NEPS
National Educational Panel Study

NEPS Working Papers

Christoph Duchhardt & Annkathrin Gerdes

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 3 in Fifth Grade

NEPS Working Paper No. 19

Bamberg, December 2012

SPONSORED BY THE

Federal Ministry
of Education
and Research

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

NEPS Technical Report for Mathematics –

Scaling Results of Starting Cohort 3 in Fifth Grade

Christoph Duchhardt[1] *& Annkathrin Gerdes*[1]

[1]*IPN – Leibniz Institute for Science and Mathematics Education, Kiel*

E-mail address of the lead author:

duchhardt@ipn.uni-kiel.de

Bibliographic Data:

Duchhardt, C. & Gerdes, A. (2012): NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

We also thank Natalie Boonyaprasop for proofreading our text in English.

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 3 in Fifth Grade

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span and designs tests for assessing these different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses have been performed based on item response theory (IRT). This paper describes the data on mathematical competence for starting cohort 3 – fifth grade. Besides presenting descriptive statistics for the data, the scaling model applied to estimate competence scores and analyses performed to investigate the quality of the scale, as well as the results of these analyses are also explained. The mathematics test in fifth grade consisted of 25 items which represented different content areas as well as different cognitive components and used different response formats. The test was administered to 5,208 students. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and the tests' dimensionality were evaluated to ensure the quality of the test. The results show that the items exhibited good item fit and measurement invariance across various subgroups. Moreover, the test showed a high reliability. As the correlations between the four content areas are very high in a multidimensional model, the assumption of unidimensionality seems adequate. Among the challenges of this test are the relatively high omission rates in some items and the lack of very difficult items. But overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. This paper describes the data available in the Scientific Use File and provides ConQuest-Syntax for scaling the data.

Keywords

item response theory, scaling, mathematical competence, Scientific Use File

Content

1 Introduction

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

Most of the competence data are scaled using models based on item response theory (IRT). Since most of the competence tests had been developed specifically for implementation in NEPS, several analyses were performed to evaluate the quality of the test. 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 (2012a). This paper presents the results of these analyses for mathematical competence in starting cohort 3.

The present report has been modeled on the technical reports of Pohl, Haberkorn, Hardt, & Wiegand (2012) and Haberkorn, Pohl, Hardt, & Wiegand (2012). Please note that the analyses of this report are based on the data set available at some time before data release. Due to data protection and data cleaning issues, the data set in the Scientific Use File (SUF) may differ slightly from the data set used for analyses in this paper. We do not, however, expect any major changes in results.

2 Testing Mathematical Competence

The framework and test development for the test of mathematical competence are described in Weinert et al. (2011), Neumann et al. (2012) and Ehmke et al. (2009). In the following, we briefly describe specific aspects of the mathematics test that are necessary for understanding the scaling results presented in this paper.

The items are not arranged in units. Thus, in the test, students usually face a certain situation followed by only one task related to it; sometimes there are two, once there are three tasks. Each of the items belongs to one of the following content areas:

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

The framework also describes as a second, independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

In the mathematics test there are three types of response formats. These are simple multiple-choice (MC), complex multiple-choice (CMC), and short constructed response (SCR). In MC items the test taker has to find the correct answer from several, usually four, response options. In CMC tasks a number of subtasks with two response options are presented. SCR items require the test taker to write down an answer into an empty box.

3 Data

3.1 The Design of the Study

Among others, two “life-span” domains were assessed in this study – namely, reading and mathematical competence. In order to control for effects of position and order, the two tests were assigned to test takers in different order. Half of the subjects received a booklet that contained the reading test first followed by the mathematics test in second place, while the other half of the sample received the two tests in the opposite order. The subjects were randomly assigned to one of the two booklets. Note that no multi-matrix design was applied regarding the choice and order of the items *within* a test. All students received the same mathematics items in the same order.

The mathematics test in grade 5 consists of 25 items which represent different content-related and process-related components[1] and use different response formats. Prior to any analyses, one item had been excluded from analyses due to a severe technical problem in the finalization of the test. The characteristics of the remaining 24 items are depicted in the following tables. Table 1 shows the distribution of the four content areas, whereas Table 2 shows the distribution of response formats. The CMC item consists of four subtasks. The SCR items required the subjects to write down either a number (plus unit of measurement, e.g., meter or kilogram) or a single word. Two of these items require two SCR that are closely related. They are treated as one single item and not as a complex one.

Table 1: Content Areas of the Items in the Mathematics Test Grade 5

Content area	**Frequency**
Quantity	8
Space and shape	5
Change and relationships	6
Data and chance	5
Total number of items	24

Table 2: Response Formats of the Items in the Mathematics Test Grade 5

Response format	**Frequency**
Simple multiple-choice	12
Complex multiple-choice	1
Short constructed response	11
Total number of items	24

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

3.2 Sample

A general description of the study and the sample can be found on the NEPS website[2].

5,208 persons took the mathematics test[3]. 15 of these cases had less than three valid responses to the test items (with the CMC item treated as a single item). Since no reliable mathematics competence score may be estimated on the basis of such few responses, these cases were excluded from further analyses. The results of the remaining 5,193 test takers are presented in the following sections.

4 Analyses

In order to carry out first analyses, the SCR items were scored, rating each answer either as correct or wrong or some kind of missing. The two SCR items that required two short answers were scored as correct, if (and only if) both subtasks were correct.

4.1 Missing Responses

There are different kinds of missing responses. These are a) invalid responses, b) missing responses due to omitted items, c) missing responses due to items that have not been reached, d) missing responses due to items that have not been administered, and e) multiple kinds of missing responses that occur within one item and are not determined. In this study, all subjects received the same set of items. As a consequence, there are no items that were not administered to a person. Invalid responses are, for example, selecting two response options in simple MC items where only one is required or simply illegible answers in the SCR format. Missing responses occur when persons skip some items. Due to time limits, it may happen that not every person finishes the test within the given time. Consequently, this results in missing responses due to items that have not been reached.

