NEPS
National Educational Panel Study

NEPS Working Papers

Anne-Katrin Jordan and Christoph Duchhardt

NEPS Technical Report for Mathematics—Scaling Results of Starting Cohort 6–Adults

NEPS Working Paper No. 32

Bamberg, November 2013

SPONSORED BY THE
Federal Ministry of Education and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

Editorial Board:

Jutta Allmendinger, WZB Berlin
Cordula Artelt, University of Bamberg
Jürgen Baumert, MPIB Berlin
Hans-Peter Blossfeld, EUI Florence
Wilfried Bos, University of Dortmund
Edith Braun, DZHW Hannover
Claus H. Carstensen, University of Bamberg
Henriette Engelhardt-Wölfler, University of Bamberg
Frank Kalter, University of Mannheim
Corinna Kleinert, IAB Nürnberg
Eckhard Klieme, DIPF Frankfurt
Cornelia Kristen, University of Bamberg
Wolfgang Ludwig-Mayerhofer, University of Siegen
Thomas Martens, DIPF Frankfurt
Manfred Prenzel, TU Munich
Susanne Rässler, University of Bamberg
Marc Rittberger, DIPF Frankfurt
Hans-Günther Roßbach, University of Bamberg
Hildegard Schaeper, DZHW Hannover
Thorsten Schneider, University of Leipzig
Heike Solga, WZB Berlin
Petra Stanat, IQB Berlin
Volker Stocké, University of Kassel
Olaf Struck, University of Bamberg
Ulrich Trautwein, University of Tübingen
Jutta von Maurice, University of Bamberg
Sabine Weinert, University of Bamberg

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 6 for Adults in Main Study 2010/11

Anne-Katrin Jordan and Christoph Duchhardt
Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of lead author:

jordan@ipn.uni-kiel.de

Bibliographic data:

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.

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 6–Adults

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 have been performed on the basis of item response theory (IRT). This paper describes the data and scaling procedures on mathematical competence for Starting Cohort 6–Adults. 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 for adults consisted of 22 items representing different content areas as well as different cognitive components and using different response formats. The test was administered to 5,245 adults. A Rasch model was applied to 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—including the necessary item parameters.

Keywords

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

Contents

1. Introduction

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

Most of the competence data are scaled using models that are based on Item Response Theory (IRT). Since most of the competence tests were developed specifically for implementation in NEPS, several analyses have been conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scales are described in Pohl and Carstensen (2012). In this paper the results of these analyses are presented for reading competence in the second wave of starting cohort 6 (adults). We will first introduce the main concepts of the mathematical competence test. Then, we will describe the mathematical competence data of starting cohort 6 and the analyses performed to estimate competence scores and to check the quality of the test. The results of these analyses will be presented and discussed. Finally, we will describe the data that are available for public use in the Scientific Use File.

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 in 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 major changes in the 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 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 two types of response formats. These are simple multiple-choice (MC), and short constructed response (SCR). In MC items the test taker has to find the correct answer from several, usually four, response options. SCR items require the test taker to write down an answer into an empty box.

3 Data

3.1 The Design of the Study

In the main study 2010/11, reading speed, reading competence, mathematical competence as well as procedural metacognition were assessed. In order to investigate the effects of test duration and to control for position and order effects, the tests were administered to participants in different selection and sequence. For this purpose, the sample was split into four groups receiving different test booklets (see Figure 1). Assignment to test booklets was random. Reading speed and procedural metacognition were assessed of all participants. In order to assess the effects of test duration, half of the sample additionally received both the reading and mathematics test, while the other half received only one of these two competence tests. The sample receiving only one of the two tests was split in two groups. In one group reading competence and in the other group mathematical competence was assessed. In order to control for position and order effects in the group receiving both tests, the two tests were assigned to the participants in different order. One testing group first completed the reading test followed by the mathematic test, while the other group completed the two tests in the opposite order. Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test. All subjects received the same set of reading items in the same order.

Table 1

Design of the study. Reading – reading competence, Math – mathematical competence, meta-p – procedural metacognition for the respective competence

Booklet 1 (in SUF: testv_a = 2)	**Booklet 2 (in SUF: testv_a = 1)**	**Booklet 3 (in SUF: testv_a = 4)**	**Booklet 4 (in SUF: testv_a = 3)**
Reading speed	Reading speed	Reading speed	Reading speed
Reading (+ meta-p)	Math (+ meta-p)	Reading (+ meta-p)	Math (+ meta-p)
Math (+ meta-p)	Reading (+ meta-p)		

The mathematics test for adults was conducted as an individual test and consists of 22 items, which represent different content-related and process-related components[1] and use different response formats. One item (mag9v132s) was eliminated from further analysis because of differential item functioning regarding duration of education (see 4.3.3 for explanation).

