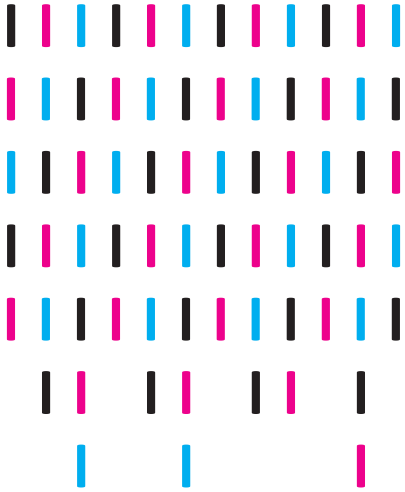

NEPS SURVEY PAPERS

Anna-Lena Gerken and Insa Schnittjer

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 5 FOR FIRST-YEAR STUDENTS

NEPS Survey Paper No. 17
Bamberg, January 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

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

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

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

NEPS Technical Report for Mathematics:

Scaling Results of Starting Cohort 5 for First-Year Students

Anna-Lena Gerken and Insa Schnittjer

Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the lead author:

gerken@ipn.uni-kiel.de

Bibliographic Data:

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. doi:10.5157/NEPS:SP17:1.0

Acknowledgements:

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

The present report has been modeled along previous reports published by the NEPS. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Duchhardt, 2015; Haberkorn, Pohl, Hardt, & Wiegand, 2012; Jordan & Duchhardt, 2013; Koller, Haberkorn, & Rohm, 2014; Pohl, Haberkorn, Hardt, & Wiegand, 2012).

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 5 for First-Year Students

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and designs tests for assessing these different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedure for the mathematical competence test in first-year students of starting cohort 5. The mathematics test contained 21 items with different response formats representing different content areas and different cognitive components. The test was administered to 5,915 first-year students. Their responses were scaled using the partial-credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test's dimensionality were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the large number of items targeted toward a lower mathematical ability as well as the relatively high omission rates in some items. Overall, the mathematics test had acceptable psychometric properties that allowed for an estimation of reliable mathematics competence scores. Besides the scaling results, this paper also 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 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 competence domains measured in the NEPS is given by Weinert et al. (2011).

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the 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 (2012).

In this paper, the results of these analyses are presented for mathematical competence in first-year students of starting cohort 5 (students). First, the main concepts of the mathematical test are introduced. Then, the mathematical competence data of starting cohort 5 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 of this report are based on the dataset available some time before data release. Due to data protection and data cleaning issues, the data in the Scientific Use File (SUF) may differ slightly from the data used for the analyses in this paper. However, fundamentally different results are not expected.

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. (2013), 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 tasks. Each of the items belongs to one of the following content areas:

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

Each item was constructed in such a way as to primarily address a specific content area. The framework also describes as a second and 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

The study assessed different competence domains including, among others, reading competence and mathematical competence. 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 order. Half of the subjects received a booklet that contained the reading test first followed by the mathematics test, 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. There was no multi-matrix design regarding order of the items *within* the mathematics test. All subjects received the same mathematics items in the same order. The test was administered as a group test in rooms at different universities.

The mathematics test for first-year students consisted of 21 items which represented different content-related and process-related components and used different response formats. One item (mas1q051_c) was eliminated from further analysis because of differential item functioning with regard to gender (see 4.3.3 for an explanation).The characteristics of the final set of 20 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.

Table 1: Number of Items by Content Areas

Content area	Frequency
Quantity	4
Space and shape	4
Change and relationships	6
Data and chance	6
Total number of items	20

Table 2: Number of Items by Response Formats

Response format	Frequency
Simple Multiple-Choice	16
Complex Multiple-Choice	1
Short-constructed response	3
Total number of items	20

3.2 Sample

A total of 5,915[1] students received the mathematics test. For ten of them less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 5,905 test takers. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and finally e) multiple kinds of missing responses within CMC items that are not determined.

In this study, all respondents received the same set of items. As a consequence, there are no items that were not administered to a person. Invalid responses occurred, for example, when two response options were selected where only one was required or when simply illegible answers were provided in the SCR format. Omitted items occurred when test takers skipped some items. Due to time limits not all persons finished the test within the given time limit. All responses after the last valid response were coded as not reached. As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. 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). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well the items functioned.

