional model, AIC =119,029 and BIC =119,190 for Grade 7, and AIC = 137,210 and BIC = 137,494 for grade 9. Therefore, we also examined the residual correlations for the one-dimensional models. The corrected Q_3 statistics indicated largely unidimensional scales in Grade 7, $M(Q_3) = 0$, $SD(Q_3) = 0.02$, and Grade 9, $M(Q_3) = 0.02$, $SD(Q_3) = 0.03$. This indicates that unidimensional scales can be assumed for the mathematics tests in Grades 7 and 9.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 3 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 14.

Table 14: *DIF Analyses for the common items in the tests for mathematical competence in Grades* 7 *and* 9.

Grade 5	**Grade 7**	**Δσ**	$SE_{\Delta\sigma}$	***F***
mag9d151_sc3g7_c	mag9d151_sc3g9_c	-0.31	0.06	32.02
mag9q071_sc3g7_c	mag9q071_sc3g9_c	-0.23	0.07	9.95
mag9q181_sc3g7_c	mag9q181_sc3g9_c	-0.10	0.07	2.13
mag9v011_sc3g7_c	mag9v011_sc3g9_c	0.01	0.05	0.03
mag9v012_sc3g7_c	mag9v012_sc3g9_c	0.05	0.06	0.72
mag9v091_sc3g7_c	mag9v091_sc3g9_c	0.59	0.06	109.42

Note. Δσ = Difference in item difficulty parameters between Grades 7 and 9 (positive values indicate easier items in Grade 7); $SE_{\Delta\sigma}$ = 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; 4,719) = 104.98. A non-significant test indicates measurement invariance.

The analyses of differential item functioning identified one item with significant DIF (mag9v091_sc3g7_c / mag9v091_sc3g9_c). Therefore, this item was excluded as anchor item. The mathematics competence tests administered in the two grades were linked using the "mean/mean" method using the five measurement invariant anchor items (see Fischer et al., 2016).

In the longitudinal subsample, the mean item difficulty parameters for the five common items were -0.198 in Grade 7 and -0.992 in Grade 9 (see Table 14). Mean/mean linking (Loyd

& Hoover, 1980) resulted in a correction term of $c_{7\text{-}9}$ = -0.198 – (-0.992) = 0.794. The correction term for linking Grade 5 to Grade 7 was $c_{5\text{-}7}$ = 0.771 (Fischer et al., 2016). The sum of the correction terms $c_{5\text{-}7}$+ $c_{7\text{-}9}$= 1.565 was added to each item difficulty parameter derived in grade 9 and, thus, resulted in the linked item parameters (see Table 15).

Table 15: Original and linked item difficulty parameters for the mathematics test in Grade 9.

item	Common item	Original Item difficulties	Linked Item difficulties
mag9d151_sc3g9_c	yes	-1.499	0.066
mag9d201_sc3g9_c	no	0.068	1.633
mag9d05s_c	no	-1.923	-0.358
mag9d061_c	no	-1.679	-0.114
mag9d111_c	no	0.028	1.593
mag9d09s_c	no	0.792	2.357
mag9d131_c	no	0.236	1.801
mag9q021_sc3g9_c	no	0.113	1.678
mag9q071_sc3g9_c	yes	-0.537	1.028
mag9q081_sc3g9_c	no	0.432	1.997
mag9q101_sc3g9_c	no	-0.365	1.200
mag9q181_sc3g9_c	yes	-2.23	-0.665
mag9q211_sc3g9_c	no	0.036	1.601
mag9q121_c	no	0.577	2.142
mag9q151_c	no	-0.051	1.514
mag9q161_c	no	0.404	1.969
mag9q021_c	no	-1.643	-0.078
mag9q041_c	no	1.375	2.940
mag9q011_c	no	-0.619	0.946
mag9q031_c	no	2.015	3.580
mag9r051_sc3g9_c	no	0.2	1.765
mag9r061_sc3g9_c	no	1.14	2.705
mag9r111_sc3g9_c	no	-0.509	1.056
mag9r191_sc3g9_c	no	-0.82	0.745
mag9r261_sc3g9_c	no	2.342	3.907
mag9r10s_c	no	-1.096	0.469

mag9r14s_c	no	-1.276	0.289
mag9v011_sc3g9_c	yes	-1.042	0.523
mag9v012_sc3g9_c	yes	-0.288	1.277
mag9v091_sc3g9_c	no	0.142	1.707
mag9v121_sc3g9_c	no	1.187	2.752
mag9v131_sc3g9_c	no	0.018	1.583
mag9v13s_sc3g9_c	no	-1.087	0.478
mag9v081_c	no	1.115	2.680