Missing responses provide information on how well the test worked (e.g., regarding the time limits or understandability of the instructions) and need to be accounted for in the estimation of item and person parameters. We therefore thoroughly investigated the occurrence of missing responses in the test. First we looked at the occurrence of the different types of missing responses per person. This gives an indication on how well the persons were coping with the test. We then examined the occurrence of missing responses per item, in order to get some information on how well the items worked.

4.2 Scaling Model

To estimate item and person parameters for mathematical competence, a partial credit model was used and estimated in ConQuest (Wu, Adams, & Wilson, 1997). A detailed description of the scaling model can be found in Pohl and Carstensen (2012a).

The CMC item consists of a set of subtasks that were aggregated to a polytomous variable, specifying the number of correctly answered subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item was scored as missing. Due to a

[2] www.neps-data.de

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

low category frequency, the lowest two categories (0 or 1 solved subtask, respectively) were collapsed into a single category for this analysis. (Note that, in the Scientific Use File, the values of the polytomously scored CMC item do contain the number of correctly answered subtasks; they cannot directly be interpreted as (partial) credit, cf. Appendix A.)

In the following analyses, each category of the collapsed polytomous item was scored with 0.5 points, while simple MC items and SCR items were scored as 1 (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2012b, for studies on the scoring of different response formats). The two items which require two short answers that are closely related were scored as 1 if, and only if, both parts were solved correctly.

Item difficulties for dichotomous variables and location parameters for polytomous parameters are estimated using the partial credit model. Ability estimates for mathematical competence will be estimated as weighted maximum likelihood estimates (WLEs, Warm, 1989) and later also in the form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl & Carstensen (2012a), while the data available in the SUF are described in section 7. Plotting the item parameters to the ability estimates of the persons had to be done in order to judge how well the item difficulties were targeted to the test persons' abilities. The test targeting gives us some information about the precision of the ability estimates at different levels of ability.

4.3 Checking the Quality of the Scale

The mathematics test had been specifically constructed to be implemented in NEPS. In order to ensure appropriate psychometric properties, the quality of the test was checked by several analyses.

The responses to the subtasks of the CMC item were aggregated to a polytomous variable. In order to justify such an aggregation, the fit of the single subtasks was checked by a first analysis. For this purpose, the single subtasks were separately included in a Rasch model together with all the other items and the fit of the subtasks was evaluated on the basis of the weighted mean square error (WMNSQ), the respective t-value, point biserial correlations of the responses with total correct score, and the item characteristic curve. Only if the subtasks had a satisfactory item fit, were they used to construct the polytomous variable corresponding to the CMC item.

The MC items contain a number of distractors (incorrect response options). We investigated whether the distractors worked well, that is, whether they were chosen by students with a lower ability than those that gave a correct response. To this end, we evaluated the point biserial correlation between giving a certain incorrect response and the total score, thereby treating all subtasks of the CMC item as single items. We judged correlations below zero as good, correlations below 0.05 as acceptable and correlations above 0.05 as problematic.

Item fit was then evaluated for the MC items, the SCR items, and the polytomous CMC item based on results of a partial credit model. Again, the weighted mean square error (WMNSQ), the respective t-value, correlations of the item score with total score, and the item characteristic curve were evaluated for each item. Items with a WMNSQ > 1.15 or WMNSQ < 0.85 were considered as having a noticeable items misfit, and items with a WMNSQ > 1.2 or WMNSQ < 0.8 were judged as a considerable item misfit. Their performance was further investigated. Correlations of the item score with the total score (equal to the discrimination value as computed in ConQuest) greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

We aim at constructing a test of mathematical competence that measures the same construct for all students. If there were any items that favored certain subgroups (e.g., that were easier for males than for females), measurement invariance would be violated, a comparison of competence scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. Test fairness was investigated for the variables test position, gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). In order to test for measurement invariance, differential item functioning was estimated using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty are estimated. Differences in the estimated item difficulties between the subgroups were evaluated. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties that were greater than 1 logit as very strong DIF, absolute differences between .6 and 1 noteworthy of further investigation, and differences smaller than .4 as no considerable DIF. Additionally, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

The competence data in NEPS are scaled using the partial credit model (1PL), in which Rasch-homogeneity is assumed. The partial credit model was chosen because it preserves the weighting of the different aspects of the framework intended by the test developers (Pohl & Carstensen, 2012a). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. We therefore checked for deviations from a uniform discrimination by estimating item discrimination using the generalized partial credit model (2PL)l (Muraki, 1992) and by comparing model fit indices of the 2PL model to those obtained when applying the partial credit model.

The mathematics test has been constructed to measure a unidimensional score of mathematical competence. The assumption of unidimensionality was, nevertheless, tested in the data by a four-dimensional model, the different dimensions being the content areas. Every item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional (MD) model, Gauss quadrature estimation in ConQuest was used (the number of nodes per dimension was chosen in such a way that a stable parameter estimation was obtained). The correlations between the subdimensions were used to evaluate the unidimensionality of the scale.

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

The number of invalid responses per person (counting the CMC item as one item) is shown in Figure 1. The number of invalid responses is very small. In fact, 75% of test persons have no invalid response. Only about 6% of the subjects gave more than one invalid response.