The characteristics of the final set of 21 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.

[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.

Table 2

Content Areas of Items in the Mathematics Test Adults

Content area	Frequency
Quantity	4
Space and shape	7
Change and relationships	5
Data and chance	5
Total number of items	21

Table 3

Response Formats of Items in the Mathematics Test Adults

Response format	Frequency
Simple multiple-choice	19
Complex multiple-choice	-
Short constructed response	2
Total number of items	21

3.2 Sample

A description of the design of the study, the sample, as well as the instruments used can be found on the NEPS website.[2] Overall, 5,245[3] persons took the mathematics test. Twenty-four of them gave less than three valid responses to the test items. Because no reliable mathematical competence score may be estimated on the basis of such few responses, these cases were excluded from further analyses.

The results of the remaining 5,221 test takers are presented in the following sections.

3.3 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 just one is required or simply illegible answers in the SCR format. Missing responses due to omitted items occur when persons skip some items. Due to time limits, it may happen that not every person finishes the test within the given time. As a consequence, this results in missing responses due to items that have not been reached.

[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.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions), and they 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 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.

3.4 Scaling Model

To estimate item and person parameters for mathematical competence, a Rasch 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).

Item difficulties for dichotomous variables are estimated using the Rasch 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), whereas 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 some information about the precision of the ability estimates at different levels of ability.

3.5 Checking the Quality of the Scale

The mathematics test was 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 MC items contain a number of distractors (incorrect response options). We investigated if the distractors worked well, that is, whether they were chosen by students with a lower ability rather than by those with a higher ability. To this end, we evaluated the point biserial correlation of giving a certain incorrect response and the total score. We judged correlations below zero as very good, correlations below 0.05 as acceptable, and correlations above 0.05 as problematic.

Item fit was then evaluated, whereas 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 deviant WMNSQ values (|WMNSQ – 1| > 0.15,|*t*-value| > 6) were considered as having a noticeable item misfit, and items with very deviant WMNSQ values (|WMNSQ – 1| > 0.2, |*t*-value| > 8) were considered as having a considerable item misfit, and their performance was investigated further. 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 mathematical competence test that measures the same construct for all subjects. If there were any items that favored certain subgroups (e.g., that 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. Test fairness was investigated for the variables gender, duration of education (≤ 12 years vs. > 12 years), 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 multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were estimated. Differences in the estimated item difficulties between the subgroups were evaluated. On the basis of experiences with preliminary data, we considered absolute differences in estimated difficulties that were greater than 1 logits 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, and differences smaller than 0.4 as very small and, hence, not worthy of special mention. Additionally, model fit was investigated by comparing a model including differential item functioning to a model that only included main effects and no DIF.

The competence data in NEPS are scaled using the Rasch model (1PL), in which Rasch-homogeneity is assumed. 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 Birnbaum model (2PL) (Birnbaum, 1968).

The mathematics test has been constructed to measure a unidimensional mathematical competence score. The assumption of unidimensionality 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, Monte Carlo estimation in ConQuest was used (the number of nodes per dimension was chosen in such a way that stable parameter estimation was obtained). The correlations between the subdimensions were used to evaluate the unidimensionality of the scale.

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 is very small. In fact, 97% of test persons gave no invalid response. The maximum number of invalid responses was three.

