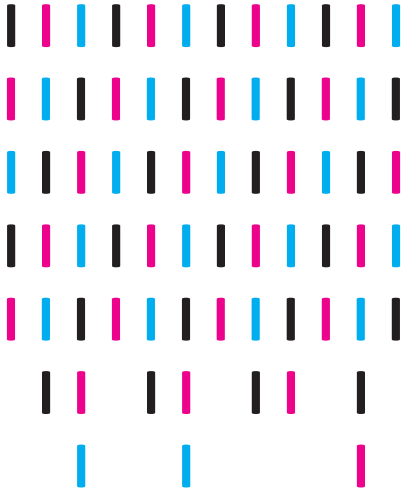

NEPS *SURVEY PAPERS*

Timo Gnambs

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORTS 4 (WAVE 10), 5 (WAVE 12), AND 6 (WAVE 9)

NEPS *Survey Paper* No. 72
Bamberg, September 2020

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 Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2020). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)* (NEPS *Survey Paper* No. 72). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP72:1.0

Acknowledgements:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Schnittjer & Gerken, 2017) to facilitate the understanding of the presented results.

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. In order to evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for a mathematical competence test that was administered in wave 10 of Starting Cohort 4 (ninth grade), wave 12 of Starting Cohort 5 (students), and wave 9 of Starting Cohort 6 (adults). The mathematical competence test included 64 items with multiple choice and open response formats that were administered in a computerized multi-stage test design. The adaptive nature of the test design allowed the administration of different items to each respondent tailored to the individual competence level. The test was administered to a total of 15,925 individuals (54% women) from Starting Cohort 4 (*N* = 6,909), Starting Cohort 5 (*N* = 4,643), and Starting Cohort 6 (*N* = 4,373). In Starting Cohort 5 about half of the respondents received the test in a proctored setting at their private homes (*N* = 2,742), whereas the remaining participants (*N* = 1,901) worked on unproctored, web-based tests. Starting Cohorts 4 and 6 were limited to proctored computerized testing. The responses of the participants were scaled using a partial credit model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and a satisfactory fit to the item response model. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test pertained to the routing of respondents in the multi-stage design that assigned few participants to the most difficult stages. Moreover, many respondents were unable to finish the test in the available testing time. Overall, the mathematics test had acceptable psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

item response theory, scaling, mathematics, multi-stage test, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2019).

Most of the competence data are scaled using models based on item response theory (IRT). Because these 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 a mathematical competence test that was administered in wave 10 of Starting Cohort 4 (ninth grade), wave 12 of Starting Cohort 5 (students), and wave 9 of Starting Cohort 6 (adults). First, the main concepts of the mathematics test and the test design are introduced. Then, the competence data of the three starting cohorts 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 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 cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, no fundamental changes in the presented results are expected.

2 Testing Mathematical Competence

2.1 Conceptual Framework

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

In the test, the items are not arranged in units. Rather, respondents usually received a description of a situation followed by a single task (or sometimes two tasks) related to this situation. Each of the items belonged to one of the following content areas:

- units and measuring (17 items),
- space and shape (18 items),
- change and relationships (16 items),
- data and chance (13 items).

Each item was constructed in such a way as 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 are distributed across the items.

2.2 Test Design

The mathematics test administered in the present study adopted a multi-stage design that included a total of 64 items (see Figure 1). However, each respondent received only a subset of 21 items depending on his or her estimated mathematical competence. The multi-stage test (MST) included five distinct stages that were sequentially administered to each respondent (Gnambs & Carstensen, in prep). Each stage consisted of different levels that included items of different difficulty. Lower levels (e.g., Levels 1 and 2) included easier items, whereas higher levels (e.g., Levels 4 and 5) included more difficult items. The difficulty of each item was estimated from previously published competence data in different starting cohorts and unpublished data from various developmental studies. To make sure that the MST respected the theoretical testing framework (see above), the different content areas were approximately uniformly distributed across the different levels within each stage. Thus, the assignment of the items to the different stages and levels was based on the respective item difficulties and the five content areas. To ensure a common scale for all test takers, we linked the different levels within each stage by assigning some items to two adjacent levels. For example, in Stage 2 item `maa9v193_c` was assigned to Level 3 and to Level 4. In this case, item `maa9v193_c` acts as link between the two levels and allows for the estimation of a common score for respondents receiving different levels. For items that were administered in a previous wave of Starting Cohort 4, 5, or 6 we made sure that these items were administered at the same position to prevent item position effects when using these items to link competence scores across waves (see section 7.2). There was no multi-matrix design regarding the order of the items *within* a specific level. All respondents assigned to a specific level received the test items in the same order.

Five different stages were sequentially administered to the respondents. Whereas Stage 1 and Stage 2 included one and three levels, respectively, the other stages had five levels each. Because Stage 1 included only one level including five items of medium difficulty, all respondents received the same items at the beginning of the test. Depending on their answers on these initial items, each respondent received one of three levels (including four items each) from Stage 2. Respondents that solved at least four of the five items from Stage 1 were assigned to Level 4 in Stage 2 that included more difficult items. In contrast, respondents that solved no more than one item from Stage 1 were assigned to Level 2 in Stage 2 that included easier items. Finally, respondents with a score of two or three in Stage 1 received items of medium difficulty in Stage 2 (i.e., Level 3). The number of items within each level of Stages 2 to 5 was four. The scoring rules for the assignment of respondents to the different levels of each stage (see Figure 1) were derived from simulation studies based on the expected ability distributions in the three starting cohorts.

To evaluate the quality of the administered items, extensive preliminary analyses were conducted. These showed that the MST routed rather few respondents to the most difficult levels. Consequently, some items were assigned to rather few respondents. Because small samples can yield unstable parameter estimates, items that were answered by less than 200 respondents were excluded from further analyses (cf. Pohl & Carstensen, 2012). In total 12 of the 64 items had to be excluded from the final scaling procedure (see Figure 1). Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 52 items.

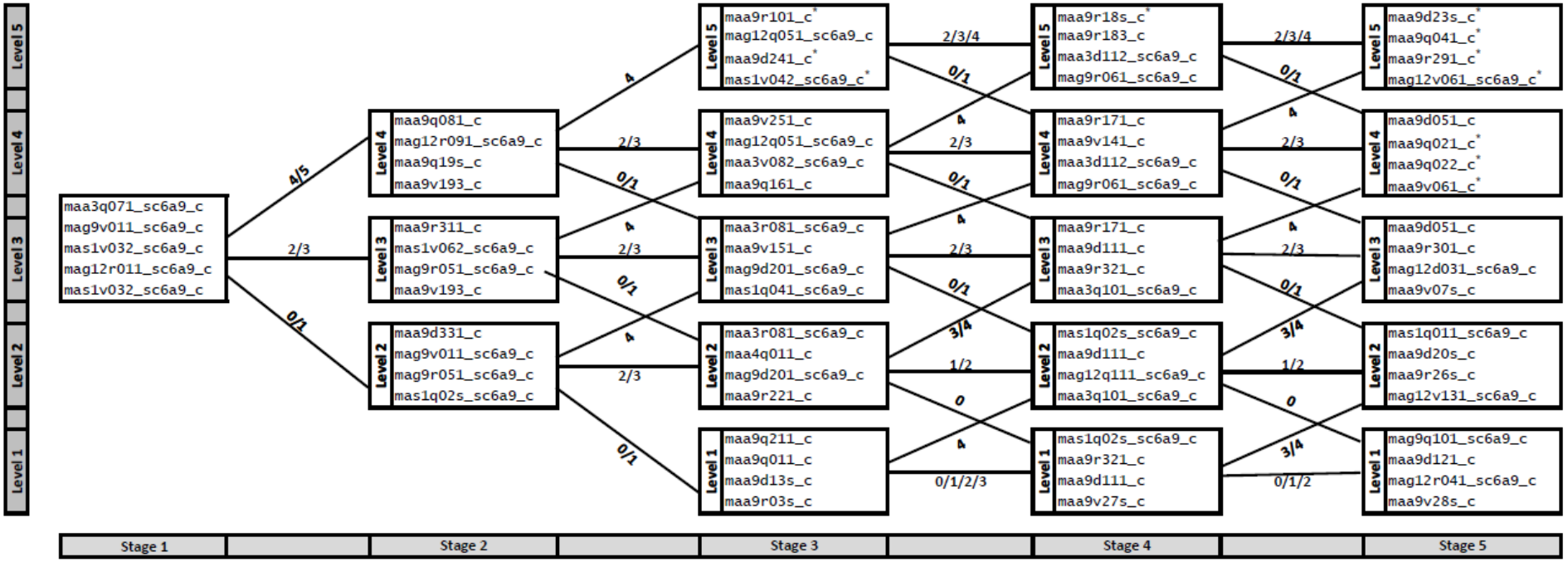

Figure 1. Multi-Stage Test Design with Items and Scoring Rules

Note. Item names correspond to those from Starting Cohort 6 (adults). An equivalency table for the variable names across the three starting cohorts is given in Appendix B.

The test items were accompanied by different response formats (see Table 1). The most common response format were short constructed responses (SCR) that required test takers to give mostly single-word answers, such as a number. All SCR items were scored dichotomously. Simple multiple choice formats included three to four response options with one being correct and three response options functioning as distractors (i.e., they were incorrect). Finally, complex multiple choice items included a number of subtasks with two response options that were subsequently combined into a single polytomous variable (see section 4.2). Because preliminary analyses identified a poor fit for one subtask of item `maa9r03s_c`, this subtask was excluded from further analyses. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt, and Weinert (2012).

Table 1. Number of Items by Different Response Formats

Response format	**Number of Items**
Short constructed responses	8
Simple multiple choice items	34
Complex multiple choice items	10
Total number of items	52

In Starting Cohorts 4 and 6, the MST was administered as a computer-based test (CBT). The test administrators visited the respondents at their private homes and presented the MST on a laptop. Thus, the respondents were administered the MST in a proctored setting. In Starting Cohort 5, an experimental design was implemented. About half of the respondent received the MST as CBT (identical to the other starting cohorts), whereas the remaining respondents were administered a web-based test (WBT). These respondents finished the MST as an unproctored web-based test.

The study assessed different competence domains including, among others, mathematical competence, reading competence, and English as a foreign language. The competence tests for these domains were always presented first within the test battery. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the mathematics test was either administered as the first or the second test (i.e., after the reading or English test). A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

3 Data

A total of 15,925[1] students (54% women) had at least three valid responses on the mathematical competence test and, thus, were used for the psychometric analyses (cf. Pohl & Carstensen, 2012). Of these, *N* = 6,909 (50% women) were from Starting Cohort 4, *N* = 4,643 (60% women) were from Starting Cohort 5, and *N* = 4,373 (52% women) were from Starting Cohort 6. The age of the respondents ranged from 18 to 73 years. Basic sociodemographic information of the different samples is summarized in Table 2.

Table 2. Sample Descriptions

	Starting Cohort 4	**Starting Cohort 5 (CBT)**	**Starting Cohort 5 (WBT)**	**Starting Cohort 6**
Sample size	6,909	2,742	1,901	4,373
Women	50%	61%	60%	52%
Migration background	13%	8%	8%	8%
Mean age (*SD*)	21.09 (0.59)	27.79 (3.10)	27.99 (3.30)	53.01 (10.37)

Note. CBT = Computer-based test, WBT = Web-based test

4 Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and, finally, e) multiple kinds of missing responses within complex multiple choice items that are not determined. Invalid responses occurred, for example, when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits or lack of motivation, not all persons finished the test. All missing responses after the last valid response within a level of the current stage were coded as not reached, whereas the remaining items in the following stages were coded as missing due to test abortion. Because of the MST design, most available items were not administered to any given participant. These items were missing by design. Because complex multiple choice items were aggregated from several subtasks, different kinds of missing responses or a

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

mixture of valid and missing responses might be found in these items. A complex multiple choice item was coded as missing if at least one subtask contained a missing response. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a 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). 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.

4.2 Scaling Model

Item and person parameters were estimated using a partial credit model (PCM; Masters, 1982) with Gauss-Hermite quadrature (21 nodes). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

Complex multiple choice items consisted of a set of subtasks that were aggregated to a polytomous variable for each item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the partial credit item was scored as missing. Response categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category.