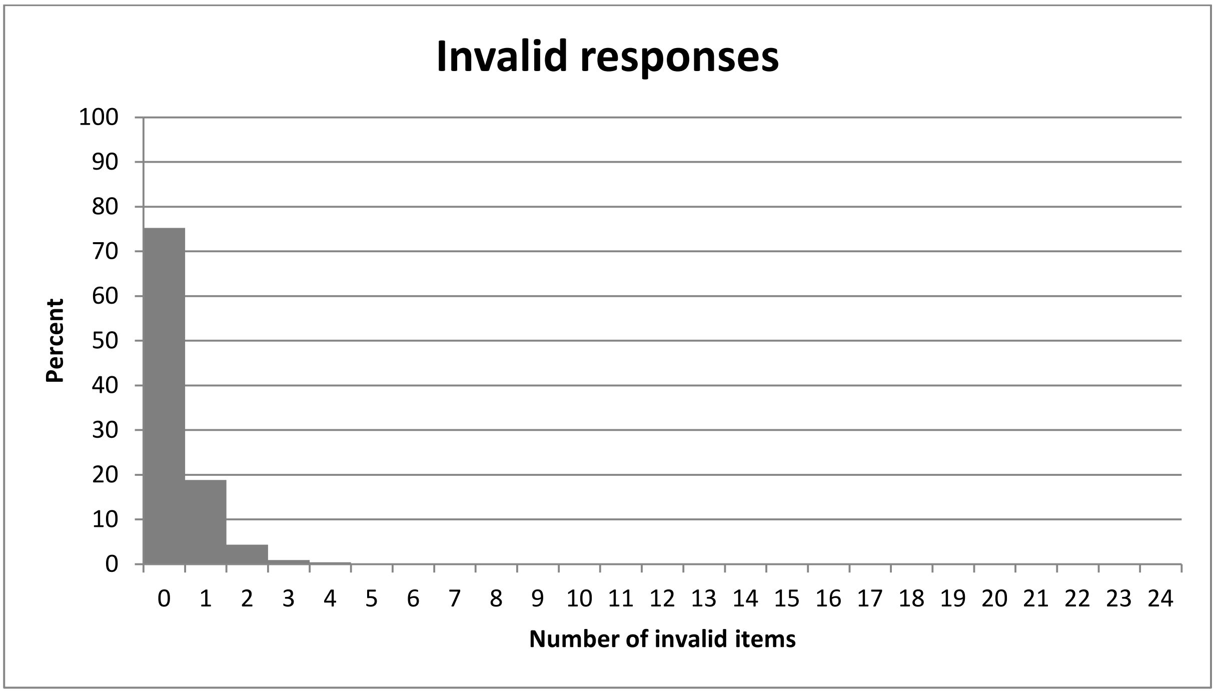

Figure 1: Number of invalid responses

Missing responses may also occur when persons skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. The figure shows that there is some tendency to omit items. However, 53% of the subjects omit no item at all. About 4.5% of the subjects omit more than five items.

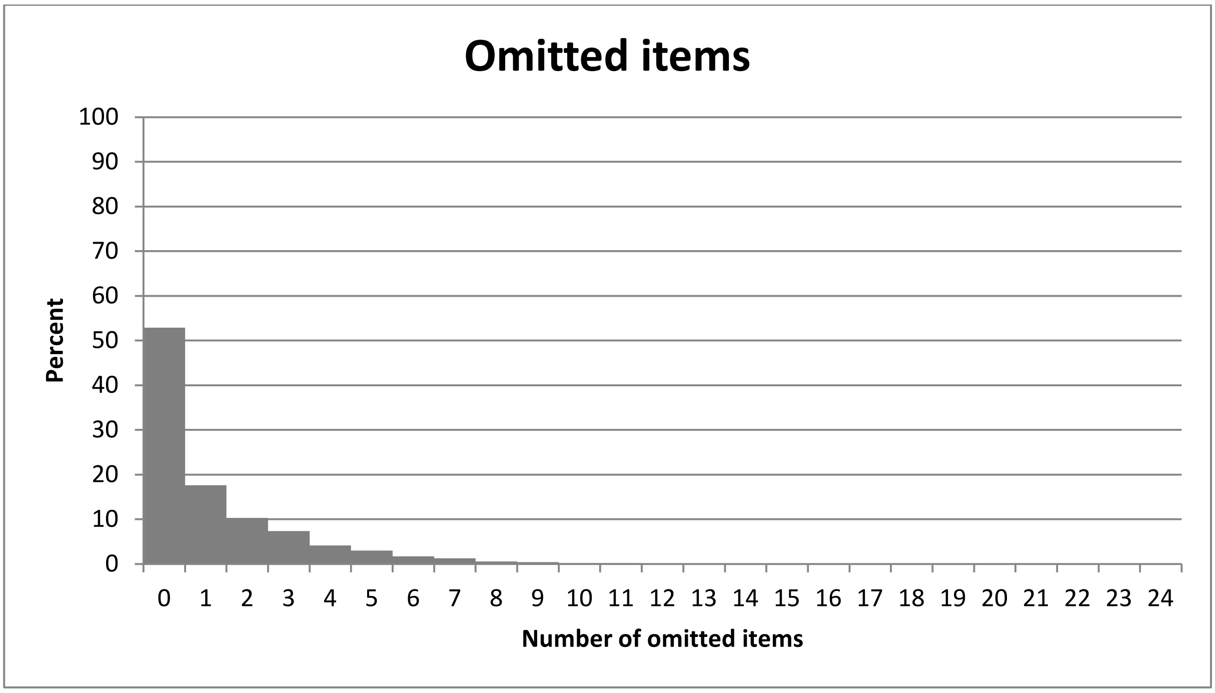

Figure 2: Number of omitted items

All missing responses after the last valid response are defined as not reached. Figure 3 shows the number of items that were not reached by the persons. As can be seen, items that were not reached are quite rare and pose no problem for this test. Only about 9% of the subjects did not reach the end of the test.