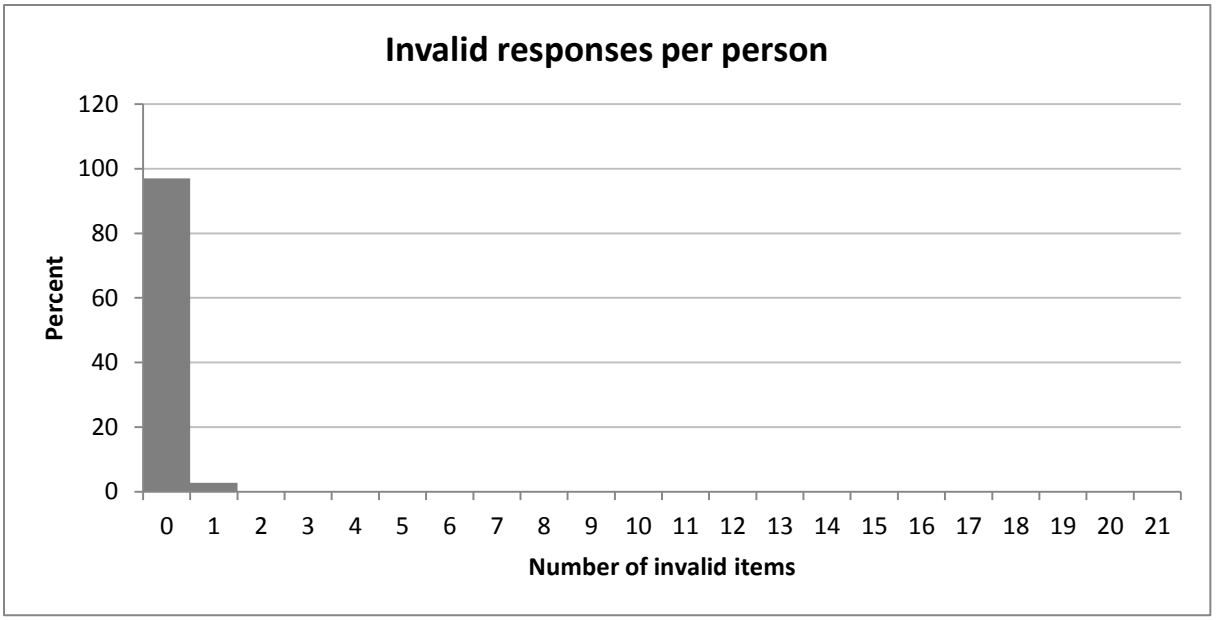

Figure 1. Number of invalid responses.

Missing responses may also occur when persons skip (omit) some items. The number of omitted responses per person is depicted in Figure 2. It shows that 40.1% of the subjects omit no item at all and that 8.9% of the subjects omit more than five items.

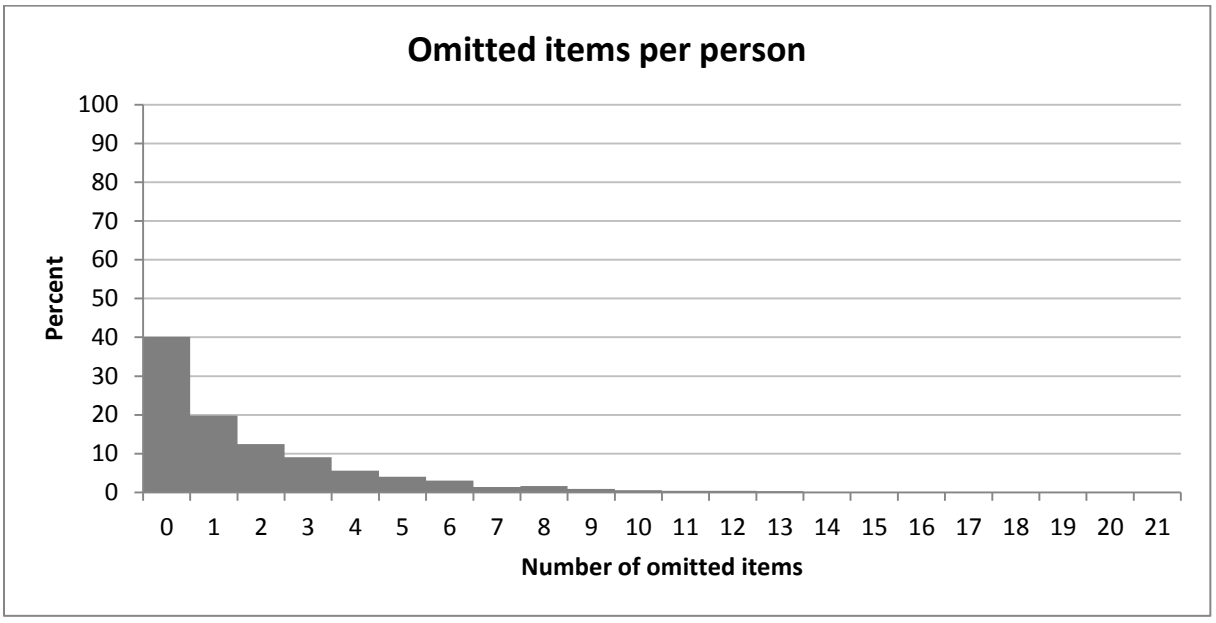

Figure 2. Number of omitted items.

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 a person. As can be seen, only 69.2% reached the end of the test. Only 8.4% of the subjects did not reach the fifth last item.