3.4 Scaling Model

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

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC item was scored as missing.

[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.

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. For item mas1q02s_c categories were collapsed.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Mathematical competencies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in the form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 6.

3.5 Checking the Quality of the Scale

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

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

The MC items consisted of one correct response option and one or more distractors (incorrect response options). The quality of the distractors within MC items was evaluated using the point-biserial correlation of selecting an incorrect response and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.2 (*t*-value > |8|) were judged as a considerable item misfit, and their performance was investigated further. Correlations of the item score with the total correct 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.

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 the subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, the position of the test within the test battery, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) was examined, using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, 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 competence data in the NEPS are scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the mathematics test was evaluated by specifying a four-dimensional model based on the four content areas. Every item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Monte Carlo estimation in ConQuest was used (the number of nodes per dimension was chosen in such a way as to obtain stable parameter estimates). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The IRT models were estimated in ConQuest version 2.0 (Wu, Adams, & Wilson, 1997). The 2PL model was estimated in MDLTM (Matthias von Davier, 2005).

4. Results

4.1 Missing Responses

4.1.1 Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was very small. In fact, 97.6% of test takers gave no invalid response at all. Less than 3% of the respondents had more than one invalid response.

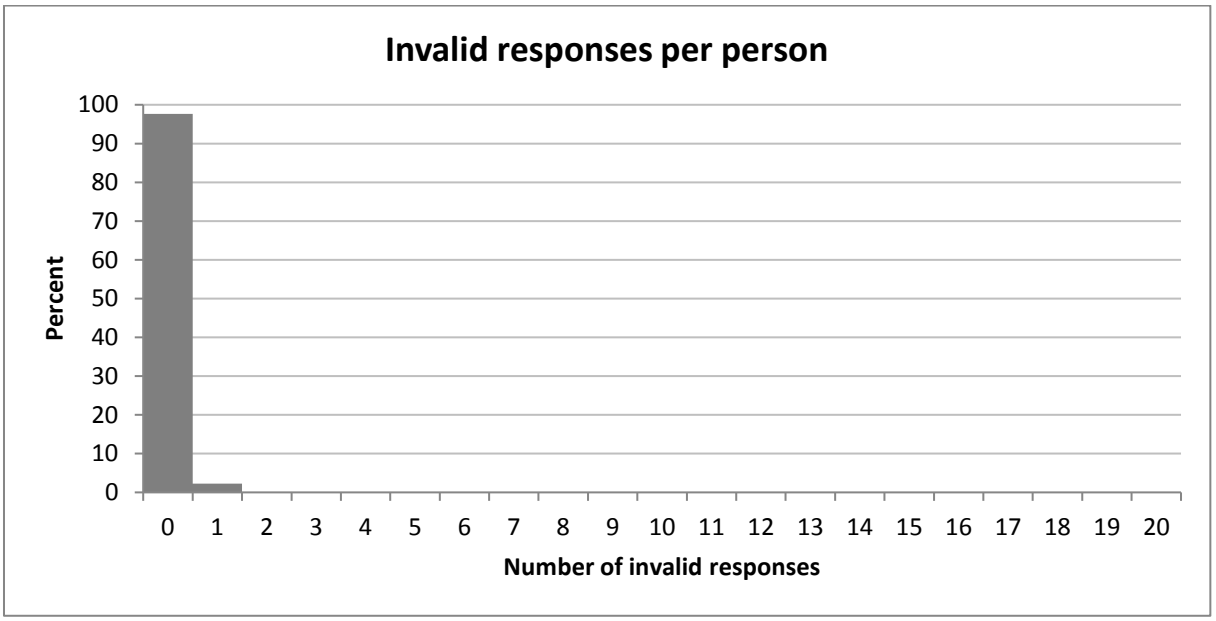

Figure 1: Number of invalid responses

Missing responses may also occur when test takers skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. It shows that 40.4% of the respondents omitted no item at all, whereas 4.6% of the respondents omitted more than five items.