Mathematical competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple multiple choice items and short constructed responses 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). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 3.

4.3 Checking the Quality of the Test

The mathematical competence test was specifically constructed for administration in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

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

After aggregating the subtasks to polytomous variables, the fit of the dichotomous and polytomous items to the partial credit model (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective *t*-value, and the item characteristic curves (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. 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 sex, age, the number of books at home (as a proxy for socioeconomic status), migration background (see Pohl & Carstensen, 2012, for a description of these variables), starting cohort (4, 5, or 6), and assessment mode (CBT versus WBT). Differential item functioning (DIF) was examined using a multigroup item response model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Minimum hypothesis tests (see Fischer, Rohm, Gnambs, & Carstensen, 2016) were used to statistically test whether the observed differences were significantly larger than 0.4 and, thus, were at least small in size. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The mathematics test was scaled using the PCM (Masters, 1982) 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 might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations as indicated by Yen's (1984) Q_3 indicate unidimensionality. 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. Moreover, we examined a multidimensional model based on the five content areas (see Appendix A) using quasi-Monte Carlo integration (with 2,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

4.4 Software

The item response models were estimated with the *TAM* package version 2.12-18 (Robitzsch, Kiefer, & Wu, 2018) in *R* version 3.5.0 (R Core Team, 2018).

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Invalid responses were extremely rare: Less than 0.20% of participants had one invalid response; the remaining respondents did not have a single invalid response. Similar, not determinable missing values (e.g., for different subtasks of complex multiple choice items) were negligible. Most respondents had no missing values of this type, whereas less than 0.50% of the respondents exhibited a single not determinable missing value.

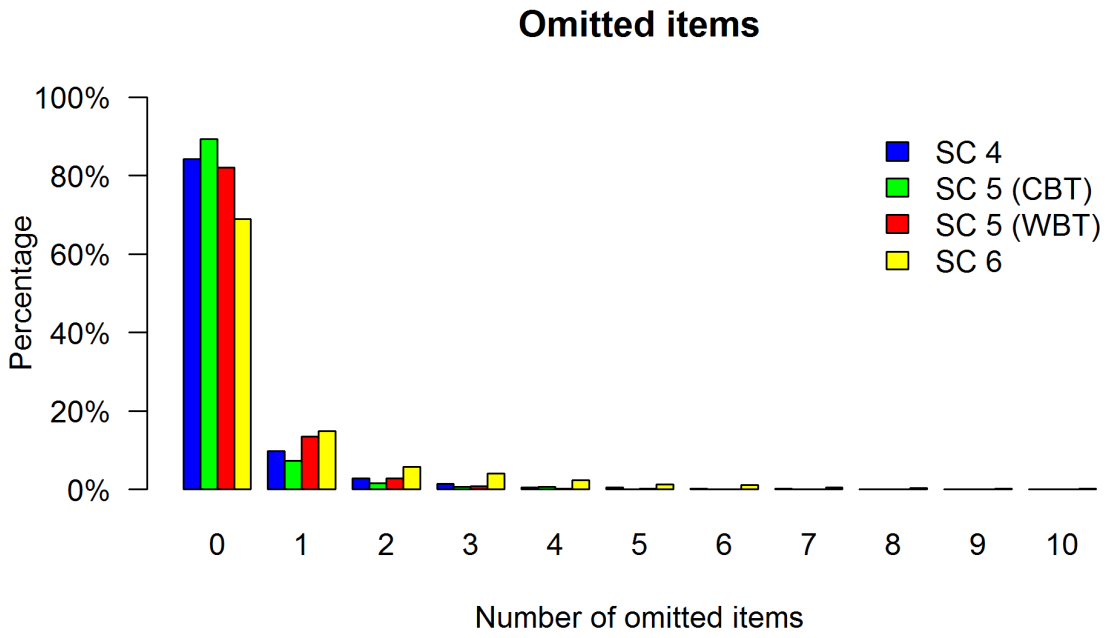

Figure 2. Number of omitted items by sample

Missing responses can also occur when respondents omit items. As illustrated in Figure 2 most respondents (about 81%) did not skip any item, only about 11% skipped one item. Participants with multiple omitted items were rare (about 8% of the sample). However, there were some differences between the three starting cohorts. The adult sample (Starting Cohort 6) exhibited substantially more omitted items; only about 69% of respondents in Starting Cohort 6 had no omitted item, whereas 15% exhibited a single omitted item. Similar, the two assessment modes in Starting Cohort 5 showed different omission rates. In the proctored CBT condition fewer people omitted at least one item (about 11%) as compared to the unproctored WBT condition (about 18%).

Another source of missing responses is items that were not reached by the respondents because they aborted the test, for example, because the time limit was reached or a lack of motivation. These missing values refer to items after the last valid response. As illustrated in Figure 3, only about 23% of the respondents did not abort the test and were administered all 21 items. About half of the sample received at least 16 items and 10 or more items were reached by 80% of the sample. This indicates that the test was too difficult for the limited testing time. The results given in Figure 3 show that this was the case for all three starting cohorts. Again, different missing rates were observed for the two assessment modes in Starting Cohort 5. In the proctored CBT condition about 13% of the respondents received 21 items, whereas this ratio was 26% in the unproctored condition (WBT).

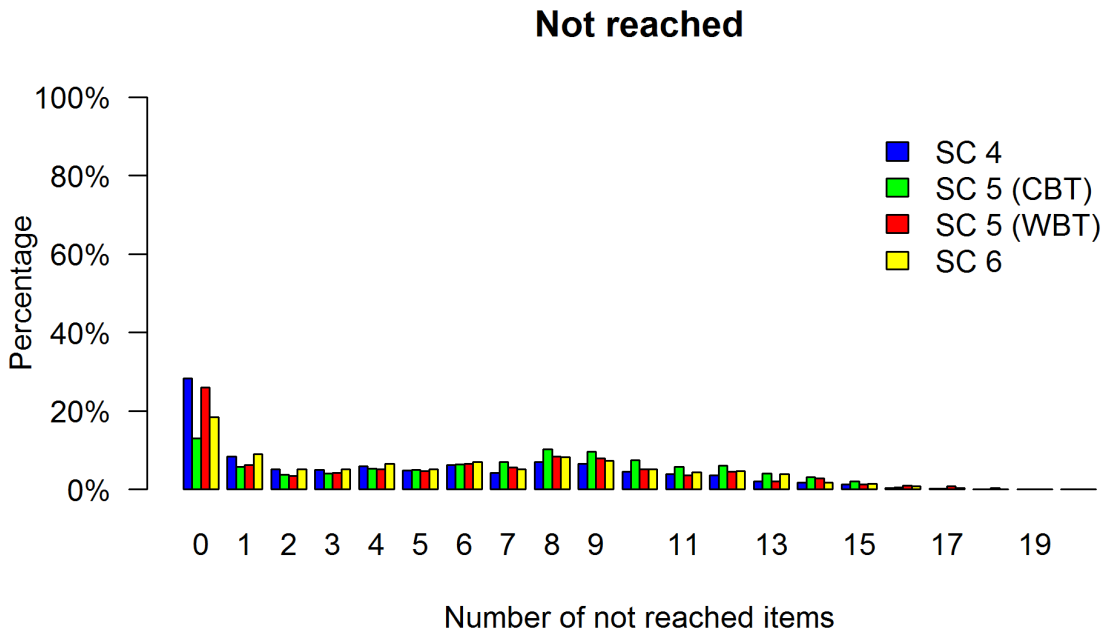

Figure 3. Number of not reached items by sample

With an item's progressing position in the test, the amount of persons that did not reach an item rose to about 77%. In all samples, the last items were reached by only few respondents. As illustrated in Figure 4, in the proctored CBT condition of Starting Cohort 5 substantially more persons did not reach the last item of the test (about 87%) as compared to the unproctored WBT condition (about 74%), Starting Cohort 4 (about 72%), or Starting Cohort 6 (about 82%). Thus, it seems that many respondents were unable to finish the MST within the allocated time span. This indicates that the testing time might have been too short for the difficulty of the administered test.

The total number of missing responses, aggregated over invalid, omitted, not reached, and not determinable missing responses per person, is illustrated in Figure 5. Because the majority of the sample did not reach the end of the test, there was a substantial number of missing values. The median number of missing responses was 5; only about 18% of the respondents had no missing response at all. Again, in Starting Cohort 6 fewer participants had no missing value (about 12%) as compared to Starting Cohort 4 (about 22%) and the

unproctored WBT sample in Starting Cohort 5 exhibited a larger percentage of respondents with no missing value (about 23%) as compared to the proctored CBT sample (about 10%).

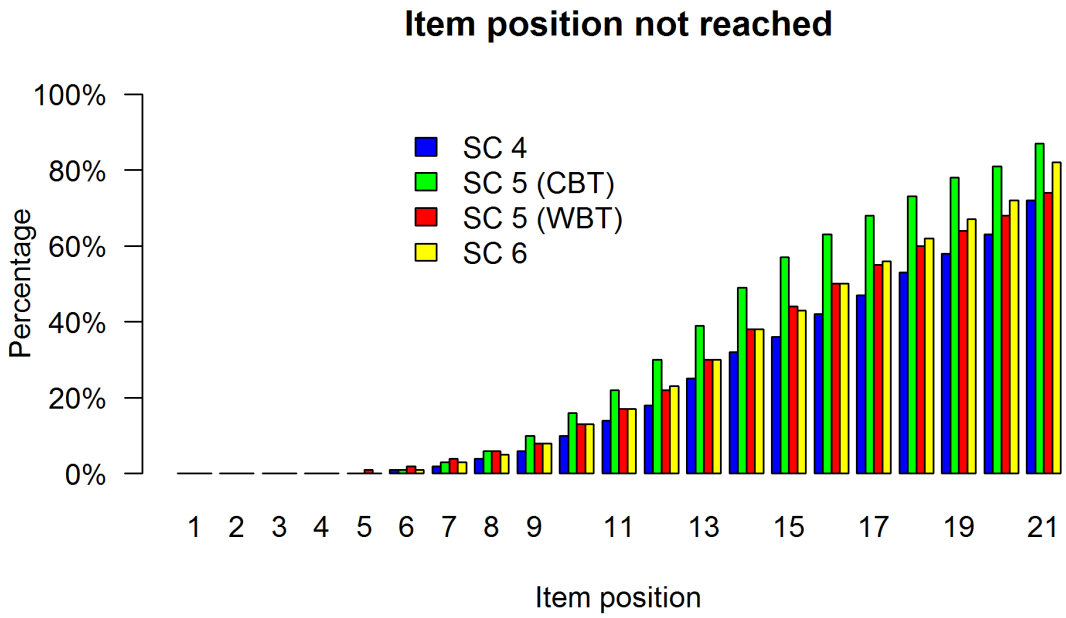

Figure 4. Item position not reached by sample

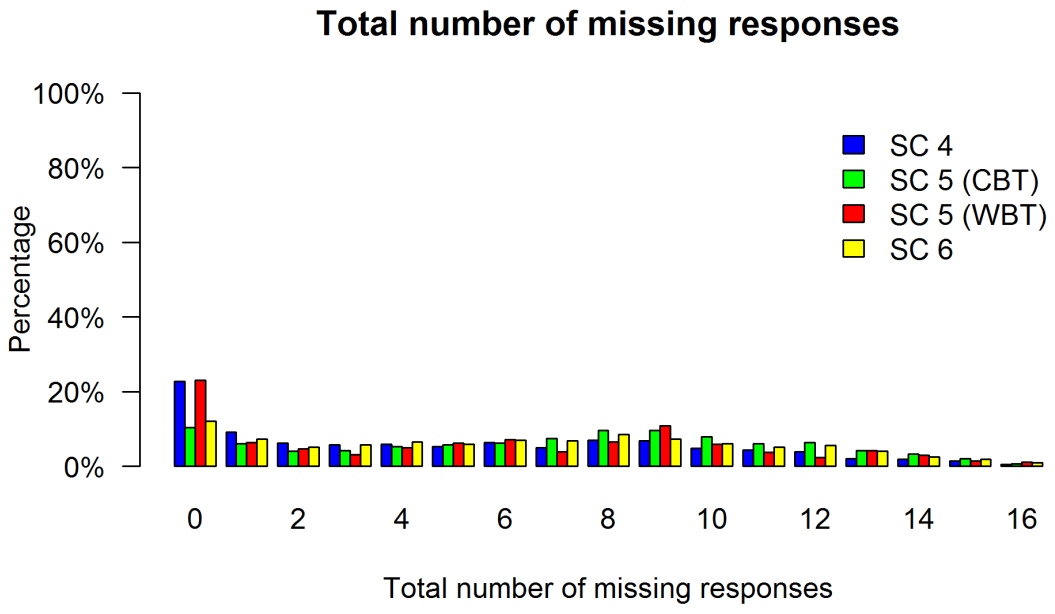

Figure 5. Total number of missing responses by sample

In sum, the amount of missing responses was rather large because many respondents did not reach the end of the test. On average, respondents from Starting Cohort 4 or the web-based condition of Starting Cohort 5 exhibited more valid responses than respondents in Starting Cohort 6 or in the proctored condition of Starting Cohort 5.