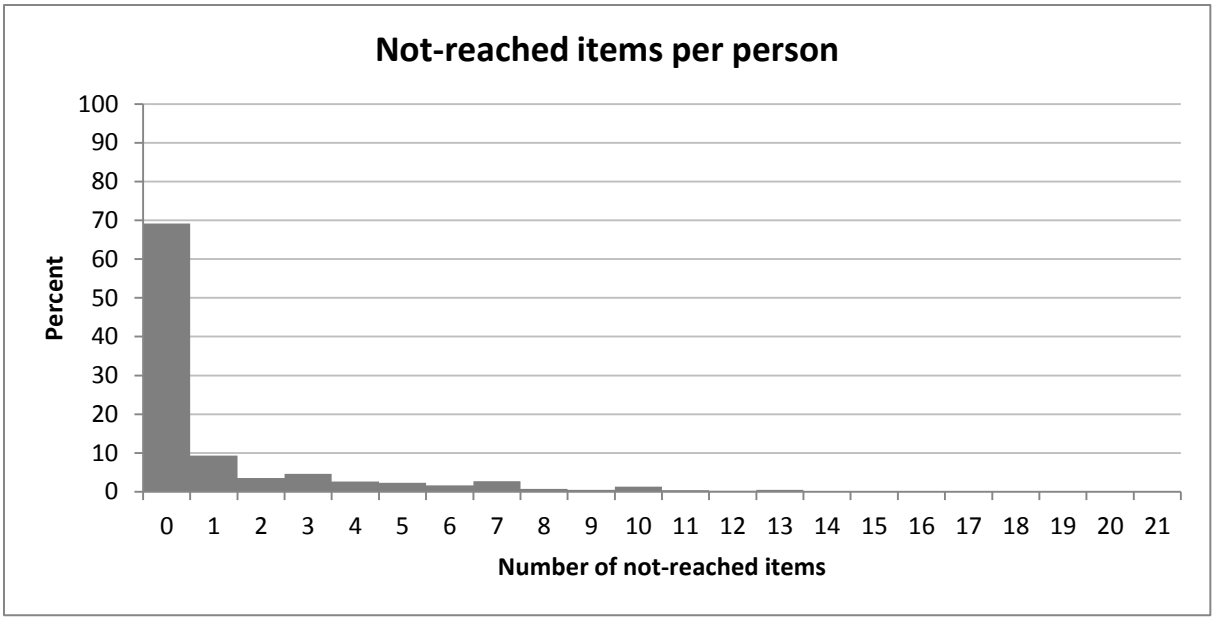

Figure 3. Number of not-reached items.

Figure 4 shows the total number of missing responses per person, which is the sum of nonvalid, omitted, not-reached, and not-determinable missing responses. In total, 31.4% of the subjects show no missing response at all. However, there is also a rather large group of participants – twenty-two percent – showing more than five missing responses.