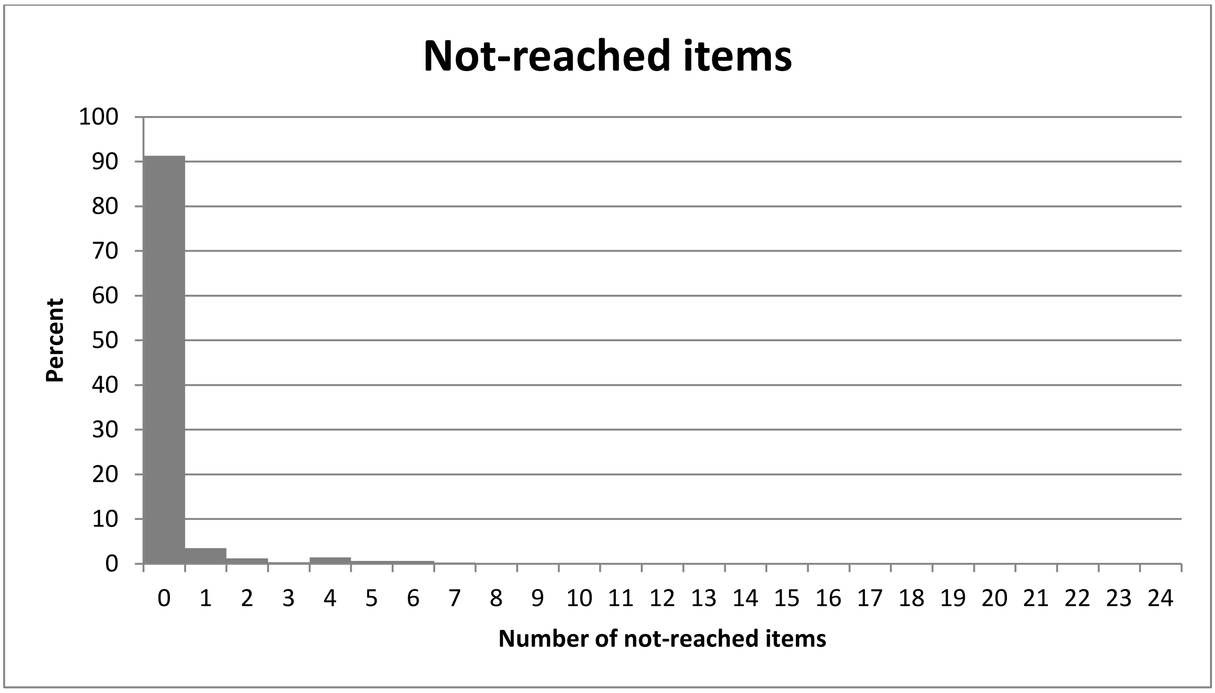

Figure 3: Number of not-reached items

The complex multiple-choice item and the two open items composed of two parts consist of a number of subtasks. Different kinds of missing responses or a mixture of valid and missing responses may occur in these items. The response to such an item was coded as missing when at least one missing response emerged. Basically, when just one kind of missing response occurred, the item response was labeled the same. (The only exception was the CMC item, which was labeled as omitted when some subtasks were answered while others were not reached.) When different kinds of missing responses occurred, the response was labeled as not determinable missing response. This latter case came up only twice (in the CMC item).

Figure 4 shows the total number of missing responses per person. The total number of missing responses is the sum of not valid, omitted, not reached, and not determinable missing responses. Figure 4 shows that only 38.5% of the subjects show no missing response at all. About 10% of the sample show more than five missing responses.

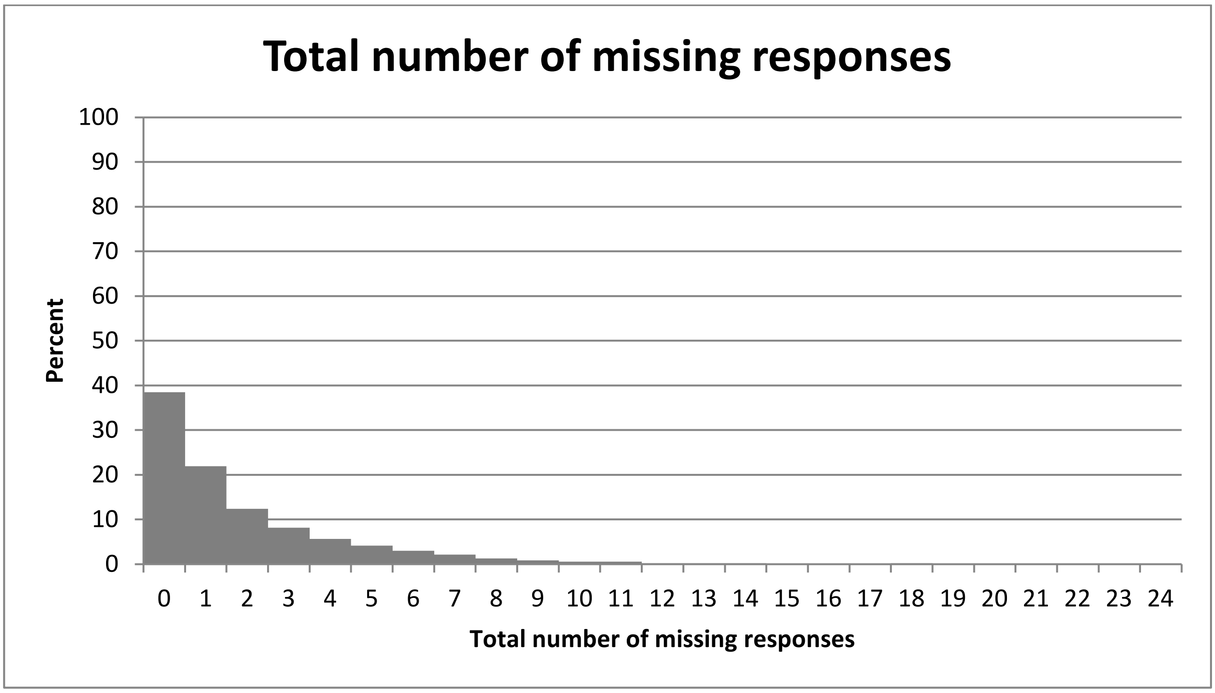

Figure 4: Total number of missing responses

Overall, there is a negligible amount of not-reached items and not-determinable responses and an acceptable amount of omitted items and invalid responses.