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (section 4.2). Linked item difficulty parameters were derived by adding $c_{5\text{-}9}$ to the original item parameters.

6.3 Mathematics competence scores

In the SUF manifest mathematics competence scores are provided in the form of two different WLEs, “ma9_sc1” and “ma9_sc1u”, including their respective standard error, “ma9_sc2” and “ma9_sc2u”. For “ma9_sc1u”, person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for differences in the test position. In grade 7, the mathematics test was either presented as the first or the second test within the test battery, whereas in grade 9 the mathematics test was always presented second (after the break) within the test battery (see section 3.1). To correct for differences in the test position, we added half of the main effect related to the test position (see Schnittjer & Gerken, 2017) to the WLE scores of respondents that received the mathematics test in grade 7 before working on another test. As a result the WLE scores provided in “ma_sc1u” can be used for longitudinal comparisons between grades 5, 7 and 9. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “ma9_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 ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the mathematics test or who did not give enough valid responses, no WLE is 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

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

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

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14.*

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.

Duchhardt, C. (2015). *NEPS Technical Report for Mathematics – Scaling Results for the Additional Study Baden-Wuerttemberg* (NEPS Working Paper No. 59). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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: Waxmann.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Jordan, A.-K., & Duchhardt, C. (2013). *NEPS Technical Report for Mathematics—Scaling results of Starting Cohort 6–Adults* (NEPS Working Paper No. 32). Bamberg: University of Bamberg, National Educational Panel Study.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.

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.

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

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

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.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *Technical report of reading – Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Paedagogiske Institut.

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

Schnittjer, I. & Gerken, A.-L. (2017): *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 in Grade* 7 (NEPS Survey Paper No. 16). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011) Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & v. Maurice, J. (Eds.). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14 (pp. 67-86).* Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi:10.1177/014662168400800201

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

Appendix

Appendix A: ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort III - Ninth Grade

```
Title SC3 G9 MATHEMATICS: Partial Credit Model;


/* load data */
datafile [FILENAME].dat;

format pid 1-10 responses 12-45;
labels << [FILENAME].nam;

codes 0,1,2,3,4,5,6;
/* collapse response categories */

recode (0,1,2,3,4)        (0,0,1,2,3)         !item (3, 26, 27);
recode (0,1,2,3,4,5,6)    (0,0,0,0,1,2,3)     !item (6);
recode (0,1,2,3)          (0,1,2,3)           !item(33);

/* scoring */

score (0,1)              (0,1)                !item (1,2,4,5,7-25,28-32,34);
score (0,1,2,3)          (0,0.5,1,1.5)        !item (3,6,26,27,33);

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


/* model specification */
set constraint=none;
model item + item*step;

/* estimate model */
estimate! method=gauss, nodes=15, iterations=1000, convergence=0.0001;

/* save results to file */
show ! estimate=latent    >> show.txt;
show cases ! estimate=wle >> wle.txt;
```