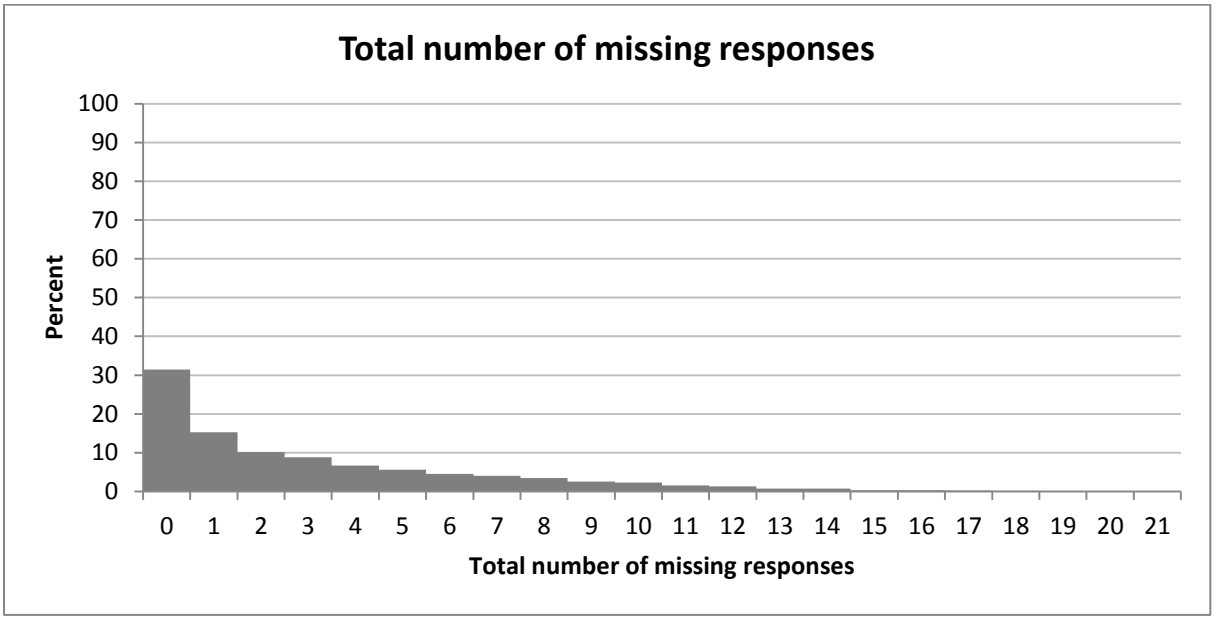

Figure 4. Total number of missing responses.

Overall, there is 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 is very small. The omission rates are acceptable, expect for six items with an omission rate higher than 10%. The highest omission rate (42.6%) occurs for item maa2r091_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. The other items with a noticeable omission rate (10.3%–19.1%) are also either SCR or quite difficult. Compared to omission rates in student tests these rates seem quite high. The number of omitted responses to an item is correlated to .55 with the difficulty of that item. Hence, subjects seem to omit difficult items, presumably because they think that they do not know the answer. They prefer to skip the item rather than to guess.

The number of persons that did not reach an item increases with the position of the item in the test to up to 30.8%.

The total number of missing responses per item varies between 0.9% (item mag9v131_c) and 46.5% (item maa2r091_c).

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
maa2q071_c	1	4925	0.10	5.57	0.00
mag9v131_c	2	5172	0.10	0.84	0.00
mag9r261_c	4[4]	5050	0.02	3.26	0.00
mag9r111_c	5	5039	0.33	3.14	0.02
maa2d131_c	6	4670	0.10	10.29	0.17
maa2d132_c	7	4248	0.04	18.39	0.21
mag9r051_c	8	4729	0.59	8.54	0.29
maa2d041_c	9	5082	0.04	2.26	0.36
maa2r081_c	10	4198	0.04	18.67	0.88
maa2v082_c	11	4263	0.00	17.22	1.13
mag9d201_c	12	4924	0.08	4.12	1.49
maa2r091_c	13	2796	1.00	42.64	2.82
mag9v121_c	14	4779	0.00	5.15	3.31
maa2r121_c	15	4812	0.13	3.66	4.04
maa2d112_c	16	3868	0.04	19.12	6.76

[4] Item 3 was removed from the analyses.

maa2r011_c	17	4436	0.00	6.67	8.37
maa2q101_c	18	4479	0.17	3.37	10.67
mag5v321_c	19	4224	0.15	5.61	13.33
mag9q021_c	20	3961	0.31	5.88	17.95
maa2v061_c	21	4019	0.02	1.55	21.45
maa2q021_c	22	3612	0.04	0.00	30.78

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 IRT analyses. The percentage of persons correctly responding to an item (relative to all valid responses) varies between 23.8% and 93.0% across all items. On average, the rate of correct responses is 55.6% (SD = 20.0%). From a descriptive point of view, the items cover a relatively wide range of difficulties.

The estimated item difficulties are depicted in Table 4.

The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties vary between -3.183 (item maa2v061_c) and 1.579 (item maa2d112_c) with a mean of -0.371. Due to the large sample size, the standard error of the estimated item difficulties (column 4) is very small (SE(ß) ≤ 0.07).

Table 4

Item Parameters

Item	Position in the test	Difficulty	SE of difficulty	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score	Discrimi-nation / 2PL
maa2q071_c	1	-0.527	0.033	1.05	3.4	0.51	0.832
mag9v131_c	2	-1.803	0.039	0.98	-0.9	0.49	1.124
mag9r261_c	4[5]	1.550	0.037	1.03	1.7	0.45	0.763
mag9r111_c	5	-2.248	0.044	1.02	0.8	0.39	0.921
maa2d131_c	6	-0.813	0.035	0.93	-4.3	0.58	1.220
maa2d132_c	7	0.702	0.036	0.90	-6.6	0.60	1.232
mag9r051_c	8	-0.229	0.034	1.06	4.2	0.50	0.798
maa2d041_c	9	-2.019	0.041	1.01	0.4	0.42	0.942