5.1.2 Missing responses per item

Table 3 provides information on the occurrence of different kinds of missing responses per item for the four samples. The number of omitted responses varied across items between 0.00% and 11.92% (*Mdn* = 3.02%) and were, thus, negligible. In contrast, there were substantially more missing responses because participants did not reach the item. On average, the items had *Mdn* = 17.99% missing values of this type. Particularly, items in the last stage of the MST were frequently not reached. The respective distributions of missing values for each starting cohort and assessment mode are summarized in Appendix C.

Table 3. Percentage of Missing Values by Item

Pos.	Item	*N*	N_v	OM	NR
1	maa3q071_sc6a9_c	15925	15868	0.36	0.00
2	mag12v101_sc6a9_c	15925	15577	2.19	0.00
3	mag12v122_sc6a9_c	15925	15656	1.69	0.00
4	mag12r011_sc6a9_c	15925	15781	0.82	0.09
5	mas1v032_sc6a9_c	15925	15039	5.11	0.45
6	maa9q081_c	4675	4615	0.09	1.20
6	maa9r311_c	7502	7349	1.57	0.47
6	maa9d331_c	3671	3600	1.58	0.11
7	mag9v011_sc6a9_c	3671	3613	1.17	0.41
7	mag12r091_sc6a9_c	4675	4416	1.16	4.39
7	mas1v062_sc6a9_c	7498	7317	0.95	1.47
8	mag9r051_sc6a9_c	11167	10846	0.77	2.10
8	maa9q19s_c	4672	4201	0.90	9.16
9	maa9d09s_c	3671	3453	3.51	2.23
9	maa9v193_c	12148	10622	3.70	8.45
10	maa3r081_sc6a9_c	9543	8852	3.47	3.77
10	maa9q211_c	1163	1118	3.18	0.69
10	maa9v251_c	2959	2781	1.12	4.90
11	maa9q011_c	5768	5387	1.77	4.84
11	mag12q051_sc6a9_c	3944	3385	0.28	13.89
11	maa9v151_c	4935	4165	5.27	10.33
12	maa3v082_sc6a9_c	2946	2189	0.44	25.25
12	maa9d13s_c	1162	975	11.27	4.39
12	mag9d201_sc6a9_c	9525	8406	0.25	11.50
13	maa9r03s_c	1162	984	8.09	7.06
13	maa9q161_c	2938	1956	0.03	33.39

Pos.	Item	*N*	N_v	OM	NR
13	mas1q041_sc6a9_c	4921	3541	2.46	25.58
13	maa9r221_c	4602	3823	6.04	10.89
14	mas1q02s_sc6a9_c	4626	3827	10.77	6.07
14	maa9r171_c	5881	4740	6.80	11.22
15	maa9v141_c	1715	1254	2.92	23.97
15	mas1d081_sc6a9_c	7424	6113	1.82	15.38
15	maa3r121_sc6a9_c	1385	1292	1.59	5.13
16	maa9d111_c	1384	1059	11.92	11.42
16	mag12q111_sc6a9_c	3238	2590	0.22	19.80
16	maa3d112_sc6a9_c	2386	1245	0.63	47.19
16	maa9r321_c	4184	3016	0.53	27.39
17	mag9r061_sc6a9_c	2376	980	0.80	57.79
17	maa9v27s_c	1384	1065	7.30	15.75
17	maa3q101_sc6a9_c	7417	5117	0.57	30.44
18	maa9d051_c	3045	2480	1.74	15.27
18	mas1q011_sc6a9_c	2557	2325	1.68	7.39
18	mag9q101_sc6a9_c	1404	1219	3.42	9.76
19	maa9d121_c	1404	1181	1.64	14.25
19	maa9d20s_c	2553	1886	5.37	20.64
19	maa9r301_c	2221	1736	0.05	21.79
20	mag12d031_sc6a9_c	2215	1472	0.36	33.18
20	mag12r041_sc6a9_c	1403	1076	0.78	22.52
20	maa9r26s_c	2549	1643	3.65	31.58
21	maa9v07s_c	2212	878	0.09	59.99
21	maa9v28s_c	1401	737	3.78	41.90
21	mag12v131_sc6a9_c	2543	1551	0.04	38.97

Note. Pos. = Item position within test. *N* = Number of respondents the item was administered to, N_v = Number of valid responses, NR = Percentage of respondents that did not reach item within a level plus percentage of respondents that aborted the test in a previous stage, OM = Percentage of respondents that omitted the item.
Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix B.

5.2 Parameter Estimates

To avoid potentially biased parameter estimates resulting from mode effects (unproctored versus proctored settings), the following analyses are limited to the proctored CBT samples. Thus, the unproctored WBT sample from Starting Cohort 5 was excluded from the scaling

procedure. Information on the measurement invariance across assessment modes is given in section 5.3.2. Moreover, preliminary analyses identified rather low discriminations for the dichotomous items `maa9r301_c` and `mas1v032_sc6a9_c`. Therefore, a 0.5 point scoring was used for these items in the scaling procedure (similar to the scoring rule for PCM items, see section 4.2).

5.2.1 Item parameters

The fifth column in Table 4 presents the percentage of correct responses (for simple multiple choice items) in relation to all valid responses for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty. The percentage of correct responses varied between 37% and 86% with an average of 57% (*SD* = 11%) correct responses.

Table 4. Item Parameters for CBT samples

Item	Pos.	Stage	Item format	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	Discr.	aQ3
maa3q071_sc6a9_c	1	1	MC	69.08	-0.87	0.02	1.01	0.80	1.03	0.03
mag12v101_sc6a9_c	2	1	MC	60.86	-0.43	0.02	0.97	-4.50	1.27	0.03
mag12v122_sc6a9_c	3	1	MC	53.10	-0.05	0.02	1.07	10.12	0.75	0.02
mag12r011_sc6a9_c	4	1	MC	47.19	0.24	0.02	1.01	1.55	1.04	0.02
mas1v032_sc6a9_c	5	1	MC	37.74	0.62	0.02	1.01	2.15	0.45	0.03
maa9q081_c	6	2	MC	69.69	-0.00	0.04	0.99	-0.58	1.20	0.12
maa9r311_c	6	2	SCR	61.78	-0.51	0.03	0.97	-3.61	1.41	0.07
maa9d331_c	6	2	SCR	85.86	-2.93	0.05	0.97	-0.89	1.33	0.07
mag9v011_sc6a9_c	7	2	MC	62.83	-1.51	0.04	0.95	-3.92	1.69	0.07
mag12r091_sc6a9_c	7	2	MC	54.94	0.77	0.04	1.01	1.09	1.01	0.11
mas1v062_sc6a9_c	7	2	MC	42.89	0.36	0.03	1.03	3.40	0.81	0.05
mag9r051_sc6a9_c	8	2	MC	55.86	-0.56	0.02	0.97	-4.39	1.32	0.06
maa9q19s_c	8	2	PCM	73.29	-0.13	0.04	0.98	-1.02	1.29	0.08
maa9d09s_c	9	2	PCM	NA	-0.32	0.02	1.01	0.24	0.53	0.07
maa9v193_c	9	2	SCR	48.47	0.48	0.02	1.05	6.56	0.76	0.04
maa3r081_sc6a9_c	10	3	MC	53.74	-0.37	0.02	0.97	-3.35	1.32	0.06
maa9q211_c	10	3	MC	60.11	-1.93	0.07	0.95	-2.23	1.69	0.20
maa9v251_c	10	3	MC	56.85	0.62	0.04	0.97	-2.19	1.57	0.08
maa9q011_c	11	3	MC	72.55	-1.90	0.03	0.96	-2.37	1.45	0.04
mag12q051_sc6a9_c	11	3	MC	46.91	1.22	0.04	0.98	-1.30	1.22	0.07
maa9v151_c	11	3	MC	37.91	0.76	0.04	0.98	-1.39	1.37	0.07
maa3v082_sc6a9_c	12	3	MC	71.90	-0.13	0.06	0.98	-0.67	1.35	0.07
maa9d13s_c	12	3	PCM	NA	-1.16	0.04	0.98	-0.63	0.79	0.21
mag9d201_sc6a9_c	12	3	MC	65.23	-0.91	0.03	0.94	-6.12	1.68	0.06
maa9r03s_c	13	3	PCM	NA	-1.01	0.04	0.93	-2.42	1.07	0.21
maa9q161_c	13	3	MC	55.83	0.70	0.05	0.98	-1.38	1.43	0.09

Item	Pos.	Stage	Item format	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	Discr.	aQ3
mas1q041_sc6a9_c	13	3	MC	57.72	-0.11	0.04	1.00	0.02	1.07	0.08
maa9r221_c	13	3	MC	36.83	-0.01	0.04	1.01	0.82	0.96	0.05
mas1q02s_sc6a9_c	14	4	PCM	NA	-0.53	0.02	1.04	2.09	0.38	0.03
maa9r171_c	14	4	SCR	56.58	0.08	0.03	0.98	-1.89	1.27	0.04
maa9v141_c	15	4	MC	67.22	0.21	0.07	1.00	0.20	1.05	0.10
mas1d081_sc6a9_c	15	4	SCR	53.53	-0.41	0.03	0.99	-1.04	1.24	0.05
maa3r121_sc6a9_c	15	4	MC	51.55	-1.56	0.06	1.05	2.56	0.71	0.04
maa9d111_c	16	4	SCR	52.41	-1.54	0.07	0.96	-1.97	1.64	0.08
mag12q111_sc6a9_c	16	4	MC	54.05	-0.87	0.04	1.01	0.85	1.00	0.08
maa3d112_sc6a9_c	16	4	MC	44.82	1.31	0.07	0.97	-1.42	1.68	0.09
maa9r321_c	16	4	MC	47.78	0.19	0.04	1.01	0.58	0.96	0.04
mag9r061_sc6a9_c	17	4	SCR	66.02	0.37	0.08	1.01	0.39	0.92	0.13
maa9v27s_c	17	4	PCM	NA	-0.86	0.04	1.02	0.55	0.44	0.05
maa3q101_sc6a9_c	17	4	MC	49.99	-0.28	0.03	1.03	2.99	0.86	0.08
maa9d051_c	18	5	SCR	56.53	0.10	0.05	1.01	0.54	0.96	0.11
mas1q011_sc6a9_c	18	5	MC	52.43	-0.72	0.04	1.00	0.17	1.09	0.06
mag9q101_sc6a9_c	18	5	MC	59.06	-1.89	0.06	0.94	-2.96	1.78	0.12
maa9d121_c	19	5	MC	76.29	-2.76	0.07	0.99	-0.23	1.17	0.12
maa9d20s_c	19	5	PCM	NA	-0.32	0.02	1.05	1.98	0.20	0.04
maa9r301_c	19	5	MC	66.07	-0.57	0.06	1.01	0.40	0.35	0.12
mag12d031_sc6a9_c	20	5	MC	56.18	-0.09	0.06	0.95	-3.04	2.35	0.14
mag12r041_sc6a9_c	20	5	MC	39.78	-1.04	0.07	1.03	1.57	0.69	0.09
maa9r26s_c	20	5	PCM	NA	-0.21	0.03	1.01	0.49	0.42	0.05
maa9v07s_c	21	5	PCM	57.29	-0.21	0.08	0.97	-1.15	1.64	0.13
maa9v28s_c	21	5	PCM	64.86	-2.21	0.08	1.04	1.35	0.63	0.11
mag12v131_sc6a9_c	21	5	MC	42.81	-0.37	0.06	1.02	1.23	0.78	0.06

Note. Pos. = Item position, Stage = Stage of MST, Format = Response format (MC = Multiple Choice, SCR = Short constructed response, PC = Partial Credit), Difficulty = Item difficulty / location, SE = Standard error of item difficulty / location, WMNSQ = Weighted mean square, t = t-value for WMNSQ, 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 item scores and, therefore, are not reported. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix B.