5.1.2 Missing responses per item

Table 3 shows the number of valid responses for each item, as well as the percentage of missing responses.

The amount of invalid responses is acceptable. Items with rather many (≥ 3%) invalid responses are mostly short constructed response items (mag5q231_c, mag5d02s_c (composed of two parts), mag5v024_c), but also include one MC item (mag5r191_c), whereas the CMC item (mag5v01s_c) is inconspicuous. One reason for the surprisingly high amount of invalid responses in this particular MC item might be that the question contains a negation.

Items are omitted quite frequently. There are ten items with an omission rate > 5%, four of them being omitted by more than 10%. These are all open items (mag5q231_c, mag5q14s_c (composed), mag5v024_c and mag5v321_c). The maximum omission rate is 14.0%. Also, the items' omission rate is correlated to .35 with their difficulty, indicating that the subjects tend to omit items that are difficult.

The number of persons that did not reach an item increases with the position of the item in the test to up to 8.7%. This is a very small amount.

Table 3: Missing Values in the Items

Item	Position in the test	Number of valid responses	Relative frequency of invalid responses	Relative frequency of omitted missings	Relative frequency of not-reached missings
mag5d041_c	1	5110	0.3	1.3	0
mag5q291_c	2	4919	0.8	4.4	0
mag5q292_c	3	4901	0.7	4.9	0
mag5v271_c	4	4753	0.1	8.4	0
mag5r171_c	5	5002	0.2	3.4	0.0
mag5q231_c	6	4292	5.6	11.7	0.0
mag5q301_c	7	4983	1.8	2.2	0.0
mag5q221_c	8	5007	0.6	3.0	0.0
mag5d051_c	9	5103	0.1	1.6	0.0
mag5d052_c	10	4983	2.4	1.6	0.0
mag5q14s_c	11	4479	2.4	11.2	0.1
mag5q121_c	13[4]	4746	0.2	8.3	0.1
mag5r101_c	14	4986	1.3	2.5	0.2
mag5r201_c	15	5059	0.1	2.2	0.3
mag5q131_c	16	4866	1.3	4.6	0.4
mag5d02s_c	17	4690	3.0	6.1	0.6
mag5d023_c	18	4736	0.9	7.2	0.7
mag5v024_c	19	4186	4.4	14.0	1.0
mag5r251_c	20	4737	0.3	6.9	1.6
mag5v01s_c	21	4580	0.3	9.2	2.3
mag5v321_c	22	4397	1.1	10.6	3.7
mag5v071_c	23	4886	0.6	1.3	4.0
mag5r191_c	24	4660	4.2	0.9	5.2
mag5v091_c	25	4720	0.4	0	8.7

[4] The item in position 12 was excluded from analyses.

5.2 Parameter Estimates

5.2.1 Item parameters

In order to a) get a first rough descriptive measure of item difficulty and b) check for possible estimation problems, we evaluated the relative frequency of given responses before performing IRT analyses. Regarding each subtask of the CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varies between 22.1% and 89.7% across all items. On average, the rate of correct responses is 60.8% (SD = 18.5%). From a descriptive point of view, the items cover a relatively wide range of difficulties.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are depicted in Table 4a. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The step parameters of the polytomous item are depicted in Table 4b. The estimated item difficulties (or location parameters, respectively) vary between -2,549 (item mag5v071_c) and 1,514 (item mag5q121_c) with a mean of -0.5. Overall, the item difficulties are distributed well, yet with a slight tendency to being easy. However, there are no items with a very high difficulty. Due to the large sample size, the standard error of the estimated item difficulties (column x) is very small (SE(ß) ≤ 0.06).

Table 4a: Item Parameters

Item	Position in the test	Difficulty / location parameter	SE of difficulty / location parameter	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimi-nation / 2 PL
mag5d041_c	1	-0,37	0.031	1.02	1.6	0.46	0.92
mag5q291_c	2	-1,04	0.034	0.98	-1.3	0.47	1.06
mag5q292_c	3	-0,799	0.033	1.02	1.4	0.44	0.91
mag5v271_c	4	0,922	0.034	1.06	4.2	0.37	0.70
mag5r171_c	5	-0,082	0.031	1.00	-0.1	0.48	0.99
mag5q231_c	6	0,478	0.034	1.02	1.8	0.44	0.88
mag5q301_c	7	0,603	0.032	0.94	-5.0	0.53	1.31
mag5q221_c	8	-1,855	0.040	1.01	0.3	0.38	0.97
mag5d051_c	9	-2,471	0.047	0.93	-2.3	0.41	1.49
mag5d052_c	10	-0,5	0.032	0.94	-4.9	0.54	1.27
mag5q14s_c	11	-0,721	0.035	0.93	-4.5	0.54	1.27
mag5q121_c	13[5]	1,514	0.038	1.05	2.4	0.33	0.70
mag5r101_c	14	-0,117	0.031	1.13	10.7	0.34	0.55
mag5r201_c	15	-1,225	0.035	1.03	2.0	0.41	0.84
mag5q131_c	16	-1,401	0.037	0.99	-0.3	0.43	1.01
mag5d02s_c	17	-2,094	0.045	0.97	-0.9	0.39	1.16
mag5d023_c	18	-0,428	0.033	1.02	1.5	0.46	0.90
mag5v024_c	19	-0,127	0.034	0.99	-1.1	0.49	1.02
mag5r251_c	20	0,219	0.032	0.99	-0.5	0.48	0.98
mag5v01s_c	21	-1,176	0.032	0.98	-1.3	0.52	1.13
mag5v321_c	22	0,977	0.036	1.00	0.3	0.44	0.92
mag5v071_c	23	-2,549	0.050	1.00	0.0	0.32	0.99
mag5r191_c	24	-0,271	0.033	0.97	-2.3	0.51	1.12
mag5v091_c	25	0,333	0.033	0.99	-0.9	0.48	0.99

Table 4b: Step Parameters of Polytomous Item

Item	Position in the test	location parameter	step 1 (SE)	step 2 (SE)	step 3
mag5v01s_c	21	-1,176	-0.511 (0.030)	0.530 (0.036)	-0.019

[5] The item in position 12 was excluded from analyses.