[5] Item 3 was removed from the analyses.

maa2r081_c	10	-0.215	0.036	0.98	-1.2	0.56	1.027
maa2v082_c	11	0.691	0.036	1.15	9.0	0.40	0.562
mag9d201_c	12	-0.349	0.033	0.95	-3.5	0.58	1.142
maa2r091_c	13	1.481	0.047	0.99	-0.2	0.48	0.871
mag9v121_c	14	-0.086	0.033	0.91	-6.7	0.62	1.287
maa2r121_c	15	-1.195	0.036	1.06	3.3	0.46	0.817
maa2d112_c	16	1.579	0.042	1.11	4.8	0.35	0.548
maa2r011_c	17	-0.124	0.035	0.91	-6.4	0.61	1.287
maa2q101_c	18	-0.075	0.035	1.12	8.1	0.45	0.643
mag5v321_c	19	-1.120	0.039	0.97	-1.4	0.53	1.032
mag9q021_c	20	-0.824	0.038	0.97	-1.6	0.55	1.069
maa2v061_c	21	-3.183	0.066	0.94	-1.3	0.36	1.508
maa2q021_c	22	1.007	0.041	0.87	-7.3	0.60	1.374

4.2.2 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 0. An overview of how to work with competence data can be found in Pohl and Carstensen (2012a).

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 1.679, indicating that the test differentiated well between subjects. The reliability of the test (EAP/PV reliability = .814, WLE reliability = .78) 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
	XX	
	X	
	X	
3	XXX	
	XXX	
	XXXXX	
	XXXXXXX	
	XXXXXX	
	XXXXXXXXX	
2	XXXXXXXXXX	
	XXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXX	16
	XXXXXXXXXXXXXXXXXX	4 13
	XXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXX	
1	XXXXXXXXXXXXXXXXXXXXXX	22
	XXXXXXXXXXXXXXXXXXXXXXXX	7
	XXXXXXXXXXXXXXXXXXXXXXXXXX	11
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	14 17 18
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	8 10
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	12
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	1
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	6 20
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
-1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15 19
	XXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXX	2
	XXXXXXXXX	9
-2	XXXXXXXXX	
	XXXXXX	5
	XXXX	
	XXX	
	XXX	
-3	XXX	
	X	21
	X	
	X	
	X	
-4		

Figure 5. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 7.2 cases. The difficulty of the items is depicted on the right-hand side of the graph. Each number represents one item (see Table 4).

4.3 Quality of the test

Because almost all of the items of the mathematical competence test refer to different stimuli (there are only two small units consisting of two items each), the assumption of local item independence is plausible.

4.3.1 Distractor analyses

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

Two distractors of the MC items mag9r261_c and maa2d111 had a positive point biserial correlation (0.15 and 0.02), which we considered critical. However, choosing these distractors indicates a failure to meet the central cognitive demand of that particular item. We therefore decided that, from a theoretical point of view, it is nonetheless desirable to include this item in the analysis.

Table 5 shows a summary of point biserial correlations between response and ability for correct and incorrect responses restricted to MC items (the only items where subjects were asked to choose between distractors).

Table 5

Point Biserial Correlations of Correct and Incorrect Response Options

Parameter	Correct responses (MC items only)	Incorrect responses (MC items only)
Mean	0.50	-0.22
Minimum	0.35	-0.49
Maximum	0.62	0.15

4.3.2 Item fit

The item fit is very good. WMNSQ is close to 1 with the lowest value being 0.87 (item maa2q021_c) and the highest being 1.15 (item maa2v082_c). The correlation of the item score with the total score varies between 0.35 (item maa2d111_c) and 0.62 (item mag9v121_c) with an average correlation of 0.50. Almost all item characteristic curves (ICC) showed a good or very good fit of the items. The three items with the highest WMNSQs (items maa2q101_c, maa2d112_c and maa2v082_c) showed an acceptable, slightly flat ICC. The item with the lowest WMNSQ (maa2q021_c) showed an acceptable but slightly steep ICC.

4.3.3 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 gender, duration of education (≤ 12 years vs. > 12 years) 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). 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. Analogously, a positive main effect implies that the test is more difficult for the first group, for example, for females.