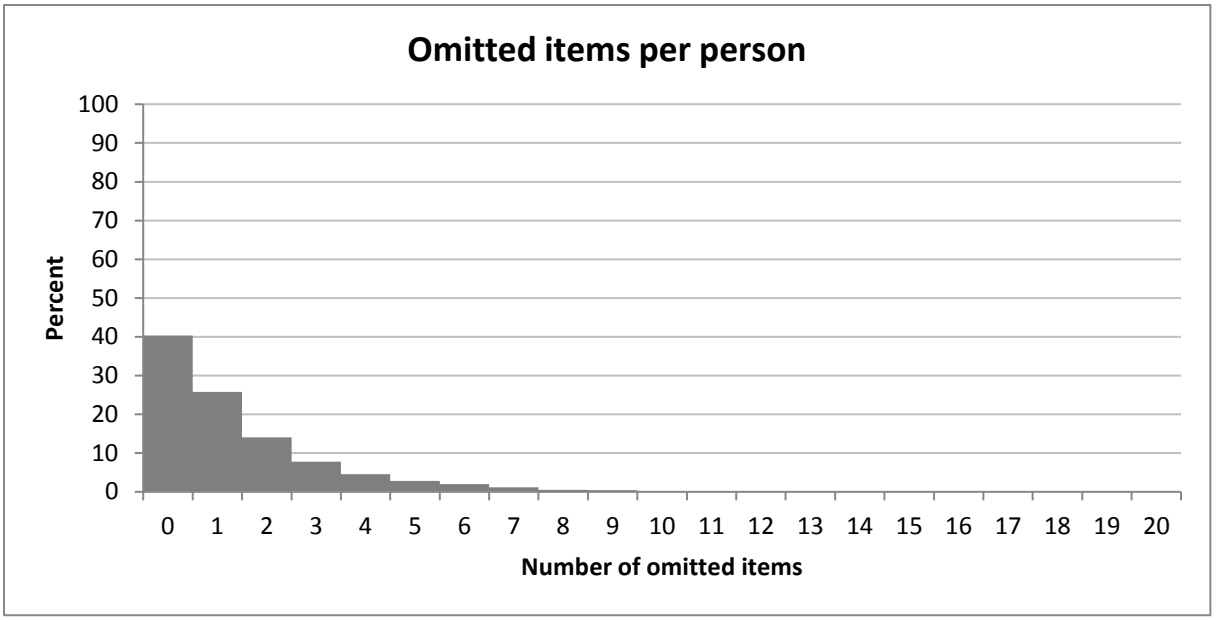

Figure 2: Number of omitted items

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 a person. As can be seen, only 62.2% reached the end of the test, whereas 26.6% of the test takers did not reach one to five items. Only 11.2% of the subjects did not reach more than five items.