5.2.2 Person parameters

Person parameters are estimated as WLEs and PVs (Pohl & Carstensen, 2012a). WLEs will be provided in the first release of the SUF. PVs will be provided in later analyses. A description of the data in the SUF can be found in section 7. An overview of how to work with competence data can be found in Pohl and Carstensen (2012a).

5.2.3 Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In these analyses, the mean of ability was constrained to be zero. The variance was estimated to be 1.098, indicating that the test differentiated well between subjects. The reliability of the test (EAP/PV reliability = .802, WLE reliability = .778) is good.

The extent to which the item difficulties and location parameters were targeted toward the test persons' ability is shown in Figure 5. The figure shows that the items cover a wide range of the ability distribution of test persons. However, there are no very difficult items, making the test a little too easy. As a consequence, subjects with a medium and low ability will be measured relatively precisely, while subjects with a high mathematical competence will have a larger standard error.

Scale in logits	Person ability	Item difficulty
3		
	X	
	X	
	XX	
2	XX	
	X	
	XX	
	XX	
	XXXX	13
	XXXX	
	XXXXX	
1	XXXXXXX	
	XXXXX	4 22
	XXXXXXX	
	XXXXXX	7
	XXXXXXXX	6
	XXXXXXXX	25
	XXXXXXXXXX	20
	XXXXXXXX	
0	XXXXXXXX	5 14 19
	XXXXXXXX	24
	XXXXXXXX	1
	XXXXXXXXX	10 18
	XXXXXXXX	
	XXXXXXX	3 11
	XXXXXXX	
-1	XXXXX	2
	XXXXX	15 21
	XXXXX	
	XXXX	16
	XX	
	XXX	
	XX	8
-2	X	
	X	17
	X	
	X	9
		23
-3		

Figure 5: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 31 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item[6] (see Table 4).

[6] Again, item 12 is missing because it was excluded from analyses.

5.3 Quality of the Test

Since the items of the mathematical competence test refer to many different stimuli (there are only two units with two items (or three, respectively) referring to the same stimulus), the assumption of local item independence is plausible.

5.3.1 Fit of the subtasks of the complex multiple-choice item

Before the responses to the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks had been checked by analyzing the subtasks together with the simple multiple-choice and the scored SCR items via a simple Rasch model. There were 27 variables altogether.

The rates of correct responses given to the four subtasks of the CMC item varied from 69.7% to 76.8%. The subtasks showed a good item fit with WMNSQ ranging between 1.01 and 1.09 and the respective t-values between 0.5 and 5.9. Hence, the aggregation of all of the subtasks to one polytomous variable (mag5v01s_c) was considered to be justified.

5.3.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point biserial correlation between each incorrect response (distractor) and the students' total score.

The distractor analysis was performed on the basis of preliminary analyses (see section 5.3.1) treating all subtasks of the CMC item as single items. As the mathematics test uses quite a lot of SCR items (where there are no distractors that can be chosen), Table 5 only shows the summary of point biserial correlations between response and ability for correct and incorrect responses restricted to MC items.

One distractor of an MC item (mag5v271_c) had a positive point biserial correlation of 0.01, one distractor of another MC item (mag5q121_c) had a point biserial correlation of 0. However, after inspecting the two distractors closely, these findings were not considered problematic. All other distractors had a point biserial correlation with the total score below zero. These results indicate that the distractors work reasonably well.

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

Parameter	**Correct responses (MC items only)**	**Incorrect responses (MC items only)**
Mean	0.422	-0.183
Minimum	0.310	-0.340
Maximum	0.540	0.100

5.3.3 Item fit

The item fit is very good. WMNSQ is close to 1 with the lowest value being 0.93 (items mag5d051_c and mag5q14s_c) and the highest being 1.13 (item mag5r101_c). The correlation of the item score with the total score varies between .32 (item mag5v071_c) and .54 (items mag5d052_c and mag5q14s_c) with an average correlation of .44. Almost all item characteristic curves (ICC) showed a good or very good fit of the items. The two items with the highest positive WMNSQs (mag5v271_c and mag5r101_c) showed an acceptable, slightly flat ICC.

5.3.4 Differential item functioning

We examined test fairness to different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables test position, gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). Unlike other cohorts, no DIF on school type was examined here. This is due to the fact that not all Federal States start their school tracking system before grade 5. Table 6 shows the difference between the estimated difficulties of the items in different subgroups. Female versus male, for example, indicates the difference in difficulty *ß*(female) – *ß*(male). A positive value indicates a higher difficulty for females, a negative value a lower difficulty for females compared to males.