Table 6

Differential Item Functioning (Absolute Differences Between Difficulties)

	Gender	**Duration of education**	**Migration status**	**Books**	**Booklet**
Item	Female vs male	≤ 12 years vs > 12 years	With vs without	≤ 100 vs > 100	Math second test vs math third test[6]
maa2q071_c	-0.172	-0.360	0.018	-0.392	0.096
mag9v131_c	-0.334	0.226	0.294	0.402	0.334
mag9r261_c	0.068	-0.442	0.068	-0.158	0.072
mag9r111_c	-0.358	-0.006	-0.214	0.144	-0.060
maa2d131_c	-0.292	0.338	-0.160	0.020	0.170
maa2d132_c	-0.116	0.380	-0.070	0.036	0.112
mag9r051_c	-0.194	0.230	0.076	0.290	-0.274
maa2d041_c	-0.328	-0.154	-0.066	0.146	0.010
maa2r081_c	0.350	-0.132	0.072	-0.144	-0.060
maa2v082_c	-0.142	-0.448	0.126	-0.130	-0.052
mag9d201_c	0.026	0.202	0.052	0.014	0.056
maa2r091_c	-0.102	0.068	0.188	0.070	0.012
mag9v121_c	0.390	0.002	0.022	-0.048	-0.006
maa2r121_c	0.594	-0.322	-0.216	-0.234	-0.178
maa2d112_c	-0.524	-0.194	-0.034	0.106	-0.044
maa2r011_c	0.546	0.288	-0.024	-0.146	-0.132
maa2q101_c	-0.032	-0.470	0.076	-0.174	0.150
mag5v321_c	-0.070	-0.168	0.100	-0.150	0.040
mag9q021_c	-0.148	0.632	-0.090	0.484	-0.120
maa2v061_c	0.006	0.370	0.266	0.446	0.078
maa2q021_c	0.486	0.348	-0.374	0.060	-0.156
Main effect	**0.816**	**1.232**	**0.186**	**0.584**	**0.064**

[6] i.e. Booklets 2 + 4 vs Booklet 1, cf. 3.1

Overall, 2,649 (50.7%) of the test takers were female and 2,572 (49.3%) were male. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.816 logits, Cohen's *d* = 0.664). There was no item with a considerable gender DIF. For four items the difference in item difficulties between the two groups exceeded 0.4 logits, the maximum being 0.594 logits (item maa2r121_c).

Regarding the duration of education, 2,726 (54.1%) of the students had received 12 years of education or less. The other group of 2,309 (45.9%) test takers had undergone an education period of more than 12 years. On average, those students who had spent more time in education performed much better in the mathematics test (main effect = 1.232 logits, Cohen's *d* = 1.082). There was one item (mag9q021_c) with a considerable DIF (0.632 logits). For another three items (mag9r261_c, maa2v082_c, maa2q101_c) the DIF exceeded 0.4 logits. The other items show no considerable DIF in terms of duration of education.

There were 4,169 (82.8%) participants without migration background, 756 (15%) participants with migration background, and 110 (2.2%) participants without a valid response. Only the first two groups were used for investigating DIF of migration. On average, participants without migration background performed slightly better in the mathematics test than those with migration background (main effect = 0.186 logits, Cohen's *d* = 0.144). There was no considerable DIF comparing the two groups. The highest difference in item difficulties between groups was 0.374 logits (item maa2q021_c).

The number of books at home was used as a proxy for socioeconomic status. There were 1,908 (36.5%) test takers with 0 to 100 books at home, and 3,312 (63.5%) test takers with more than 100 books at home. One person did not answer this question. Group differences and DIF were investigated by using these two groups. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.584 logits (Cohen's *d* = 0.462) lower in mathematics than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books. For two items, the difference in item difficulties between the two groups exceeded 0.4 logits, the maximum being 0.484 logits (item mag9q021).

According to the design of the study (cf. section 3.1), the math test was either taken directly after the reading speed test – 3525 participants (67.5%) – or after completing the reading speed and reading tests – 1696 participants (32.5%). Group differences and DIF were investigated using these two groups. Surprisingly, the second group (math = third test) slightly outperformed the first (main effect = 0.064 logits, Cohen's *d* = 0.049). There was no considerable DIF comparing the two groups. The highest difference in item difficulties between groups was 0.334 logits (item mag9v131_c).