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 4. The step parameters for polytomous variables are summarized in Table 5. The item difficulties and location parameters were estimated by constraining the mean of the ability distribution to be zero. Due to the large sample size, the standard errors (SE) of the estimated parameters (see Tables 5 and 6) were rather small (all SEs ≤ 0.08). The estimated item difficulties and location parameters ranged from -2.93 (item `maa9d331_c`) to 1.31 (item `maa3d112_sc6a9_c`). Thus, they covered a rather wide range including easy as well as difficult items.

Table 5. Step Parameters (with Standard Errors) for Polytomous Items in CBT samples

Item	Step 1	Step 2	Step 3	Step 4
maa9d09s_c	-1.19 (0.05)	-0.48 (0.04)	0.20 (0.04)	1.47
maa9d13s_c	0.26 (0.07)	-0.26		
maa9r03s_c	-0.12 (0.07)	0.12		
mas1q02s_sc6a9_c	-0.43 (0.03)	-0.33 (0.03)	0.76	
maa9v27s_c	-0.30 (0.06)	0.30		
maa9d20s_c	-0.43 (0.05)	-0.02 (0.05)	0.45	
maa9r26s_c	-0.53 (0.05)	-0.48 (0.05)	1.01	

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix B.

5.2.2 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. Because some items in the mathematics test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson, & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 6, the category thresholds of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of category thresholds. The respective thresholds ranged from -3.71 (item `maa4d09s_c`) to 2.76 (item `maa4d09s_c`) and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.16, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .75, WLE reliability = .69) was acceptable. The mean of the category threshold distribution was about 0.62 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

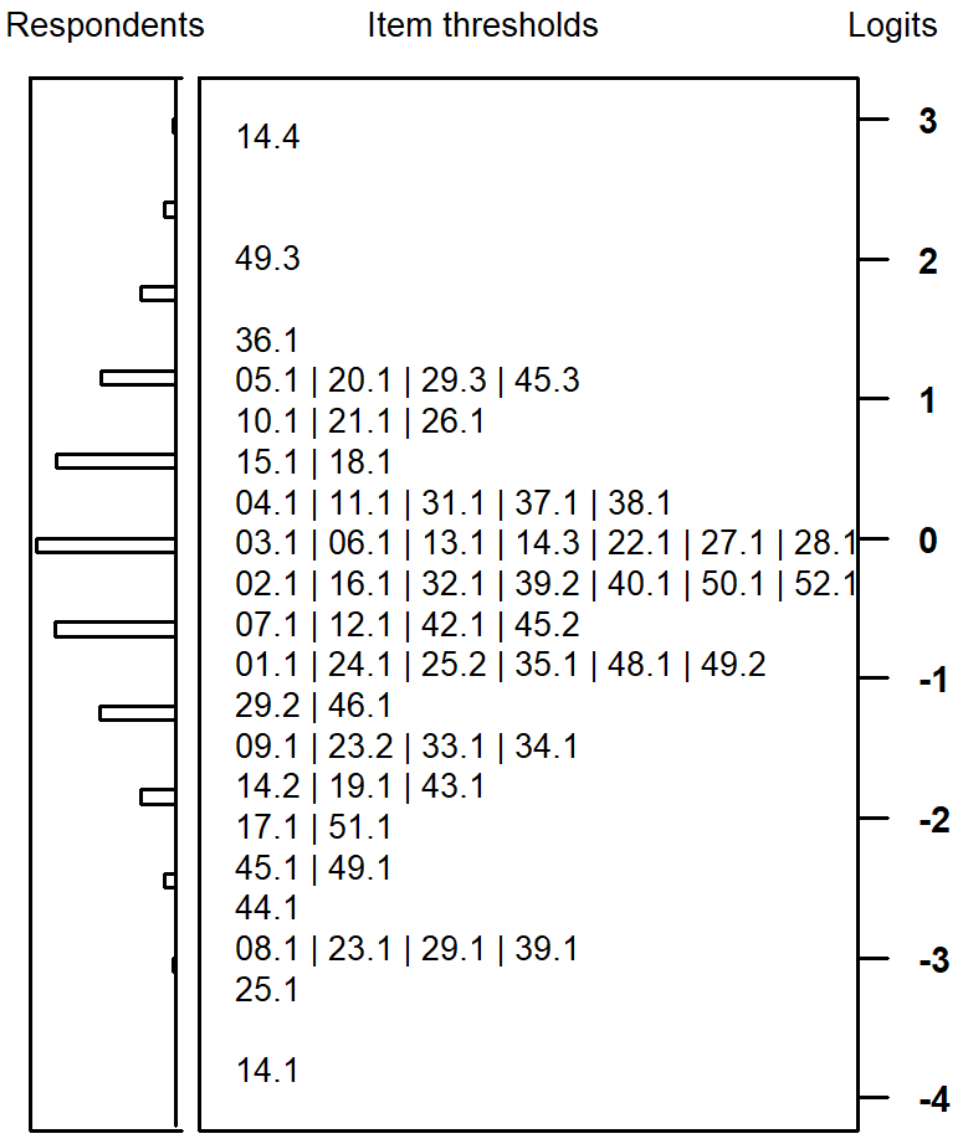

Figure 6. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item number in Table 4 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Item fit

The evaluation of the item fit was performed based on the final scaling model, the PCM. Again, the test quality was examined for the CBT samples only, while excluding the unproctored WBT sample from Starting Cohort 5. Altogether, item fit was good (see Table 5). No item exhibited a WMNSQ greater than 1.10. Moreover, a visual inspection of the item characteristic curves (ICC) showed no pronounced deviation from the expected ICC for the items. For the remaining items, values of the WMNSQ ranged from 0.93 (item `maa9r03s_c`) to 1.07 (item `mag12v122_sc6a9_c`). Only one item exhibited a *t*-value of the WMNSQ greater than 8 (item `mag12v122_sc6a9_c`). However, a visual inspection of the ICC showed no noticeable problems.

5.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables sex, the number of books at home (as a proxy for socioeconomic status), migration background, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, we examined DIF effects between the three CBT samples from Starting Cohorts 4, 5, and 6. Moreover, mode effects were studied by comparing the proctored CBT sample and the unproctored WBT sample in Starting Cohort 5. All analyses were limited to items with at least 100 valid responses for each response category in each group. Because of varying sample sizes in the different subgroups, the reported DIF analyses included different item sets. The differences between the estimated item difficulties in the various groups are summarized in Table 6. For example, the column “Male vs. female” reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed 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).

Sex: The sample included 6,617 men and 7,405 women. On average, male participants had a slightly higher estimated mathematical competence than females (main effect = 0.59 logits, Cohen’s d = 0.57). No item showed DIF greater than 0.50 logits (or a d greater than 0.45). An overall test for DIF (see Table 7) 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) favored the DIF model over the more parsimonious model including only the main effect. Similar results were obtained using the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models. However, the estimated main effects for sex were rather similar in both models (Cohen’s d = 0.45 versus 0.57). Thus, there was no pronounced DIF with regard to sex.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 7,386 test takers with less than 100 books at home and 8,537 test takers with 100 or more books at home. There were small average differences between the two groups. Participants with fewer books at home performed on average 0.33 logits (Cohen’s d = 0.31)

lower in mathematical competence than participants with more books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.58 for item `maa9q211_c`). As a consequence, also the overall test for DIF using the BIC favored the main effects model (Table 7).

Table 6. Differential Item Functioning

Item	Sex	Books	Migration	Position	Sample			Mode
	male vs. female	< 100 vs. ≥ 100	without vs. with	first vs. second	SC 4 vs. SC 6	SC 5 vs. SC 6	SC 4 vs. SC 5	CBT vs. WBT
maa3q071_sc6a9_c	0.12 (0.11)	-0.12 (-0.12)	0.12 (0.12)	0.25 (0.23)	-0.05 (-0.06)	-0.17 (-0.18)	0.11 (0.12)	0.27 (0.29)
mag12v101_sc6a9_c	-0.34 (-0.33)	-0.04 (-0.03)	0.06 (0.06)	0.11 (0.10)	0.21 (0.23)	-0.04 (-0.04)	0.24 (0.27)	-0.08 (-0.08)
mag12v122_sc6a9_c	0.17 (0.16)	-0.16 (-0.15)	0.21 (0.22)	0.12 (0.11)	-0.47 (-0.52)	-0.14 (-0.15)	-0.34 (-0.37)	0.11 (0.12)
mag12r011_sc6a9_c	-0.13 (-0.13)	0.07 (0.06)	-0.04 (-0.04)	0.11 (0.10)	-0.00 (-0.00)	0.03 (0.04)	-0.04 (-0.04)	0.05 (0.06)
mas1v032_sc6a9_c	0.37 (0.36)	-0.19 (-0.18)	0.21 (0.22)	-0.07 (-0.07)	-0.06 (-0.07)	0.31 (0.34)	-0.37 (-0.41)	0.16 (0.17)
maa9q081_c	-0.39 (-0.37)	-0.18 (-0.17)		0.07 (0.06)	0.28 (0.31)	0.28 (0.31)	-0.00 (-0.00)	-0.16 (-0.17)
maa9r311_c	-0.49 (-0.48)	0.04 (0.03)	0.15 (0.16)	0.03 (0.03)	-0.36 (-0.39)	-0.31 (-0.34)	-0.05 (-0.05)	-0.02 (-0.02)
maa9d331_c	0.39 (0.38)	0.36 (0.34)	-0.04 (-0.05)	0.16 (0.15)				
mag9v011_sc6a9_c	-0.19 (-0.18)	0.03 (0.03)	0.04 (0.05)	0.03 (0.03)				
mag12r091_sc6a9_c	-0.05 (-0.05)	-0.18 (-0.17)	0.11 (0.12)	0.07 (0.07)	0.19 (0.21)	0.30 (0.33)	-0.11 (-0.12)	0.11 (0.12)
mas1v062_sc6a9_c	-0.04 (-0.03)	0.06 (0.05)	0.01 (0.01)	-0.02 (-0.02)	0.34 (0.38)	0.19 (0.21)	0.15 (0.17)	-0.08 (-0.09)
mag9r051_sc6a9_c	0.24 (0.23)	0.34 (0.32)	-0.23 (-0.24)	0.01 (0.01)	-0.10 (-0.11)	-0.43 (-0.47)	0.33 (0.36)	-0.23 (-0.25)
maa9q19s_c	0.01 (0.01)	-0.20 (-0.19)		0.03 (0.03)	0.39 (0.42)	0.20 (0.22)	0.18 (0.20)	0.14 (0.15)
maa9d09s_c				-0.10 (-0.09)				
maa9v193_c	0.31 (0.30)	-0.07 (-0.07)	-0.10 (-0.10)	-0.04 (-0.04)	0.17 (0.19)	0.14 (0.15)	0.04 (0.04)	0.08 (0.09)
maa3r081_sc6a9_c	-0.28 (-0.27)	0.03 (0.02)	-0.04 (-0.04)	-0.03 (-0.03)	0.25 (0.28)	0.21 (0.23)	0.04 (0.04)	-0.06 (-0.07)
maa9q211_c	-0.02 (-0.02)	0.58 (0.55)	-0.10 (-0.10)	0.43 (0.40)				
maa9v251_c	-0.35 (-0.34)	0.04 (0.04)		-0.00 (-0.00)	0.27 (0.30)	0.24 (0.26)	0.04 (0.04)	-0.01 (-0.01)
maa9q011_c	0.26 (0.25)	0.08 (0.07)	0.02 (0.03)	-0.01 (-0.01)				