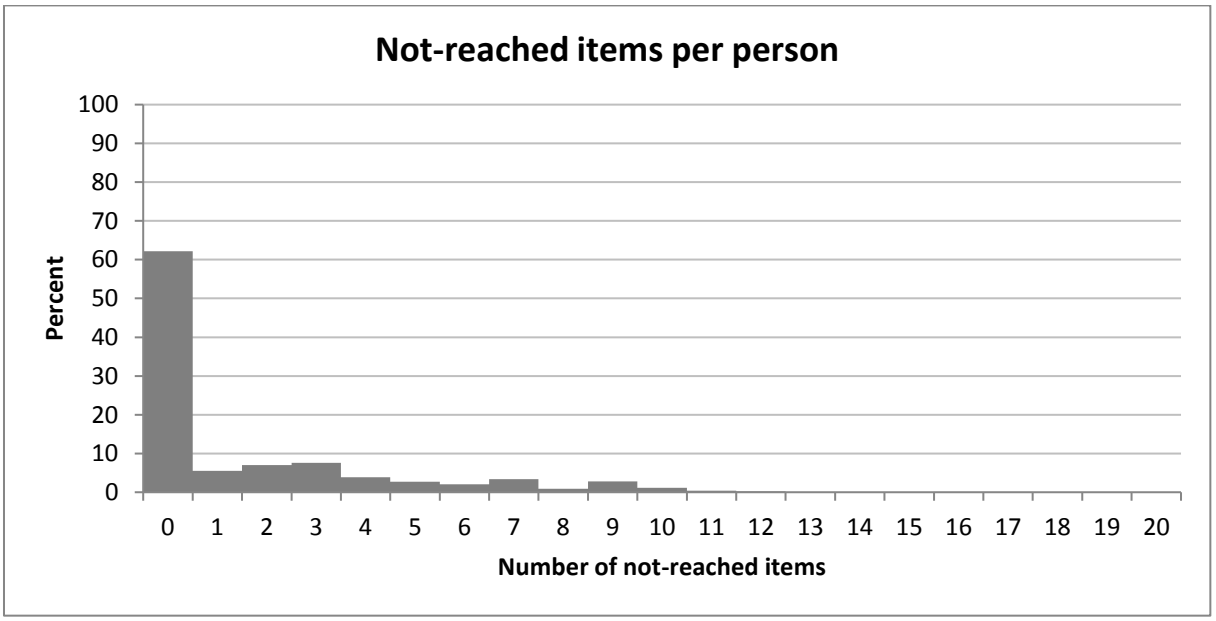

Figure 3: Number of not-reached items

Figure 4 shows the total number of missing responses per person, which is the sum of invalid, omitted, not-reached, and not-determinable missing responses. In total, 28.2% of the test takers showed no missing response at all, whereas 22.1% showed more than five missing responses.

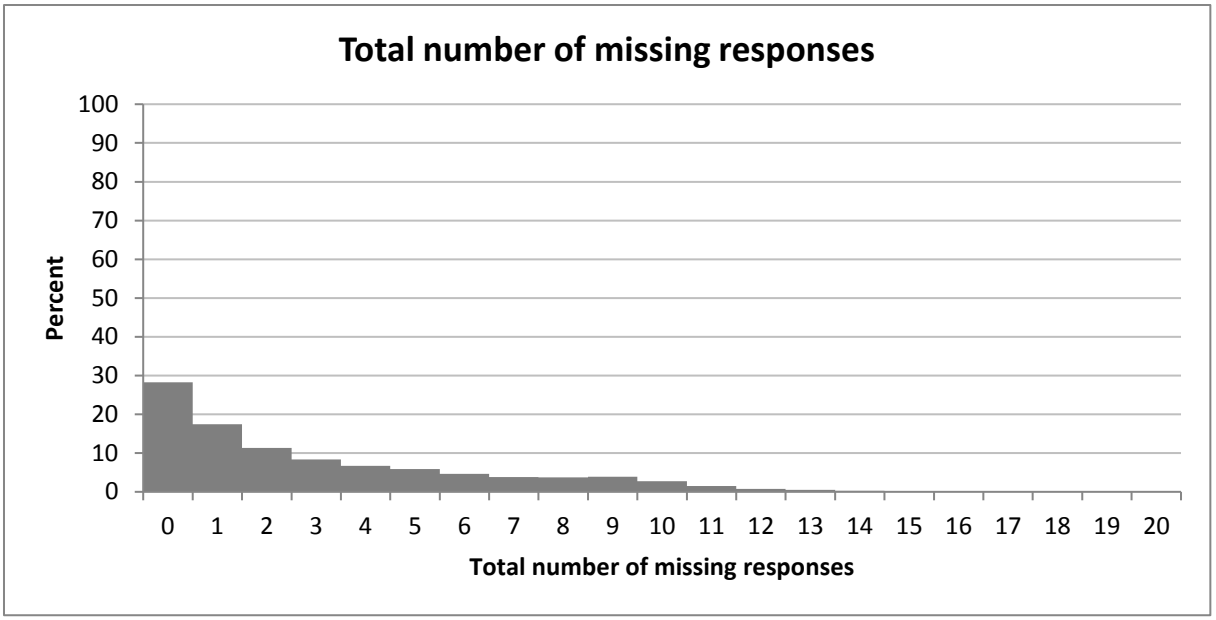

Figure 4: Total number of missing responses

Overall, there was a negligible amount of invalid, and a reasonable amount of not-reached or omitted items.

4.1.2 Missing responses per item

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

Overall, the number of invalid responses per item was very small. The omission rates were acceptable, except for three items with an omission rate higher than 10%. The highest omission rate (41.95%) occurred for item mas1v042_c. As this item is a SCR item, the subjects might have preferred to skip this item rather than to guess. Furthermore, this item is one of the most difficult ones. In this test, subjects tended to omit difficult items, supposedly because they did not to know the answer. They preferred to skip these items rather than to guess an answer.