Appendix B: Fixed Item Parameters cross-sectional

```
 1   -1.49887  /* item mag9d151_sc3g9_c */
 2    0.06801  /* item mag9d201_sc3g9_c */
 3   -1.92309  /* item mag9d05s_c */
 4   -1.67856  /* item mag9d061_c */
 5    0.02839  /* item mag9d111_c */
 6    0.79221  /* item mag9d09s_c */
 7    0.23623  /* item mag9d131_c */
 8    0.11305  /* item mag9q021_sc3g9_c */
 9   -0.53666  /* item mag9q071_sc3g9_c */
10    0.43206  /* item mag9q081_sc3g9_c */
11   -0.36526  /* item mag9q101_sc3g9_c */
12   -2.22958  /* item mag9q181_sc3g9_c */
13    0.03613  /* item mag9q211_sc3g9_c */
14    0.57702  /* item mag9q121_c */
15   -0.05112  /* item mag9q151_c */
16    0.40402  /* item mag9q161_c */
17   -1.64329  /* item mag9q021_c */
18    1.37547  /* item mag9q041_c */
19   -0.61902  /* item mag9q011_c */
20    2.01506  /* item mag9q031_c */
21    0.19961  /* item mag9r051_sc3g9_c */
22    1.14016  /* item mag9r061_sc3g9_c */
23   -0.50949  /* item mag9r111_sc3g9_c */
```

```
24     -0.82049  /* item mag9r191_sc3g9_c */
25      2.34211  /* item mag9r261_sc3g9_c */
26     -1.09609  /* item mag9r10s_c */
27     -1.27638  /* item mag9r14s_c */
28     -1.04235  /* item mag9v011_sc3g9_c */
29     -0.28779  /* item mag9v012_sc3g9_c */
30      0.14204  /* item mag9v091_sc3g9_c */
31      1.18738  /* item mag9v121_sc3g9_c */
32      0.01768  /* item mag9v131_sc3g9_c */
33     -1.08717  /* item mag9v13s_sc3g9_c */
34      1.11516  /* item mag9v081_c */
35     -0.11944  /* item mag9d05s_c step 1 */
36       0.5586  /* item mag9d05s_c step 2 */
37     -0.74302  /* item mag9d09s_c step 1 */
38     -0.34452  /* item mag9d09s_c step 2 */
39     -0.43528  /* item mag9r10s_c step 1 */
40     -0.61795  /* item mag9r14s_c step 1 */
41      0.20371  /* item mag9r14s_c step 2 */
42      0.02559  /* item mag9v13s_sc3g9_c step 1 */
43     -0.68072  /* item mag9v13s_sc3g9_c step 2 */
```

Appendix C: Fixed Item Parameters longitudinal (cross-sectional parameters + correction term)

```
  1      -0.705  /*  mag9d151_sc3g9_c  */
  2       0.862  /*  mag9d201_sc3g9_c  */
  3      -1.129  /*        mag9d05s_c  */
  4      -0.885  /*        mag9d061_c  */
  5       0.822  /*        mag9d111_c  */
  6       1.586  /*        mag9d09s_c  */
  7        1.03  /*        mag9d131_c  */
  8       0.907  /*  mag9q021_sc3g9_c  */
  9       0.257  /*  mag9q071_sc3g9_c  */
 10       1.226  /*  mag9q081_sc3g9_c  */
 11       0.429  /*  mag9q101_sc3g9_c  */
 12      -1.436  /*  mag9q181_sc3g9_c  */
 13        0.83  /*  mag9q211_sc3g9_c  */
 14       1.371  /*        mag9q121_c  */
 15       0.743  /*        mag9q151_c  */
 16       1.198  /*        mag9q161_c  */
 17      -0.849  /*        mag9q021_c  */
 18       2.169  /*        mag9q041_c  */
 19       0.175  /*        mag9q011_c  */
 20       2.809  /*        mag9q031_c  */
 21       0.994  /*  mag9r051_sc3g9_c  */
 22       1.934  /*  mag9r061_sc3g9_c  */
 23       0.285  /*  mag9r111_sc3g9_c  */
 24      -0.026  /*  mag9r191_sc3g9_c  */
```

```
25        3.136  /*  mag9r261_sc3g9_c  */
26       -0.302  /*         mag9r10s_c  */
27       -0.482  /*         mag9r14s_c  */
28       -0.248  /*  mag9v011_sc3g9_c  */
29        0.506  /*  mag9v012_sc3g9_c  */
30        0.936  /*  mag9v091_sc3g9_c  */
31        1.981  /*  mag9v121_sc3g9_c  */
32        0.812  /*  mag9v131_sc3g9_c  */
33       -0.293  /*  mag9v13s_sc3g9_c  */
34        1.909  /*         mag9v081_c  */
```

Appendix D: Residual correlations for the one-dimensional scaling of respective the difficult and medium, the easy and medium, and the difficult and easy booklet