Table 6: Differential Item Functioning (Absolute Differences Between Difficulties)

	Booklet	**Gender**	**Books**			**Migration status**		
Item	Position 1 vs Position 2	female vs male	<100 vs >100	<100 vs missing	>100 vs missing	without vs with	without vs missing	with vs missing
mag5d041_c	0.088	-0,342	0.209	0.121	-0.088	-0.152	0.122	0.274
mag5q291_c	-0.172	0,48	0.127	0.179	0.052	0.059	-0.061	-0.12
mag5q292_c	-0.092	-0,054	0.033	0.004	-0.029	0.131	-0.083	-0.214
mag5v271_c	-0.108	-0,002	-0.195	0.048	0.243	0.25	0.284	0.034
mag5r171_c	0.026	-0,15	0.23	0.151	-0.079	-0.143	-0.127	0.016
mag5q231_c	-0.066	0,024	0.127	-0.184	-0.311	-0.004	0.001	0.005
mag5q301_c	0.042	-0,256	0.266	0.316	0.05	-0.132	-0.241	-0.109
mag5q221_c	-0.022	0,098	0.134	0.096	-0.038	-0.059	-0.093	-0.034
mag5d051_c	-0.172	0,108	0.212	0.082	-0.13	-0.153	-0.322	-0.169
mag5d052_c	0.074	-0,098	0.224	-0.035	-0.259	-0.344	-0.301	0.043
mag5q14s_c	-0.142	-0,138	0.093	-0.111	-0.204	-0.001	0.193	0.194
mag5q121_c	0.146	0,006	-0.069	0.23	0.299	-0.071	-0.094	-0.023
mag5r101_c	0.158	0,112	-0.18	0.319	0.499	0.25	0.266	0.016
mag5r201_c	0.148	-0,078	-0.073	-0.232	-0.159	-0.045	0.102	0.147
mag5q131_c	-0.082	0,296	-0.11	0.032	0.142	0.042	-0.321	-0.363
mag5d02s_c	-0.194	0,09	0.013	-0.246	-0.259	-0.092	-0.463	-0.371
mag5d023_c	-0.102	0,22	0.008	0.226	0.218	-0.01	-0.137	-0.127
mag5v024_c	-0.02	0,198	-0.017	0.314	0.331	0.098	0.156	0.058
mag5r251_c	-0.056	-0,124	0.072	0.307	0.235	-0.107	-0.085	0.022
mag5v01s_c	0.08	-0,374	0.214	-0.029	-0.243	-0.023	-0.181	-0.158
mag5v321_c	-0.016	0,302	0.002	0.241	0.239	-0.143	0.287	0.43
mag5v071_c	-0.034	-0,124	0.084	0.48	0.396	0.446	0.196	-0.25
mag5r191_c	0.084	-0,188	0.084	0.174	0.09	-0.157	-0.222	-0.065
mag5v091_c	0.162	0,326	0.103	0.209	0.106	-0.12	0.118	0.238
Main effect	**-0.046**	**0.294**	**0.668**	**-0.156**	**-0.824**	**-0.641**	**-0.612**	**0.029**

The mathematical competence test was administered in two different positions (see section 3.1 for the design of the study). 2,587 (49.8%) persons received the mathematics test first and then the reading test; 2,606 (50.2%) persons received the mathematics test after having completed the reading test. The subjects were randomly assigned to either of the two design groups. There are, however, almost no average differences between the two design groups. Subjects who received the mathematics test before the reading test perform, on average, 0.046 logits (Cohen's d = 0.044) better than subjects who received the mathematics test after the reading test. Differential item functioning of the position of the test might occur, for example, if there are differential fatigue effects for certain items. But there is also no considerable DIF due to the position of the test in the study design. The highest DIF between the two design groups is 0.194 logits.

In total, 2,512 (48.4%) of the test takers were female and 2,679 (51.6%) were male. Two missing responses were given in relation to the variable gender. These cases were excluded from the DIF analysis. On average, male students showed a higher mathematical competence than female students (main effect = 0.294 logits, Cohen's d = 0.283). There was no item with a considerable gender DIF; the only item for which the difference in item difficulties between the two groups exceeded 0.4 logits was item mag5q291_c (0.480 logits).

The number of books at home was used as a proxy for socioeconomic status. There were 2,144 (41.3%) test takers with 0 to 100 books at home, 2,721 (52.4%) test takers with more than 100 books at home, and 328 (6.3%) test takers with a missing response in relation to this variable. Group differences and DIF were investigated by using these three groups. There are considerable average differences between the three groups. Participants with 100 or less books at home perform on average 0.668 logits (Cohen's d = 0.676) lower in mathematics than participants with more than 100 books. Participants without a valid response in relation to the variable books at home performed 0.156 logits (Cohen's d = 0.158) or 0.824 logits (Cohen's d = 0.834) worse than participants with up to 100 and more than 100 books, respectively. There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.266 logits). Comparing the group without valid responses to the two groups with valid responses, DIF exceeding 0.4 logits occurs in two items (mag5r101_c and mag5v071_c), the maximum being 0.499 logits.

There were 3,535 (68.1%) participants without migration background, 1,303 (25.1%) participants with migration background, and 355 (6.8%) participants without a valid response. All three groups were used for investigating DIF of migration. On average, participants without migration background performed considerably better in the mathematics test than those with migration background (main effect = 0.641 logits, Cohen's d = 0.638). Also, subjects with missing values for migration differ from those without migration background (main effect = 0.612 logits, Cohen's d = 0.609). Here, too, participants without migration background show a higher mathematical competence. Subjects with migration background performed slightly worse compared to participants with missing values for migration (main effect = 0.029 logits, Cohen's d =0.029). There is no considerable DIF comparing the three groups. Differences in item difficulties exceeding 0.4 logits were observed in items mag5d02s_c, mag5v321_c, and mag5v071_c, the maximum being 0.463 logits.

In Table 7, the models including main effects only are compared with those that additionally estimate DIF. Akaike's (1974) information criterion (AIC) favors the models estimating DIF for all four DIF variables. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters into account more strongly and, thus, prevents from overparametrization of models. Using BIC, the more parsimonious models including only the main effects of the position of the test, number of books, and migration status, respectively, are preferred over the more complex respective DIF models. However, BIC prefers the model including both main effect and DIF effect of gender to the model including the gender main effect only. (Note that the analyses including gender contain fewer cases, thus the information criteria cannot be compared across analyses with different DIF variables.)