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 into greater account the number of estimated parameters and, thus, prevents from overparametrization of models. Using BIC, the more parsimonious models including only the main effects of the number of books, migration status, and booklet, respectively, are preferred over the more complex respective DIF models. However, BIC prefers the models including both main effect and DIF effect of gender and duration of education, respectively,

to the models including only the respective main effect. (Note that the analyses including gender contain fewer cases and, thus, 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
Gender	Main effect	98258.24	23	98304.24	98455.13
	DIF	97914.93	44	98002.93	98291.59
Duration of education	Main effect	94181.25	23	94227.25	94377.31
	DIF	93836.14	44	93924.14	94211.21
Migration status	Main effect	93237.92	23	93283.92	93433.47
	DIF	93201.33	44	93289.33	93575.42
Books	Main effect	98480.04	23	98526.04	98676.93
	DIF	98316.62	44	98404.62	98693.27
Booklet	Main effect	98695.80	23	98741.80	98892.69
	DIF	98635.99	44	98723.99	99012.65

4.3.4 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 between 0.55 (item maa2d112_c) and 1.51 (item maa2v061_c). The 2PL model (AIC = 97969.219, BIC = 98316.922, number of parameters = 53) fits the data better than the Rasch model (1PL) (AIC = 98711.408, BIC = 98941.024, number of parameters = 35). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework (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.

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 method of estimation implemented in ConQuest was used. The number of nodes was set to 2,000. (Due to convergence problems even with 25 nodes per dimension, model parameters could not be estimated using the Gauss-Hermite quadrature method. This might be caused by very high correlations between the four dimensions.) From this analysis all persons with less than two

valid responses in one or more dimensions were excluded. The resulting sample consisted of *N* = 4,488 persons.

The variances and correlations of the four dimensions are shown in Table 8. All four dimensions exhibit a substantial variance. The correlation between the four dimensions is—as expected—very high, varying between .898 and .981.

Table 8

Results of Four-Dimensional Scaling

	Quantity	**Space and shape**	**Change and relationships**	**Data and chance**
Quantity (4 items)	1.677			
Space and shape (7 items)	0.951	1.745		
Change and relationships (5 items)	0.981	0.959	1.887	
Data and chance (5 items)	0.956	0.898	0.946	2.306

Note. Variance of the dimensions are depicted in the diagonal, correlations are given in the off-diagonal.

Model fit between the unidimensional and the four-dimensional model is compared in Table 9. (Here, the same sample of *N* = 4488 persons was taken for the unidimensional model.)

Table 9

Comparison of the Unidimensional and the Four-Dimensional Model

Model	**Deviance**	**Number of parameters**	**AIC**	**BIC**
Unidimensional	89069.023	22	89113.023	89254.025
Four-dimensional	89034.761	31	89096.761	89295.445

The comparison shows that, using BIC, the unidimensional model describes the data better than the four-dimensional model. Taking into account the high correlations between the four content areas as well, the assumption of unidimensionality seems very reasonable.

5. Discussion

The analyses in the previous sections have aimed to provide information on the quality of the mathematics test in Starting Cohort 6 and describe how the mathematics competence score was 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 = .814, WLE reliability = .780). It distinguishes well between test takers, indicated by the test's variance (= 1.679). However, very difficult items are rare; hence, test targeting is somewhat suboptimal. The test measures mathematical competence of high-performing subjects a little less accurately. This can be explained by the heterogeneity of this target group of adults.

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 the total score) are acceptable. Different variables were used for testing measurement invariance. One item was excluded from the main analyses due to high DIF regarding duration of education. In the remaining items, almost no considerable DIF became evident for any of the examined variables, indicating that the test is reasonably fair to the subgroups considered.

Fitting a four-dimensional model (between-item-multidimensionality, the dimensions being the content areas) yields a slightly worse model-fit regarding BIC (slightly better when using AIC) than the unidimensional model. Very high correlations of 0.9 and higher between the four dimensions also indicate that the unidimensional model describes the data very well.

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

6. Data in the Scientific Use File

There are 21 items in the data set that are scored as dichotomous variables (MC and SCR items) with 0 indicating an incorrect response and 1 indicating a correct response. The variables are marked with a '_c' behind their variable name.

Manifest scale scores are provided in the form of WLE estimates (ma_sc1) including the respective standard error (ma_sc2). The ConQuest Syntax for estimating the WLE scores from the items are provided in the Appendix.

Test takers that did not take part in the math test or those that did not give enough valid responses to estimate a scale score (24 cases as mentioned in Section 0) 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 (2012a).

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* (p. 397–479). Reading, MA: MIT Press.

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.

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). *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. (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 Rreading – Scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper No. 15). 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, & 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

ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort 6–Adults

```
Title Starting Cohort VI, MATHEMATICS: Rasch Model;

data filename.dat;
format pid 4-10 responses 12-32 position 34; /* insert number of columns with data*/

labels << filename_with_labels.txt;

codes 0,1;

set constraint=cases;

model item + position;
estimate;

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