The number of persons that did not reach an item increased with the position of the item in the test up to 37.78%.

The total number of missing responses per item varied between 1.3% (item maa3d131_sc5s1_c) and 48.09% (item mas1v042_c).

Table 3: Percentage of Missing Values

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted responses	Percentage of not-reached items
maa3q071_sc5s1_c	1	5,778	0.00	2.15	0.00
mas1r092_c	2	5,746	0.00	2.69	0.00
mas1v093_c	3	5,635	0.00	4.57	0.00
mas1v032_c	5	5,531	0.03	6.27	0.02
maa3d131_sc5s1_c	6	5,828	0.02	1.24	0.05
maa3d132_sc5s1_c	7	5,635	0.02	4.50	0.05
mas1v062_c	8	4,981	0.56	14.94	0.15
mas1v063_c	9	5,608	0.05	4.78	0.20
maa3r081_sc5s1_c	10	5,348	0.00	8.98	0.46
maa3v082_sc5s1_c	11	5,223	0.05	10.62	0.85
mas1q041_c	12	5,271	0.02	8.74	1.98
mas1v042_c	13	3,065	0.95	41.95	4.79
mas1q02s_c	14	5,047	0.24	8.45	5.72
maa3d111_sc5s1_c	15	5,239	0.00	2.18	9.09
maa3d112_sc5s1_c	16	4,917	0.00	5.57	11.16
maa3r011_sc5s1_c	17	4,938	0.00	2.54	13.84
mas1q011_c	18	4,685	0.17	2.78	17.71
mag9r061_sc5s1_c	19	3,854	0.34	9.13	25.27
mas1d071_c	20	3,893	0.05	1.74	32.28
mas1d072_c	21	3,665	0.02	0.14	37.78

Note. The item on position 4 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

4.2 Parameter Estimates

4.2.1 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 the responses given before performing any IRT analyses. Using each subtask of a CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varied between 19.34% and 92.52% across all items. On average, the rate of correct responses was 62.19% (*SD* = 20.08%). From a descriptive point of view, the items covered 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 varied between -2.458 (item mas1q02s_c) and 1.748 (item mas1r092_c) with a mean of -0.466. Due to the large sample size, the standard errors of the estimated item difficulties (Table 4a) were very small, *SE*(ß) ≤ 0.04.

Table 4a: Item Parameters

Item	Position	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.
maa3q071_sc5s1_c	1	-1.206	0.032	1.03	2.0	0.43	0.91
mas1r092_c	2	1.748	0.036	1.00	0.0	0.39	0.84
mas1v093_c	3	-0.822	0.031	0.96	-2.8	0.52	1.23
mas1v032_c	5	0.450	0.030	1.06	5.2	0.42	0.69
maa3d131_sc5s1_c	6	-2.203	0.041	1.01	0.3	0.36	1.02
maa3d132_sc5s1_c	7	-0.646	0.031	0.91	-7.2	0.59	1.55
mas1v062_c	8	0.461	0.032	1.10	7.4	0.39	0.63
mas1v063_c	9	0.344	0.030	0.99	-1.1	0.50	0.99
maa3r081_sc5s1_c	10	-0.863	0.032	0.97	-2.4	0.52	1.22
maa3v082_sc5s1_c	11	-0.472	0.031	1.02	1.5	0.48	0.96
mas1q041_c	12	-0.768	0.032	1.02	1.3	0.47	0.93
mas1v042_c	13	0.710	0.041	0.93	-4.2	0.55	1.16
mas1q02s_c	14	-2.458	0.037	0.97	-1.3	0.47	0.65
mas1081_c	15	-1.279	0.034	1.02	1.3	0.43	0.98
maa3d112_sc5s1_c	16	0.147	0.032	0.99	-0.6	0.51	1.02
maa3r011_sc5s1_c	17	-1.212	0.035	0.92	-4.7	0.54	1.58
mas1q011_c	18	-1.754	0.040	0.94	-2.6	0.48	1.57
mag9r061_sc5s1_c	19	-0.522	0.037	1.05	3.4	0.43	0.76
mas1d071_c	20	-0.147	0.036	1.07	4.7	0.44	0.75
mas1d072_c	21	1.171	0.041	1.09	4.6	0.35	0.57

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, rit = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).