Table 7: Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	132410.214	28	132466.214	132649.756
	DIF	132354.456	52	132458.456	132799.319
Gender	main effect	132273.559	28	132329.559	132513.090
	DIF	132033.776	52	132137.776	132478.619
Books	main effect	131934.698	29	131992.698	132182.795
	DIF	131816.193	77	131970.193	132474.933
Migration	main effect	132065.207	29	132123.207	132313.304
	DIF	131936.188	77	132090.188	132594.928

5.3.5 Rasch-homogeneity

In order to test for the assumption of Rasch-homogeneity, we also fit a generalized partial credit model (2PL) to the data. The estimated discrimination parameters are depicted in Table 4a. They range from 0.55 (item mag5r101_c) to 1.49 (item mag5d051_c), most of them (17 out of 24) are very close (between 0.8 and 1.2) to 1. Nevertheless, the 2PL model (AIC = 132046.23, BIC = 132446.09, number of parameters = 61) fits the data better than the partial credit model (1PL) (AIC = 132466.18, BIC = 132843.17, number of parameters = 27). Nevertheless, the theoretical aim was to construct a test that represents the different aspects of the framework equally (see Pohl & Carstensen, 2012a, 2012b, for a discussion of this issue), and, thus, the partial credit model was used to model the data and to estimate competence scores.

5.3.6 Unidimensionality

The unidimensionality of the test was investigated by specifying a four-dimensional model based on the four different content areas. Every item was assigned one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Gauss-Hermite quadrature estimation implemented in ConQuest was used. The number of nodes per dimension was chosen in such a way that a stable parameter estimation was obtained. The variances and correlations of the four dimensions are shown in Table 8.

Table 8: Results of Four-Dimensional Scaling. Variances of Dimensions are Depicted in the Diagonal, Correlations Are Given in the Off-Diagonal.

	Dim 1	Dim 2	Dim 3	Dim 4
Quantity (8 items)	1.327			
Space and shape (5 items)	0.894	0.978		
Change and relationships (6 items)	0.908	0.943	1.115	
Data and chance (5 items)	0.872	0.933	0.942	1.532

All four dimensions show a substantial variance. The correlation between the four dimensions is – as expected – high, varying between .87 and .94.

Model fit between the unidimensional model and the four-dimensional model is compared in Table 9.

Table 9: Comparison of the Unidimensional and the Four-Dimensional Model.

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	132412.185	27	132466.185	132643.171
Four-dimensional	132279.004	36	132351.004	132586.986

6 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 3 and at describing how the mathematics competence score had been estimated.

Fortunately, the amount of invalid responses and not-reached items is rather low. Some items show higher omission rates, although, in general, the amount of omitted items is acceptable, too.

The test has a good reliability (EAP/PV-reliability = .802, WLE reliability = .778). It distinguishes well between test takers, indicated by the test's variance (= 1.098). However, very difficult items are missing, hence, test targeting is somewhat suboptimal. The test measures mathematical competence of high-performing students a little less accurately.

Indicated by various fit criteria – WMNSQ, t-value of the WMNSQ, ICC – the items exhibit a good item fit. Also, discrimination values of the items (either estimated in a 2PL model or as a correlation of the item score with total score) are acceptable. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test is fair to the considered subgroups.

Fitting a four-dimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas) yields a slightly better model fit than the unidimensional partial credit model. However, high correlations of about 0.9 between the four dimensions indicate that a unidimensional model describes the data reasonably well.

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

7 Data in the Scientific Use File

There are 24 items in the data set that are either scored as dichotomous variables (MC and SCR items), with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (corresponding to the CMC item) indicating the number of correctly answered subtasks. The dichotomous variables are marked with a '_c' behind their variable name, whereas the polytomous variable is marked with a 's_c' behind its variable name. Please note that for the purpose of this analysis, the two lowest categories of this polytomous variable have been collapsed (see section 4.2 on the aggregation of CMC items). In the scaling model, the collapsed polytomous variable is scored in steps of 0.5 – 0 for the lowest category, 1.5 denoting the highest. Manifest scale scores are provided in the form of WLE estimates (ma_sc1) including the respective standard error (ma_sc2). Also, note that for the estimation of the WLE scores, the effect of test position in the booklet has been controlled for. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. Students that did not take part in the test or those that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on their WLE score for reading competence.

Plausible values that allow us to investigate latent relationships of competence scores with other variables will be provided in later data releases. Users interested in investigating latent relationships may alternatively either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012a).

References

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

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 (313-327). Münster: Waxmann.

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

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). Technical report of reading– Scaling results of starting cohort 4 in ninth grade (NEPS Working Paper No. x). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

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

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2012). Modeling and assessing of mathematical competence over the lifespan. Manuscript submitted for publication.

Pohl, S. & Carstensen, C. H. (2012a). Scaling the data of the competence tests. NEPS Working Papers.

Pohl, S. & Carstensen, C. H. (2012b). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. Manuscript submitted for publication.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6 (2)*, 461–464.

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

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

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

Appendix

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

```
Title Starting Cohort III, MATHEMATICS: Partial Credit Model;

data filename.dat;
format pid 4-10 booklet 12 responses 14-37; /* insert number of columns with data*/

labels << filename_with_labels.txt;

codes 0,1,2,3,4;

recode (0,1,2,3,4) (0,0,1,2,3) !item (20); /* collapsing the lowest categories */

score (0,1) (0,1)                       !items (1-19,21-24);
score (0,1,2,3) (0,0.5,1,1.5)           !item (20);

set constraint=cases;

model item + item*step + booklet;
estimate;

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