Item	Sex	Books	Migration	Position	Sample			Mode
	male vs. female	< 100 vs. ≥ 100	without vs. with	first vs. second	SC 4 vs. SC 6	SC 5 vs. SC 6	SC 4 vs. SC 5	CBT vs. WBT
mag12q051_sc6a9_c	0.03 (0.03)	-0.23 (-0.22)		-0.01 (-0.01)	-0.19 (-0.21)	-0.25 (-0.28)	0.06 (0.06)	-0.26 (-0.28)
maa9v151_c	0.01 (0.01)	-0.10 (-0.09)	0.21 (0.22)	-0.05 (-0.04)	-0.54 (-0.59)	-0.41 (-0.45)	-0.12 (-0.14)	-0.09 (-0.10)
maa3v082_sc6a9_c	-0.39 (-0.37)	-0.03 (-0.03)		-0.26 (-0.24)	-0.31 (-0.34)	-0.05 (-0.05)	-0.27 (-0.29)	-0.04 (-0.04)
maa9d13s_c								
mag9d201_sc6a9_c	-0.08 (-0.08)	0.14 (0.13)	-0.08 (-0.09)	-0.04 (-0.04)	-0.34 (-0.37)	-0.71 (-0.78)	0.37 (0.40)	-0.18 (-0.20)
maa9r03s_c				0.01 (0.01)				
maa9q161_c	-0.50 (-0.48)	-0.22 (-0.21)		0.10 (0.09)	0.42 (0.46)	0.16 (0.18)	0.26 (0.28)	-0.00 (-0.00)
mas1q041_sc6a9_c	0.02 (0.02)	-0.05 (-0.05)	-0.01 (-0.01)	0.06 (0.05)	-0.17 (-0.19)	-0.26 (-0.28)	0.08 (0.09)	0.13 (0.14)
maa9r221_c	-0.19 (-0.18)	-0.04 (-0.04)	0.29 (0.30)	-0.06 (-0.06)	-0.07 (-0.07)	-0.02 (-0.02)	-0.04 (-0.05)	0.43 (0.46)
mas1q02s_sc6a9_c	0.47 (0.45)	-0.24 (-0.23)		-0.14 (-0.13)				
maa9r171_c	-0.03 (-0.03)	-0.06 (-0.06)	0.01 (0.01)	-0.07 (-0.07)	0.29 (0.32)	0.16 (0.18)	0.13 (0.14)	-0.06 (-0.06)
maa9v141_c	0.38 (0.36)	0.00 (0.00)		0.07 (0.06)				
mas1d081_sc6a9_c	0.25 (0.24)	0.18 (0.17)	-0.19 (-0.20)	-0.11 (-0.10)	-0.58 (-0.63)	-0.50 (-0.55)	-0.08 (-0.09)	-0.19 (-0.20)
maa3r121_sc6a9_c	-0.22 (-0.22)	-0.18 (-0.17)	0.03 (0.04)	-0.26 (-0.24)				
maa9d111_c	-0.04 (-0.04)	0.26 (0.25)		0.12 (0.12)				
mag12q111_sc6a9_c	0.04 (0.04)	0.01 (0.01)	-0.06 (-0.06)	-0.03 (-0.03)	0.38 (0.42)	0.55 (0.60)	-0.16 (-0.18)	
maa3d112_sc6a9_c	-0.14 (-0.13)	-0.27 (-0.25)		0.07 (0.06)				-0.34 (-0.36)
maa9r321_c	0.04 (0.04)	0.05 (0.05)		-0.20 (-0.18)	0.01 (0.01)	0.25 (0.27)	-0.24 (-0.26)	0.14 (0.15)
mag9r061_sc6a9_c				0.01 (0.01)				
maa9v27s_c		-0.23 (-0.22)		-0.02 (-0.02)				
maa3q101_sc6a9_c	-0.12 (-0.11)	0.07 (0.07)	-0.15 (-0.16)	0.05 (0.04)	0.18 (0.19)	0.34 (0.37)	-0.16 (-0.17)	0.08 (0.09)
maa9d051_c	0.04 (0.04)	0.16 (0.15)		0.01 (0.01)	-0.15 (-0.17)	-0.09 (-0.10)	-0.06 (-0.06)	-0.09 (-0.09)
mas1q011_sc6a9_c	-0.06 (-0.06)	0.11 (0.10)	-0.02 (-0.02)	0.05 (0.05)				

Item	Sex	Books	Migration	Position	Sample			Mode
	male vs. female	< 100 vs. ≥ 100	without vs. with	first vs. second	SC 4 vs. SC 6	SC 5 vs. SC 6	SC 4 vs. SC 5	CBT vs. WBT
mag9q101_sc6a9_c	0.13 (0.13)	0.39 (0.37)	0.42 (0.44)	0.04 (0.04)				
maa9d121_c		0.13 (0.12)		0.11 (0.10)				
maa9d20s_c	0.34 (0.32)	-0.12 (-0.12)		-0.22 (-0.21)				
maa9r301_c	0.12 (0.12)	-0.12 (-0.11)		-0.11 (-0.10)				
mag12d031_sc6a9_c	-0.20 (-0.20)	0.13 (0.12)		-0.01 (-0.00)				
mag12r041_sc6a9_c	0.05 (0.05)	0.09 (0.08)		0.14 (0.13)				
maa9r26s_c								
maa9v07s_c	0.32 (0.30)	-0.06 (-0.05)		-0.14 (-0.13)				
maa9v28s_c		-0.20 (-0.19)		0.01 (0.01)				
mag12v131_sc6a9_c	0.08 (0.08)	0.07 (0.06)		-0.31 (-0.29)				
Main effect (DIF model)	0.59 (0.57)	-0.33 (-0.31)	0.57 (0.59)	-0.20 (-0.19)	0.18 (0.20)	0.94 (1.03)	-0.76 (-0.82)	-0.11 (-0.12)
Main effect (Main effect model)	0.46 (0.45)	-0.24 (-0.23)	0.54 (0.57)	-0.16 (-0.15)	0.22 (0.24)	0.96 (1.05)	-0.74 (-0.80)	-0.13 (-0.14)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix B.
* Absolute standardized difference is significantly, $p < .05$, greater than 0.40 (see Fischer et al., 2016).

Migration background: There were 12,508 participants without migration background and 1,460 respondents with a migration background. In comparison to subjects without migration background, participants with migration background had, on average, a slightly lower mathematical competence (main effect = 0.57 logits, Cohen's d = 0.59). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.4 logits for most items (highest DIF = 0.42 for item `mag9q101_sc6a9_c`). Moreover, the overall test for DIF using the BIC also favored the main effects model that did not include item-level DIF.

Test position: There were 6,529 participants that received the MST first and 7,495 respondents that received the test after finishing another competence test. Participants receiving the mathematics test second had, on average, a slightly higher mathematical competence (main effect = 0.20 logits, Cohen's d = 0.19). There was no noteworthy item DIF due to test position; the largest difference in estimated difficulties was 0.43 logits for item `maa9q211_c`. Moreover, the overall test for DIF using the BIC (see Table 7) also favored the main effects model that did not include item-level DIF.

Table 7. Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF model	14022	261824	93	262010	262712
	Main effect	14022	262518	50	262618	262996
Books	DIF model	13412	255491	100	255691	256442
	Main effect	13412	255728	54	255836	256241
Migration	DIF model	13968	211531	53	211637	212037
	Main effect	13968	211595	28	211651	211862
Position	DIF model	14024	279569	110	279789	280620
	Main effect	14024	279722	61	279844	280305
Starting cohorts	DIF model	14024	215434	82	215598	216217
	Main effect	14024	216088	30	216148	216375
Mode	DIF model	4643	70178	55	70288	70642
	Main effect	4643	70235	29	70293	70479

Starting cohort: Analyses of differences between the three starting cohorts were based on 6,909 participants from Starting Cohort 4, 2742 participants from Starting Cohort 5, and 4,373 participants from Starting Cohort 6. On average, respondents from Starting Cohort 5 exhibited higher mathematical abilities than respondents from Starting Cohort 4 (0.76 logits, Cohen's d = 0.82) or respondents from Starting Cohort 6 (0.94 logits, Cohen's d = 1.03). One item (`mag9d201_sc6a9_c`) exhibited noteworthy item DIF (DIF = 0.71 logits). However, DIF did not affect the average mean level effects between starting cohorts that were highly comparable between the DIF model and the main effect model that did not acknowledge item level DIF. Moreover, the overall model test using the BIC indicated a slightly better fit for the more parsimonious main effects model that did not account for item level DIF. Thus, no severe DIF was observed for the starting cohorts.

Assessment mode: In Starting Cohort 5 participants received either a proctored computerized test (CBT) or an unproctored web-based test (WBT). Therefore, we examined mode effects in Starting Cohort 5. There were 2,742 respondents in the CBT condition and 1,901 respondents in the WBT condition. As expected, there were no pronounced differences in the subjects' mean abilities between the two modes (0.11 logits, Cohen's d = 0.12). There was also no noteworthy DIF (largest DIF = 0.43 logits for `item maa9r221_c`). Also, the overall tests for DIF favored the main effects model that did not include item-level DIF (see Table 7).

5.3.3 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately among items (see Table 4). The average discrimination parameter fell at 1.10 (*SD* = 0.44). Particularly, the discrimination parameter of 0.20 for item `maa9d20s_c` was rather low. However, an inspection of the respective item characteristic curve of the PCM indicated an adequate fit. Model fit indices

suggested a slightly better model fit of the GPCM (AIC = 284,986, BIC = 285,862, number of parameters = 116) as compared to the PCM (AIC = 285,788, BIC = 286,278, number of parameters = 65). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.4 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics (see Table 4) were quite low (M = 0.08, SD = 0.04)—the largest individual residual correlation was 0.21 (item `maa9d13s_c`)—and, thus, indicated an essentially unidimensional test. Because the mathematics test is constructed to measure a single dimension, a unidimensional mathematical competence score was estimated.

Table 8. Results of Four-Dimensional Scaling

	Units and measuring	**Change and relationships**	**Data and chance**	**Space and shape**
Units and measuring (13 items)	1.297			
Change and relationships (14 items)	0.950	0.966		
Data and chance (11 items)	0.965	0.949	1.779	
Space and shape (14 items)	0.940	0.950	0.932	1.414

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

We also examined the dimensionality of the test by specifying a four-dimensional model based on the four different content areas (see section 2.1). Each item was assigned to one content area (between-item-multidimensionality). The multidimensional model was estimated using Quasi Monte Carlo method with 5,000 nodes. The variances and correlations of the four dimensions are summarized in Table 8. All dimensions exhibited substantial variances. As expected, the correlations between the four dimensions were rather high, falling between 0.93 and 0.95. Thus, they did not deviate substantially from a perfect correlation (i.e., r = .95, see Carstensen, 2013). Still, according to model fit indices, the four-dimensional model fitted the data slightly better (AIC =285,637, BIC = 286,195, number of parameters = 74) than the unidimensional model (AIC = 285,788, BIC = 286,278, number of parameters = 65). These results indicate that the four content areas measure a common construct, although they are not completely unidimensional.

6 Discussion

The analyses in the previous sections reported information on the quality of the mathematical test that was administered in Starting Cohorts 4, 5, and 6. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number of missing responses was rather large because many respondents did not finish the test in time. The test had a satisfactory reliability and distinguished well between test takers. However, the test was slightly better targeted at mediocre- and low-performing students and covered the high ability spectrum less well. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students. In summary, the test had acceptable psychometric properties that allowed the estimation of a unidimensional mathematical competence.

7 Data in the Scientific Use Files

7.1 Naming Conventions

The SUFs for the three starting cohorts contain 64 items, of which 52 were scored dichotomously (multiple choice items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 12 items were scored as polytomous variables (matching items) that are marked with a 's_c' at the end of the variable names. For further details on the naming conventions of the variables see Fuß and colleagues (2019). Twelve items were removed from the final scaling procedure because few valid responses were available (see Figure 1). Nevertheless, these items are included in the SUFs.