Item 4 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

Table 4b: Step Parameters of Polytomous Item

Item	Position in the test	step 1 (*SE*)	step 2 (*SE*)	step 3
mas1q02s_c	14	-0.614 (0.032)	0.719 (0.038)	-0.104

4.2.2 Person parameters

Person parameters are estimated as WLEs and PVs (Pohl & Carstensen, 2012). 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 6. An overview of how to work with competence data can be found in Pohl and Carstensen (2012).

4.2.3 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 1.166, indicating that the test differentiated well between subjects. The reliability of the test (EAP/PV reliability =.761, WLE reliability = .720) 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 items cover a wide range of the ability distribution of test persons. However, there are no very difficult items. As a consequence, subjects with a low or medium ability will be measured relatively precisely, while subjects with a high mathematical competence will have a larger standard error.

Scale in logits	Person ability	Item difficulty
3	X	
	X	
	X	
	XX	
	XX	
2	XX	
	XX	2
	XX	
	XXXX	
	XXXX	
	XXXXX	21
	XXXXXXX	
1	XXXXX	
	XXXXXX	
	XXXXXXX	13
	XXXXXXXX	5 8
	XXXXXXXXX	9
	XXXXXXXXXX	16
	XXXXXXXX	
0	XXXXXXXX	
	XXXXXXXX	20
	XXXXXXXXX	
	XXXXXXXXX	11 19
	XXXXXXXXX	7
	XXXXXXX	3 10 12
-1	XXXXXXX	
	XXXX	
	XXXXX	1 15 17
	XXXXX	
	XXX	
	XX	
	XX	18
-2	XX	
	X	
	X	6
	X	
	X	14
-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 34.6 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4a).

4.3 Quality of the test

4.3.1 Distractor analyses

To investigate how well the distractors performed in the test, we evaluated – for the MC items – the point-biserial correlations between selecting each incorrect response (distractor) and the students' total correct scores. This distractor analysis was performed on the basis of preliminary analyses treating all subtasks of the CMC item as single items. The point-biserial correlations for the distractors ranges from -0.45 to -0.05 with a mean of -0.20. These results indicate that the distractors worked well. In contrast, the point-biserial correlations between the correct response and student's total scores range from 0.34 to 0.58 with a mean of 0.45 indicating that more proficient students were also more likely to identify the correct response option.

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

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.45	-0.20
Minimum	0.34	-0.45
Maximum	0.58	-0.05

4.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC and polytomous CMC items. Altogether, item fit can be considered to be very good (see Table 4a). Values of the WMNSQ were close to 1 with the lowest value being 0.91 (item maa3d132_sc5s1_c) and the highest being 1.10 (item mas1v062_c). Thus, there was no indication of severe item over- or underfit. The correlations of the item score with the total score varied between 0.35 (item mas1d072_c) and 0.59 (item maa3d132_sc5s1_c) with an average correlation of 0.46. Almost all item characteristic curves (ICC) showed a good or very good fit of the items. The item with the highest WMNSQs (item mas1v062_c) showed a slightly flat ICC.

4.3.3 Differential item functioning

We examined test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables gender, the position of the test, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 6 shows the differences between the estimated difficulties of the items in different subgroups. For example, female versus male indicates the difference in difficulty between women and men, β(female) – β(male). A positive value indicates a higher difficulty for females, whereas a negative value shows a lower difficulty for females as compared to males.