	Item	Q3 DM	Q3 EM	Q3 DE
1	mag9d151_sc3g9_c	0.03	0.04	0.05
2	mag9d201_sc3g9_c	0.04	0.04	0.05
3	mag9d05s_c	0.05	0.08	0.05
4	mag9d061_c		0.07	0.06
5	mag9d111_c	0.04	0.06	0.05
6	mag9d09s_c	0.04	0.07	0.05
7	mag9d131_c	0.06	0.04	0.05
8	mag9q021_sc3g9_c	0.04	0.06	0.05
9	mag9q071_sc3g9_c		0.06	0.05
10	mag9q081_sc3g9_c	0.04	0.06	0.05
11	mag9q101_sc3g9_c	0.06	0.04	0.05
12	mag9q181_sc3g9_c	0.05	0.04	0.05
13	mag9q211_sc3g9_c	0.06		0.05
14	mag9q121_c	0.07		0.05
15	mag9q151_c		0.06	0.05
16	mag9q161_c	0.07	0.07	0.05
17	mag9q021_c	0.06	0.04	0.06
18	mag9q041_c	0.04	0.05	0.05
19	mag9q011_c	0.06	0.04	0.05
20	mag9q031_c	0.06		0.04
21	mag9r051_sc3g9_c	0.05	0.06	0.05
22	mag9r061_sc3g9_c	0.04	0.04	0.05
23	mag9r111_sc3g9_c	0.04	0.04	0.05
24	mag9r191_sc3g9_c		0.07	0.05
25	mag9r261_sc3g9_c	0.03	0.04	0.04
26	mag9r10s_c	0.06	0.04	0.05
27	mag9r14s_c	0.06	0.07	0.05
28	mag9v011_sc3g9_c	0.04	0.04	0.05
29	mag9v012_sc3g9_c	0.04	0.04	0.05
30	mag9v091_sc3g9_c	0.04	0.04	0.05
31	mag9v121_sc3g9_c	0.04	0.06	0.05
32	mag9v131_sc3g9_c	0.07	0.05	0.06
33	mag9v13s_sc3g9_c	0.06	0.05	0.07
34	mag9v081_c	0.07		0.05

Q_3 =Average absolute residual correlation for item (Yen, 1983), DM= difficult vs. medium, EM = easy vs. medium, DE= difficult vs. easy.

Appendix E: Content Areas of Items in the Mathematics Test Grade 9

	Item	Content area
1	mag9d151_sc3g9_c	Data and chance
2	mag9d201_sc3g9_c	Data and chance
3	mag9d05s_c	Data and chance
4	mag9d061_c	Data and chance
5	mag9d111_c	Data and chance
6	mag9d09s_c	Data and chance
7	mag9d131_c	Data and chance
8	mag9q021_sc3g9_c	Quantity
9	mag9q071_sc3g9_c	Quantity
10	mag9q081_sc3g9_c	Quantity
11	mag9q101_sc3g9_c	Quantity
12	mag9q181_sc3g9_c	Quantity
13	mag9q211_sc3g9_c	Quantity
14	mag9q121_c	Quantity
15	mag9q151_c	Quantity
16	mag9q161_c	Quantity
17	mag9q021_c	Quantity
18	mag9q041_c	Quantity
19	mag9q011_c	Quantity
20	mag9q031_c	Quantity
21	mag9r051_sc3g9_c	Space and shape
22	mag9r061_sc3g9_c	Space and shape
23	mag9r111_sc3g9_c	Space and shape
24	mag9r191_sc3g9_c	Space and shape
25	mag9r261_sc3g9_c	Space and shape
26	mag9r10s_c	Space and shape
27	mag9r14s_c	Space and shape
28	mag9v011_sc3g9_c	Change and relationships
29	mag9v012_sc3g9_c	Change and relationships
30	mag9v091_sc3g9_c	Change and relationships
31	mag9v121_sc3g9_c	Change and relationships
32	mag9v131_sc3g9_c	Change and relationships
33	mag9v13s_sc3g9_c	Change and relationships
34	mag9v081_c	Change and relationships