7.2 Linking of competence scores

In all starting cohorts, mathematical competence was measured in the current wave and also in a previous wave. The tests in the different waves were constructed in such a way as to allow for an accurate measurement of mathematical competence within the respective age group. As a consequence, the competence scores derived in the different waves cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across waves, the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016) was adopted. Following an anchor-items design, the responses from the current wave were linked to the scale of the test that was administered in the previous wave of each starting cohort.

7.2.1 Starting Cohort 4

In Starting Cohort 4, a subsample of 3,437 respondents (53% women) participated at both measurement occasions, in wave 7 (i.e., grade 12; see Fischer, Rohm, & Gnambs, 2017) and also in wave 10 (see above). Consequently, these respondents were used to link the two tests across both waves. The test administered in wave 7 included 29 items, whereas the MST in wave 10 had 48 items (4 items were excluded because they had less than 50 valid responses). Because the two tests administered at the two waves shared 17 items, an anchor-items design was used to link both tests (see Fischer et al., 2016).

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 17 common items that were included in both waves showed a non-negligible shift in item difficulties. The differences in item difficulties between waves 7 and 10 of Starting Cohort 4 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 9. Although the minimum effects hypothesis test was not significant (α = .05) for any item, some items exhibited noticeable DIF effects (absolute difference in logits: *Min* = 0.05, *Max* = 0.89). Therefore, we selected 13 items with DIF effects that did not exceed 0.40 on the logit scale. The mathematical competence tests administered in the two waves were linked using the "mean/mean" method for the anchor-items design using these 13 common items with (see Fischer et al., 2016).

The correction term was calculated as *c* = 0.319. Previously, the correction term between wave 1 (i.e., grade 9) and wave 7 (i.e., grade 12) has been estimated as 0.496 (see Fischer et al., 2017). Therefore, the combined correction term of 0.496 + 0.319 = 0.815 was subsequently added to each difficulty parameter estimated in the three CBT samples (see Table 4) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.055 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change between the two linked measurements.

Table 9. Differential Item Functioning Analyses between Waves 7 and 10 of Starting Cohort 4

Item	Δσ	$SE_{\Delta\sigma}$	*F*
mag12r011_sc4a10_c	-0.07	0.06	1.39
mag9v011_sc4a10_c	0.39	0.13	8.88
mag9r061_sc4a10_c	-0.89	0.14	38.40
mag12d031_sc4a10_c	-0.56	0.12	21.73
mag12r041_sc4a10_c	0.23	0.17	1.86
mag9q101_sc4a10_c	-0.06	0.18	0.10
mag12q051_sc4a10_c	-0.20	0.09	4.90
maa3q071_sc4a10_c	0.05	0.06	0.74
mag12r091_sc4a10_c	-0.52	0.08	41.47
mag12v101_sc4a10_c	-0.24	0.06	17.88
mag12q111_sc4a10_c	0.07	0.13	0.25
mag12v122_sc4a10_c	0.09	0.06	2.79
maa3q101_sc4a10_c	0.15	0.12	1.71
mas1d081_sc4a10_c	-0.32	0.08	16.48
maa3d112_sc4a10_c	-0.49	0.13	13.74
mag12v131_sc4a10_c	-0.11	0.12	0.76
maa3r121_sc4a10_c	0.14	0.18	0.61

Note. Δσ = Difference in item difficulty parameters

Item	Δσ	$SE_{\Delta\sigma}$	*F*
between waves (negative values indicate easier items in wave 7); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 3437) = 80.73. A non-significant test indicates measurement invariance.			
$^{*}p < .05$			

7.2.2 Starting Cohort 5

In Starting Cohort 5, a subsample of 2,257 respondents (64% women) participated at both measurement occasions, in wave 1 (see Gerken & Schnittjer, 2017) and also in wave 12 (see above). This subsample included only respondents from the CBT condition in wave 12. Consequently, these respondents were used to link the two tests across both waves. The test administered in wave 1 included 20 items, whereas the MST in wave 10 had 52 items. Because the two tests administered at the two waves shared eight items, an anchor-items design was used for the linking (see Fischer et al., 2016).

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 eight common items that were included in both waves showed a non-negligible shift in item difficulties. The differences in item difficulties between waves 1 and 12 of Starting Cohort 5 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 10. The minimum effects hypothesis test identified one item (`mas1v062_sc5s12_c`) with substantial DIF (α = .05). Moreover, another item (`mas1q011_sc5s12_c`) had an absolute DIF greater than 0.40. Both items were significantly more difficult in wave 12 as compared to wave 1. Therefore, these items were excluded from the linking procedure. The mathematical competence tests administered in the two waves were linked using the “mean/mean” method for the anchor-items design using six common items without DIF (see Fischer et al., 2016).

The correction term was calculated as *c* = -0.010. This correction term was subsequently added to each difficulty parameter estimated in the three CBT samples (see Table 4) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.075 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change between the two linked measurements.

Table 10. Differential Item Functioning Analyses between Waves 1 and 12 of Starting Cohort 5

Item	Δσ	$SE_{\Delta\sigma}$	*F*
mas1q011_sc5s12_c	-0.723	0.173	17.458
maa3q071_sc5s12_c	0.173	0.079	4.764
maa3r081_sc5s12_c	-0.222	0.089	6.310

Item	Δσ	$SE_{\Delta\sigma}$	*F*
maa3v082_sc5s12_c	-0.075	0.126	0.358
mas1v032_sc5s12_c	-0.038	0.070	0.301
mas1q041_sc5s12_c	-0.084	0.112	0.568
mas1v062_sc5s12_c	0.682	0.086	62.253*
maa3d112_sc5s12_c	-0.372	0.152	5.959

Note. Δσ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 1); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 3437) = 57.68. A non-significant test indicates measurement invariance.

*$p < .05$

7.2.3 Starting Cohort 6

In Starting Cohort 6, a subsample of 3,212 respondents (52% women) participated at both measurement occasions, in wave 3 (see Jordan & Duchhardt, 2013) and also in wave 9 (see above). Consequently, these respondents were used to link the two tests across both waves. The test administered in wave 3 included 21 items, whereas the MST in wave 9 had 50 items (2 items were excluded because they had less than 50 valid responses). Because the two tests administered at the two waves shared seven items, an anchor-items design was used for the linking (see Fischer et al., 2016).

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 seven common items that were included in both waves showed a non-negligible shift in item difficulties. The differences in item difficulties between waves 3 and 9 of Starting Cohort 6 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 11. The minimum effects hypothesis test was not significant (α = .05) for any item. However, one item (`maa3v082_sc6a9_c`) had an absolute difference in logits of 0.46 and, thus, was excluded from the linking procedure. Therefore, the mathematical competence tests administered in the two waves were linked using the "mean/mean" method for the anchor-items design using six common items (see Fischer et al., 2016).

The correction term was calculated as c = -0.111. This correction term was subsequently added to each difficulty parameter estimated in the three CBT samples (see Table 4) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.036 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change between the two linked measurements.

Table 11. Differential Item Functioning Analyses between Waves 3 and 9 of Starting Cohort 6

Item	Δσ	$SE_{\Delta\sigma}$	*F*
maa3q071_sc6a9_c	0.14	0.06	5.64
maa3r081_sc6a9_c	0.26	0.07	13.30
maa3v082_sc6a9_c	0.46	0.14	10.40
maa3q101_sc6a9_c	0.30	0.08	12.97
maa3d112_sc6a9_c	0.34	0.21	2.74
maa3r121_sc6a9_c	0.19	0.11	2.77
mag9r051_sc6a9_c	0.12	0.06	3.41

Note. Δσ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 3); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 3210) = 76.40. A non-significant test indicates measurement invariance.

$^{*}p < .05$

7.3 Mathematical competence scores

In the SUF, manifest mathematical competence scores are provided in the form of WLEs (`maa10_sc1`, `mas12_sc1`, `maa9_sc1`) including their respective standard error (`maa10_sc2`, `mas12_sc2`, `maa9_sc2`). Because the responses for the three starting cohorts were concurrently scaled, these WLEs can be used to compare mathematical competences across starting cohorts. For `maa10_sc1u`, `mas12_sc1u`, and `maa9_sc1u` person abilities were estimated using the linked item difficulty parameters. As a result, these WLE scores can be used for longitudinal comparisons between different waves within a starting cohort. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in `maa10_sc1`, `mas12_sc1`, and `maa9_sc1` are not linked to the underlying reference scale of the previous wave. However, they are corrected for the position of the mathematical test within the test battery. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The R Syntax for estimating the WLEs is provided in Appendix C. Because no substantial DIF was found for the proctored CBT and the unproctored WBT conditions in Starting Cohort 5, WLEs for respondents receiving the WBT were estimated using the fixed item parameters from the CBT scaling (see Table 4)[2]. In the

[2] The test taking behavior in unproctored testing cannot be properly supervised and, thus, might not be comparable to proctored settings (see Kröhne, Gnambs, & Goldhammer, 2019). Therefore, we inspected the response times in for respondents in the WBT condition. For 319 respondents exhibiting breaks (with no test interaction) of more than five minutes during the test no WLEs were estimated because they were suspected to adopt different test taking strategies.

IRT scaling model, all polytomous variables and two dichotomous variables (`maa9r301_c` and `mas1v032_sc6a9_c`) were scored as 0.5 for each category. For respondents who did not take part in the mathematical test or who did not give enough valid responses no WLEs were estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, 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*, 716-723. https://doi.org/10.1109/TAC.1974.1100705

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.

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

Fischer, L., Rohm, T., & Gnambs, T., (2017). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 12). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (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. (2019). *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.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)*. Scientific Use File 2012, Version 1.0.0. Bamberg: University of Bamberg, National Educational Panel Study.

Gerken, A.-L., & Schnittjer, I. (2017). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 5 for First-Year Students* (NEPS Survey Paper No. 17). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gnambs, T., & Carstensen, C. H. (in prep). *Longitudinal Multi-Stage Testing of Mathematical Competence in the National Educational Panel Study*. Manuscript in preparation.

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

Kröhne, U., Gnambs, T., & Goldhammer, F. (2019). *Disentangling setting and mode effects for online competence assessment*. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), Education as a lifelong process (2nd ed., pp. 171-193). Wiesbaden, Germany: Springer VS. https://doi.org/10.1007/978-3-658-23162-0_10

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1177/014662169201600206

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

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, DK: The Danish Institute of Education Research.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. R package version 2.12-18. URL: https://CRAN.R-project.org/package=TAM

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

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*, 461-464. https://doi.org/10.1214/aos/1176344136

Van den Ham, A.-K. (2016). *Ein Validitätsargument für den Mathematiktest der National Educational Panel Study für die neunte Klassenstufe*. Unpublished German dissertation, Leuphana University Lüneburg, Lüneburg.