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

Item	Position in the test	Gender	Position	Migration status
		female vs. male	second vs. first	without vs. with
maa3q071_sc5s1_c	1	0.156	-0.128	-0.106
mas1r092_c	2	-0.276	-0.050	-0.036
mas1v093_c	3	-0.222	-0.110	-0.308
mas1v032_c	5	0.444	0.026	-0.014
maa3d131_sc5s1_c	6	0.362	0.168	-0.162
maa3d132_sc5s1_c	7	-0.094	0.030	0.104
mas1v062_c	8	0.250	0.020	0.820
mas1v063_c	9	-0.180	0.030	-0.286
maa3r081_sc5s1_c	10	-0.254	-0.056	-0.070
maa3v082_sc5s1_c	11	-0.142	0.074	0.140
mas1q041_c	12	-0.076	-0.036	-0.184
mas1v042_c	13	-0.216	0.128	0.012
mas1q02s_c	14	-0.044	0.114	-0.254
mas1081_c	15	0.314	-0.122	-0.322
maa3d112_sc5s1_c	16	0.144	0.154	0.064
maa3r011_sc5s1_c	17	-0.562	-0.080	0.250
mas1q011_c	18	-0.448	0.018	0.008
mag9r061_sc5s1_c	19	0.182	-0.046	0.188
mas1d071_c	20	0.244	-0.072	0.024
mas1d072_c	21	0.412	0.010	0.186
Main effect (Model with DIF)		**0.796**	**0.016**	**-0.426**
Main effect (Model without DIF)		**0.804**	**0.016**	**-0.420**

Overall, 2,042 (34.6%) of the test takers were male and 3,852 (65.2%) were female. The remaining 11 (0.2%) participants did not give an answer. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.796 logits, Cohen's *d* = 0.790). After excluding item 4 (mas1q051_c), there was no item with a considerable sex DIF. For four items the difference in item difficulties between the two groups exceeded 0.4 logits, the maximum being item maa3r011_sc5s1_c (0.56 logits).

The test takers received either the mathematic or the reading test first. If this resulted in a position effect it was analyzed through a second DIF analysis. There were 2,933 (49.7%) subjects who solved the mathematics test first and 2,973 (50.3%) students who received the mathematics test second after the reading test. There was no considerable mean difference between the two groups (main effect = 0.016 logits, Cohen's *d* = 0.015). There was also no considerable DIF for the items comparing participants with the different test position.

There were 5,647 (95.6%) participants without migration background, 154 (2.6%) participants with migration background, and 104 (1.8%) participants without a valid response. Only the first two groups were used for investigating DIF of migration. On average, participants without a migration background performed considerably better in the mathematics test than those with a migration background (main effect = -0.426 logits, Cohen's *d* = -0.396). There was only one item (mas1v062_c) with a considerable DIF (0.82 logits). The other items show no considerable DIF.

In Table 7, we compared the models that included only main effects on the three variables to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the model estimating DIF for the variable gender. For the variables migration status and test positon AIC preferred models that only included the main effects. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents from overparametrization of models. Using the BIC, the more parsimonious models including only the main effects of the migration status and test position were preferred over the more complex DIF models. For the variable gender BIC also preferred the model including both main and DIF effects.

Table 7: Comparisons of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	main effect	115,308.72	24	115,356.72	115,517.08
	DIF	115,031.86	44	115,119.86	115,413.85
Migration status	main effect	114,018.70	24	114,066.70	114,226.68
	DIF	113,988.46	44	114,076.46	114,369.75
Position	main effect	114,036.84	24	114,084.84	114,244.82
	DIF	114,006.19	44	114,094.19	114,387.48

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (2PL) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4a), ranging from 0.57 (item

mas1d072_c) to 1.58 (items maa3r011_sc5s1_c). The average discrimination parameter fell at 1.00. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 115,602.42072, BIC = 115,963.33267, number of parameters = 54) as compared to the 1PL model (AIC = 116,321.25958, BIC = 116,568.55110, number of parameters = 37). Despite the empirical preference for the 2PL model, the 1PL model 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 partial credit model (1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

Note that these calculations could not be made by conquest 2.0 so that we had to use a substitute Program called MDLTM (see Davier, 2005). As a consequence, the results for AIC and BIC using the 1PL model might differ from the later results (see 4.3.5) comparing multi-dimensionality to unidimensionality of the test.

4.3.5 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 to one content area (between-item-multidimensionality). To estimate this multidimensional model, the Monte Carlo estimation implemented in ConQuest was used. The number of nodes was set to 500. (Due to convergence problems even with 25 nodes per dimension, model parameters could not be estimated using Gauss-Hermite quadrature method.) The variances and correlations of the four dimensions are shown in Table 8. All four dimensions exhibited a substantial variance. As expected, the correlations between the four dimensions were rather high, varying between .784 and .947. However, they deviated from a perfect correlation (i.e., they were lower than $r = .95$, see Carstensen, 2013). Moreover, according to model fit indices, the four-dimensional model fits the data slightly better (AIC = 116,034.17, BIC = 116,333.24, number of parameters = 32) than the unidimensional model (AIC = 116,321.26, BIC = 116,568.55, number of parameters = 37). These results indicate that the three cognitive requirements measure a common construct, although it is not completely unidimensional.

Table 8: Results of Four-Dimensional Scaling

	Dim 1	Dim 2	Dim 3	Dim 4
Quantity (4 items)	1.356			
Space and shape (4 items)	0.875	1.746		
Change and relationships (6 items)	0.947	0.842	1.155	
Data and chance (6 items)	0.889	0.784	0.867	1.442

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

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

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	116133.52	23	116,179.52	116,333.24
Four-dimensional	115970.17	32	116,034.17	116,248.05

Note. Contrary to the calculations for the 1PL and 2PL models results in this table were achieved by using Conquest 2.0.

5. Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 5 and at describing how the mathematics competence score had been estimated.

The amount of different kinds of missing responses was evaluated and the total amount of missings was rather low. Especially the amount of invalid responses and not-reached items was rather low. Some items showed higher omission rates, although the amount of omitted items was, in general, acceptable.

Furthermore, the test had a good reliability and distinguished well between test takers, indicated by the test's variance. The item distribution along the ability scale was good, that is the test consisted of easy as well as difficult items.

Indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ, ICC – the items exhibit a good item fit. In addition, discrimination values of the items (estimated either in a 2PL model or as a correlation of the item score with the total score) were acceptable. Different variables were used for testing measurement invariance. Only one item (mas1v062_c) showed a considerable DIF for migration status, preferring subjects with migration status. However, beside this item the test was fair to all three considered subgroups.

Fitting a four-dimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model-fit than the

unidimensional partial credit model. However, high correlations ranging between 0.78 and 0.95 between the four dimensions indicated that the unidimensional model describes the data reasonably well.

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

6. Data in the Scientific Use File

There are 20 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 items) indicating the number of correctly answered subtasks. The dichotomous variables are marked with a '_c' behind their variable names; the polytomous variable is marked with a 's_c' behind its variable name. In the scaling model the polytomous variable is scored in steps of 0.5 – 0 for the lowest category, 1.5 for the highest.

Manifest scale scores are provided in the form of WLE estimates (mas1_sc1) including the respective standard error (mas1_sc2). To correct for differences in the test position of the mathematics test, we acknowledged the main effect related to the test position (see Table 6) in the estimation of the WLE scores of the respondents. Therefore, the provided WLE scores are corrected for the position of the mathematics test within the booklet and can be used for cross-sectional research questions. The ConQuest Syntax for estimating the WLE scores from the items are provided in Appendix A, the fixed item parameters are provided in Appendix B. Test takers 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 the WLE score for mathematical 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 (2012).

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: MIT Press.

Davier, M. von, (2005). A general diagnostic model applied to language testing data (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Duchhardt, C. (2015). *NEPS Technical Report for Mathematics—Scaling results for the additional study Baden Wuerttemberg* (NEPS Working Paper No. 59). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A., & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne – Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze & M. Grüßing (Eds.). *Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (pp. 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. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

Koller, I., Haberkorn, K., & Rohm, T. (2014). *NEPS Technical Report for Mathematics—Scaling results of Starting Cohort 6 for adults in main study 2012* (NEPS Working Paper No. 48). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47(2)*, 149-174.

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

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

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

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 the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, 5(2), 189-216.

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

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 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, & von Maurice, J. (Eds.). *Education as a lifelong process: The German National Educational Panel Study (NEPS).* (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 A: ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort V

```
Title Starting Cohort V, MATHEMATICS: Partial Credit;

data filename.dat;
format pid 4-10 responses 12-31;  /* insert number of columns with data*/

labels << filename_with_labels.nam

codes 0,1,2,3;
score (0,1)(0,1) !item(1-12,14-20);
score (0,1,2,3) (0,0.5,1,1.5) !item(13);

set constraint=cases;

model item + item*step + rotation;
estimate;

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