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

Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0*. Camberwell, Australia: 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. https://doi.org/10.1177/014662168400800201

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

Appendices

Appendix A: Content Areas for the Items in the Multi-Stage-Test

Position	Stage	Item	Content area
1	1	maa3q071_sc6a9_c	Units and measuring
2	1	mag12v101_sc6a9_c	Change and relationships
3	1	mag12v122_sc6a9_c	Change and relationships
4	1	mag12r011_sc6a9_c	Space and shape
5	1	mas1v032_sc6a9_c	Change and relationships
6	2	maa9q081_c	Units and measuring
6	2	maa9r311_c	Space and shape
6	2	maa9d331_c	Data and chance
7	2	mag9v011_sc6a9_c	Change and relationships
7	2	mag12r091_sc6a9_c	Space and shape
7	2	mas1v062_sc6a9_c	Change and relationships
8	2	mag9r051_sc6a9_c	Space and shape
8	2	maa9q19s_c	Units and measuring
9	2	maa9d09s_c	Data and chance
9	2	maa9v193_c	Change and relationships
10	3	maa9r101_c	Space and shape
10	3	maa3r081_sc6a9_c	Space and shape
10	3	maa9q211_c	Units and measuring
10	3	maa9v251_c	Change and relationships
11	3	maa9q011_c	Units and measuring
11	3	mag12q051_sc6a9_c	Units and measuring
11	3	maa9v151_c	Change and relationships
12	3	maa3v082_sc6a9_c	Change and relationships
12	3	maa9d13s_c	Data and chance
12	3	maa9d241_c	Data and chance
12	3	mag9d201_sc6a9_c	Data and chance
13	3	maa9r03s_c	Space and shape
13	3	maa9q161_c	Units and measuring
13	3	mas1q041_sc6a9_c	Units and measuring
13	3	mas1v042_sc6a9_c	Units and measuring
13	3	maa9r221_c	Space and shape
14	4	mas1q02s_sc6a9_c	Units and measuring

Position	Stage	Item	Content area
14	4	maa9r171_c	Space and shape
14	4	maa9r18s_c	Space and shape
15	4	maa9v141_c	Change and relationships
15	4	maa9r183_c	Space and shape
15	4	mas1d081_sc6a9_c	Data and chance
15	4	maa3r121_sc6a9_c	Space and shape
16	4	maa9d111_c	Data and chance
16	4	mag12q111_sc6a9_c	Units and measuring
16	4	maa3d112_sc6a9_c	Data and chance
16	4	maa9r321_c	Space and shape
17	4	mag9r061_sc6a9_c	Space and shape
17	4	maa9v27s_c	Change and relationships
17	4	maa3q101_sc6a9_c	Units and measuring
18	5	maa9d051_c	Data and chance
18	5	mas1q011_sc6a9_c	Units and measuring
18	5	mag9q101_sc6a9_c	Units and measuring
18	5	maa9d23s_c	Data and chance
19	5	maa9q021_c	Units and measuring
19	5	maa9q041_c	Units and measuring
19	5	maa9d121_c	Data and chance
19	5	maa9d20s_c	Data and chance
19	5	maa9r301_c	Space and shape
20	5	maa9q022_c	Units and measuring
20	5	mag12d031_sc6a9_c	Data and chance
20	5	mag12r041_sc6a9_c	Space and shape
20	5	maa9r26s_c	Space and shape
20	5	maa9r291_c	Space and shape
21	5	maa9v061_c	Change and relationships
21	5	maa9v07s_c	Change and relationships
21	5	mag12v061_sc6a9_c	Change and relationships
21	5	maa9v28s_c	Change and relationships
21	5	mag12v131_sc6a9_c	Change and relationships

Note. So far, the internal validity of the individual dimensions of mathematical competence has not yet been confirmed (van den Ham, 2016). Therefore, analyses on the content level should not be conducted.

Appendix B: Variable Names in Different Starting Cohorts

Position	Stage	Starting Cohort 4	Starting Cohort 5	Starting Cohort 6
1	1	maa3q071_sc4a10_c	maa3q071_sc5s12_c	maa3q071_sc6a9_c
2	1	mag12v101_sc4a10_c	mag12v101_sc5s12_c	mag12v101_sc6a9_c
3	1	mag12v122_sc4a10_c	mag12v122_sc5s12_c	mag12v122_sc6a9_c
4	1	mag12r011_sc4a10_c	mag12r011_sc5s12_c	mag12r011_sc6a9_c
5	1	mas1v032_sc4a10_c	mas1v032_sc5s12_c	mas1v032_sc6a9_c
6	2	maa9q081_sc4a10_c	maa9q081_sc5s12_c	maa9q081_c
6	2	maa9r311_sc4a10_c	maa9r311_sc5s12_c	maa9r311_c
6	2	maa9d331_sc4a10_c	maa9d331_sc5s12_c	maa9d331_c
7	2	mag9v011_sc4a10_c	mag9v011_sc5s12_c	mag9v011_sc6a9_c
7	2	mag12r091_sc4a10_c	mag12r091_sc5s12_c	mag12r091_sc6a9_c
7	2	mas1v062_sc4a10_c	mas1v062_sc5s12_c	mas1v062_sc6a9_c
8	2	mag9r051_sc4a10_c	mag9r051_sc5s12_c	mag9r051_sc6a9_c
8	2	maa9q19s_sc4a10_c	maa9q19s_sc5s12_c	maa9q19s_c
9	2	maa9d09s_sc4a10_c	maa9d09s_sc5s12_c	maa9d09s_c
9	2	maa9v193_sc4a10_c	maa9v193_sc5s12_c	maa9v193_c
10	3	maa9r101_sc4a10_c	maa9r101_sc5s12_c	maa9r101_c
10	3	maa3r081_sc4a10_c	maa3r081_sc5s12_c	maa3r081_sc6a9_c
10	3	maa9q211_sc4a10_c	maa9q211_sc5s12_c	maa9q211_c
10	3	maa9v251_sc4a10_c	maa9v251_sc5s12_c	maa9v251_c
11	3	maa9q011_sc4a10_c	maa9q011_sc5s12_c	maa9q011_c
11	3	mag12q051_sc4a10_c	mag12q051_sc5s12_c	mag12q051_sc6a9_c
11	3	maa9v151_sc4a10_c	maa9v151_sc5s12_c	maa9v151_c
12	3	maa3v082_sc4a10_c	maa3v082_sc5s12_c	maa3v082_sc6a9_c
12	3	maa9d13s_sc4a10_c	maa9d13s_sc5s12_c	maa9d13s_c
12	3	maa9d241_sc4a10_c	maa9d241_sc5s12_c	maa9d241_c
12	3	mag9d201_sc4a10_c	mag9d201_sc5s12_c	mag9d201_sc6a9_c
13	3	maa9r03s_sc4a10_c	maa9r03s_sc5s12_c	maa9r03s_c
13	3	maa9q161_sc4a10_c	maa9q161_sc5s12_c	maa9q161_c
13	3	mas1q041_sc4a10_c	mas1q041_sc5s12_c	mas1q041_sc6a9_c
13	3	mas1v042_sc4a10_c	mas1v042_sc5s12_c	mas1v042_sc6a9_c
13	3	maa9r221_sc4a10_c	maa9r221_sc5s12_c	maa9r221_c
14	4	mas1q02s_sc4a10_c	mas1q02s_sc5s12_c	mas1q02s_sc6a9_c
14	4	maa9r171_sc4a10_c	maa9r171_sc5s12_c	maa9r171_c

Position	Stage	Starting Cohort 4	Starting Cohort 5	Starting Cohort 6
14	4	maa9r18s_sc4a10_c	maa9r18s_sc5s12_c	maa9r18s_c
15	4	maa9v141_sc4a10_c	maa9v141_sc5s12_c	maa9v141_c
15	4	maa9r183_sc4a10_c	maa9r183_sc5s12_c	maa9r183_c
15	4	mas1d081_sc4a10_c	mas1d081_sc5s12_c	mas1d081_sc6a9_c
15	4	maa3r121_sc4a10_c	maa3r121_sc5s12_c	maa3r121_sc6a9_c
16	4	maa9d111_sc4a10_c	maa9d111_sc5s12_c	maa9d111_c
16	4	mag12q111_sc4a10_c	mag12q111_sc5s12_c	mag12q111_sc6a9_c
16	4	maa3d112_sc4a10_c	maa3d112_sc5s12_c	maa3d112_sc6a9_c
16	4	maa9r321_sc4a10_c	maa9r321_sc5s12_c	maa9r321_c
17	4	mag9r061_sc4a10_c	mag9r061_sc5s12_c	mag9r061_sc6a9_c
17	4	maa9v27s_sc4a10_c	maa9v27s_sc5s12_c	maa9v27s_c
17	4	maa3q101_sc4a10_c	maa3q101_sc5s12_c	maa3q101_sc6a9_c
18	5	maa9d051_sc4a10_c	maa9d051_sc5s12_c	maa9d051_c
18	5	mas1q011_sc4a10_c	mas1q011_sc5s12_c	mas1q011_sc6a9_c
18	5	mag9q101_sc4a10_c	mag9q101_sc5s12_c	mag9q101_sc6a9_c
18	5	maa9d23s_sc4a10_c	maa9d23s_sc5s12_c	maa9d23s_c
19	5	maa9q021_sc4a10_c	maa9q021_sc5s12_c	maa9q021_c
19	5	maa9q041_sc4a10_c	maa9q041_sc5s12_c	maa9q041_c
19	5	maa9d121_sc4a10_c	maa9d121_sc5s12_c	maa9d121_c
19	5	maa9d20s_sc4a10_c	maa9d20s_sc5s12_c	maa9d20s_c
19	5	maa9r301_sc4a10_c	maa9r301_sc5s12_c	maa9r301_c
20	5	maa9q022_sc4a10_c	maa9q022_sc5s12_c	maa9q022_c
20	5	mag12d031_sc4a10_c	mag12d031_sc5s12_c	mag12d031_sc6a9_c
20	5	mag12r041_sc4a10_c	mag12r041_sc5s12_c	mag12r041_sc6a9_c
20	5	maa9r26s_sc4a10_c	maa9r26s_sc5s12_c	maa9r26s_c
20	5	maa9r291_sc4a10_c	maa9r291_sc5s12_c	maa9r291_c
21	5	maa9v061_sc4a10_c	maa9v061_sc5s12_c	maa9v061_c
21	5	maa9v07s_sc4a10_c	maa9v07s_sc5s12_c	maa9v07s_c
21	5	mag12v061_sc4a10_c	mag12v061_sc5s12_c	mag12v061_sc6a9_c
21	5	maa9v28s_sc4a10_c	maa9v28s_sc5s12_c	maa9v28s_c
21	5	mag12v131_sc4a10_c	mag12v131_sc5s12_c	mag12v131_sc6a9_c

Appendix C: Percentage of Missing Values by Item and Starting Cohort

Percentage of Missing Values by Item in Starting Cohorts 4 and 6.

		Starting Cohort 4				Starting Cohort 6			
Pos.	**Item**	***N***	***N_v***	**OM**	**NR**	***N***	***N_v***	**OM**	**NR**
1	maa3q071_sc4a10_c	6909	6890	0.28	0.00	4373	4340	0.75	0.00
2	mag12v101_sc4a10_c	6909	6800	1.58	0.00	4373	4163	4.80	0.00
3	mag12v122_sc4a10_c	6909	6852	0.83	0.00	4373	4210	3.73	0.00
4	mag12r011_sc4a10_c	6909	6856	0.72	0.04	4373	4322	1.10	0.07
5	mas1v032_sc4a10_c	6909	6648	3.49	0.29	4373	3852	11.43	0.48
6	maa9q081_sc4a10_c	1765	1745	0.06	1.08	797	785	0.13	1.38
6	maa9r311_sc4a10_c	3368	3321	1.22	0.18	2071	1988	3.09	0.92
6	maa9d331_sc4a10_c	1752	1733	0.86	0.06	1483	1432	2.83	0.20
7	mag9v011_sc4a10_c	1752	1737	0.68	0.17	1483	1444	2.02	0.61
7	mag12r091_sc4a10_c	1765	1674	0.96	4.19	797	736	2.13	5.52
7	mas1v062_sc4a10_c	3368	3310	0.53	1.19	2071	1977	2.37	2.17
8	mag9r051_sc4a10_c	5120	5001	0.57	1.76	3554	3421	1.32	2.42
8	maa9q19s_sc4a10_c	1765	1602	0.57	8.67	797	694	2.13	10.79
9	maa9d09s_sc4a10_c	1752	1708	1.08	1.37	1483	1324	7.08	3.30
9	maa9v193_sc4a10_c	5133	4574	3.43	6.97	2868	2385	6.28	10.08
10	maa3r081_sc4a10_c	4576	4317	2.43	3.23	2732	2452	6.11	4.14
10	maa9q211_sc4a10_c	524	515	1.53	0.19	579	548	4.32	1.04
10	maa9v251_sc4a10_c	1105	1049	0.18	4.89	549	503	0.55	7.83
11	maa9q011_sc4a10_c	2811	2666	1.28	3.88	2184	1985	2.75	6.36
11	mag12q051_sc4a10_c	1383	1198	0.14	13.23	678	564	0.59	16.22
11	maa9v151_sc4a10_c	2288	1984	3.76	9.53	1127	856	12.78	11.27
12	maa3v082_sc4a10_c	1105	826	0.72	24.52	549	386	0.73	28.96
12	maa9d13s_sc4a10_c	523	484	5.16	1.91	579	439	16.93	6.74
12	mag9d201_sc4a10_c	4576	4099	0.20	10.23	2732	2368	0.40	12.92
13	maa9r03s_sc4a10_c	523	485	3.63	3.63	579	444	12.61	10.36
13	maa9q161_sc4a10_c	1105	742	0.00	32.85	549	337	0.18	38.43
13	mas1q041_sc4a10_c	2288	1711	1.27	23.95	1127	738	5.32	29.19
13	maa9r221_sc4a10_c	2288	2002	3.98	8.52	1605	1206	10.34	14.52
14	mas1q02s_sc4a10_c	2327	2060	6.45	4.77	1758	1301	17.52	7.68
14	maa9r171_sc4a10_c	2675	2157	7.59	9.98	1221	928	9.42	13.68
15	maa9v141_sc4a10_c	699	520	0.43	25.18	272	178	1.84	32.72
15	mas1d081_sc4a10_c	3650	3120	1.53	12.47	2044	1584	3.13	18.98
15	maa3r121_sc4a10_c	653	634	0.61	2.30	663	594	2.41	7.99
16	maa9d111_sc4a10_c	652	565	7.36	5.98	663	439	16.59	16.89
16	mag12q111_sc4a10_c	1674	1403	0.12	16.07	1095	796	0.27	27.03
16	maa3d112_sc4a10_c	864	463	0.58	45.83	343	159	1.75	51.90

		Starting Cohort 4				Starting Cohort 6			
Pos.	**Item**	***N***	***N_v***	**OM**	**NR**	***N***	***N_v***	**OM**	**NR**
16	maa9r321_sc4a10_c	1976	1482	0.66	24.34	949	653	0.53	30.66
17	mag9r061_sc4a10_c	864	367	0.93	56.37	343	120	0.58	64.43
17	maa9v27s_sc4a10_c	652	583	2.30	8.28	663	425	12.52	23.38
17	maa3q101_sc4a10_c	3650	2644	0.77	26.79	2044	1304	0.44	35.76
18	maa9d051_sc4a10_c	1423	1149	2.25	14.69	579	450	1.90	19.17
18	mas1q011_sc4a10_c	1383	1274	1.23	6.65	742	664	1.48	9.03
18	mag9q101_sc4a10_c	741	659	3.24	7.83	569	475	3.34	13.18
19	maa9d121_sc4a10_c	741	656	0.40	11.07	569	439	3.51	19.33
19	maa9d20s_sc4a10_c	1383	1069	4.41	18.29	742	511	7.68	23.18
19	maa9r301_sc4a10_c	1091	875	0.00	19.80	459	335	0.00	27.02
20	mag12d031_sc4a10_c	1091	736	0.37	32.17	459	277	0.22	39.43
20	mag12r041_sc4a10_c	740	610	0.95	16.62	569	388	0.70	31.11
20	maa9r26s_sc4a10_c	1383	944	2.82	28.85	742	431	5.39	35.58
21	maa9v07s_sc4a10_c	1091	436	0.18	59.76	459	143	0.00	68.85
21	maa9v28s_sc4a10_c	740	478	3.11	31.22	569	202	4.39	57.47
21	mag12v131_sc4a10_c	1383	892	0.00	35.50	742	409	0.00	44.88

Note. Pos. = Item position within test. *N* = Number of respondents the item was administered to, N_v = Number of valid responses, NR = Percentage of respondents that did not reach item within a level plus percentage of respondents that aborted the test in a previous stage, OM = Percentage of respondents that omitted the item.

Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohort 4 are given in Appendix B.

Percentage of Missing Values by Item in Starting Cohort 5.

		Proctored (CBT)				Unproctored (WBT)			
Pos.	**Item**	***N***	***N_v***	**OM**	**NR**	***N***	***N_v***	**OM**	**NR**
1	maa3q071_sc5s12_c	2742	2740	0.07	0.00	1901	1898	0.16	0.00
2	mag12v101_sc5s12_c	2742	2722	0.73	0.00	1901	1892	0.47	0.00
3	mag12v122_sc5s12_c	2742	2712	1.09	0.00	1901	1882	1.00	0.00
4	mag12r011_sc5s12_c	2742	2719	0.80	0.04	1901	1884	0.53	0.37
5	mas1v032_sc5s12_c	2742	2679	1.97	0.33	1901	1860	1.00	1.16
6	maa9q081_sc5s12_c	1170	1157	0.00	1.11	943	928	0.21	1.38
6	maa9r311_sc5s12_c	1283	1272	0.62	0.23	780	768	0.64	0.90
6	maa9d331_sc5s12_c	280	279	0.36	0.00	156	156	0.00	0.00
7	mag9v011_sc5s12_c	280	277	0.36	0.71	156	155	0.00	0.64
7	mag12r091_sc5s12_c	1170	1103	1.03	4.70	943	903	0.85	3.39
7	mas1v062_sc5s12_c	1279	1264	0.16	1.02	780	766	0.26	1.54
8	mag9r051_sc5s12_c	1557	1512	0.45	2.44	936	912	0.32	2.24
8	maa9q19s_sc5s12_c	1167	1048	0.69	9.51	943	857	0.74	8.27
9	maa9d09s_sc5s12_c	280	270	0.71	2.50	156	151	1.92	1.28
9	maa9v193_sc5s12_c	2441	2118	2.58	10.36	1706	1545	1.76	7.39
10	maa3r081_sc5s12_c	1358	1263	2.58	4.42	877	820	2.05	4.45
10	maa9q211_sc5s12_c	36	35	2.78	0.00	24	20	12.50	4.17

		Proctored (CBT)				Unproctored (WBT)			
Pos.	**Item**	***N***	***N_v***	**OM**	**NR**	***N***	***N_v***	**OM**	**NR**
10	maa9v251_sc5s12_c	754	711	0.13	5.57	551	518	4.90	1.09
11	maa9q011_sc5s12_c	481	454	0.62	4.99	292	282	1.03	2.40
11	mag12q051_sc5s12_c	1037	874	0.10	15.62	846	749	0.47	10.99
11	maa9v151_sc5s12_c	907	784	2.21	11.36	613	541	1.63	10.11
12	maa3v082_sc5s12_c	741	534	0.00	27.94	551	443	0.18	19.42
12	maa9d13s_sc5s12_c	36	31	11.11	2.78	24	21	8.33	4.17
12	mag9d201_sc5s12_c	1340	1157	0.07	13.58	877	782	0.34	10.49
13	maa9r03s_sc5s12_c	36	34	0.00	5.56	24	21	8.33	4.17
13	maa9q161_sc5s12_c	733	465	0.00	36.56	551	412	0.00	25.23
13	mas1q041_sc5s12_c	893	619	1.79	28.89	613	473	2.61	20.23
13	maa9r221_sc5s12_c	441	373	3.17	12.24	268	242	2.61	7.09
14	mas1q02s_sc5s12_c	316	265	6.01	10.13	225	201	9.33	1.33
14	maa9r171_sc5s12_c	1096	897	4.11	12.50	889	758	4.16	10.01
15	maa9v141_sc5s12_c	377	271	0.27	27.85	367	285	11.17	11.17
15	mas1d081_sc5s12_c	979	794	0.61	17.67	751	615	1.20	16.78
15	maa3r121_sc5s12_c	38	34	2.63	7.89	31	30	3.23	0.00
16	maa9d111_sc5s12_c	38	31	2.63	15.79	31	24	19.35	3.23
16	mag12q111_sc5s12_c	275	219	0.00	20.36	194	172	1.03	10.31
16	maa3d112_sc5s12_c	586	275	0.17	52.90	593	348	0.51	40.81
16	maa9r321_sc5s12_c	701	487	0.29	30.24	558	394	0.36	29.03
17	mag9r061_sc5s12_c	576	198	0.87	64.41	593	295	0.67	49.58
17	maa9v27s_sc5s12_c	38	28	7.89	18.42	31	29	0.00	6.45
17	maa3q101_sc5s12_c	972	640	0.10	34.05	751	529	0.53	29.03
18	maa9d051_sc5s12_c	499	412	0.40	15.83	544	469	1.47	12.13
18	mas1q011_sc5s12_c	252	220	1.59	11.11	180	167	6.11	1.11
18	mag9q101_sc5s12_c	47	42	2.13	8.51	47	43	8.51	0.00
19	maa9d121_sc5s12_c	47	42	0.00	10.64	47	44	0.00	6.38
19	maa9d20s_sc5s12_c	248	161	4.44	30.24	180	145	4.44	15.00
19	maa9r301_sc5s12_c	319	252	0.00	21.00	352	274	0.28	21.88
20	mag12d031_sc5s12_c	313	212	0.64	31.63	352	247	0.28	29.55
20	mag12r041_sc5s12_c	47	38	0.00	19.15	47	40	0.00	14.89
20	maa9r26s_sc5s12_c	244	134	4.10	40.98	180	134	2.22	23.33
21	maa9v07s_sc5s12_c	310	124	0.00	58.71	352	175	0.00	50.28
21	maa9v28s_sc5s12_c	46	26	6.52	36.96	46	31	4.35	26.09
21	mag12v131_sc5s12_c	238	120	0.00	49.58	180	130	0.56	27.22

Note. Pos. = Item position within test. *N* = Number of respondents the item was administered to, N_v = Number of valid responses, NR = Percentage of respondents that did not reach item within a level plus percentage of respondents that aborted the test in a previous stage, OM = Percentage of respondents that omitted the item.

Appendix C: R-Syntax for estimating WLEs in starting cohorts 4, 5, and 6

```
# load packages
library(haven) # to import SPSS files
library(doBy)  # recode variables
library(TAM)   # for IRT analyses

# load competence data
dat <- read_sav("SUF for competencies.sav")

# 52 items of the mathematical competence test
items <- c("maa3q071 sc6a9 c", "mag12v101 sc6a9 c",
           "mag12v122 sc6a9 c", "mag12r011 sc6a9 c",
           ...)

# identify polytomous items
f <- c("mas1v032_sc6a9_c", "maa9d09s_c", "maa9d13s_c",
       "maa9r03s_c", "mas1q02s_sc6a9_c", "maa9v27s_c",
       "maa9d20s_c", "maa9r301_c", "maa9r26s_c")
f <- items %in% f

# collapse response categories
dat$maa9v28s_c <- recodeVar(dat$maa9v28s_c,
                            c(0, 1, 2, 3, 4, 5, 6),
                            c(0, 0, 0, 0, 1, 1, 1))
dat$maa9v27s_c <- recodeVar(dat$maa9v27s_c,
                            c(0, 1, 2, 3),
                            c(0, 0, 1, 2))
dat$maa9r26s_c <- recodeVar(dat$maa9r26s_c,
                            c(0, 1, 2, 3, 4),
                            c(0, 0, 1, 2, 3))
dat$maa9d23s_c <- recodeVar(dat$maa9d23s_c,
                            c(0, 1, 2),
                            c(0, 0, 1))
dat$maa9r18s_c <- recodeVar(dat$maa9r18s_c,
                            c(0, 1, 2),
                            c(0, 1, 1))
dat$maa9d13s_c <- recodeVar(dat$maa9d13s_c,
                            c(0, 1, 2, 3, 4, 5, 6),
                            c(0, 0, 0, 0, 1, 2, 2))
dat$maa9d09s_c <- recodeVar(dat$maa9d09s_c,
                            c(0, 1, 2, 3, 4, 5),
                            c(0, 0, 1, 2, 3, 4))
dat$maa9v07s_c <- recodeVar(dat$maa9v07s_c,
                            c(0, 1, 2),
                            c(0, 1, 1))
```

```
dat$maa9r03s_c <- recodeVar(dat$maa9r03s_c,
                            c(0, 1, 2, 3),
                            c(0, 0, 1, 2))

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[f, 1] <- 0.5    # score of 0.5

# estimate partial credit model
mod <- tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2",
               pid = dat$ID_t)
summary(mod)

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```