NEPS
National Educational Panel Study

NEPS Working Papers

Christoph Duchhardt & Annkathrin Gerdes

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 4 in Ninth Grade

NEPS Working Paper No. 22

Bamberg, April 2013

SPONSORED BY THE

Federal Ministry
of Education
and Research

NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 4 in Ninth Grade

Christoph Duchhardt[1] & Annkathrin Gerdes[1]

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

Email address of the lead author:

duchhardt@ipn.uni-kiel.de

Bibliographic Data:

Duchhardt, C. & Gerdes, A. (2013): *NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 4 in ninth grade* (NEPS Working Paper No. 22). 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 4 in Ninth 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 4 in ninth 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 ninth grade consisted of 22 items which represented different content areas as well as different cognitive components and used different response formats. The test was administered to 14,524 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 – including the necessary item parameters.

Keywords

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

Content

1. Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. 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 4.

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 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

All participants received the mathematics test after completing the ICT test, the science test, and the reading speed test. No multi-matrix design was applied regarding the choice and order of the items *within* the mathematics test. All students received the same mathematics items in the same order.

The mathematics test in grade 9 consists of 22 items which represent different content-related and process-related components[1] and use different response formats. The characteristics of the 22 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. One of the CMC items consists of three subtasks. The other consists of four subtasks with one of them excluded from analysis due to unsatisfactory item fit. The SCR item required the subjects to write down a number.

Table 1: Content Areas of Items in the Mathematics Test Grade 9

Content area	**Frequency**
Quantity	7
Space and shape	6
Change and relationships	6
Data and chance	3
Total number of items	22

Table 2: Response Formats of Items in the Mathematics Test Grade 9

Response format	**Frequency**
Simple Multiple-Choice	19
Complex Multiple-Choice	2
Short-constructed response	1
Total number of items	22

[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].

Overall, 14,524[3] persons took the mathematics test. One of them had less than three valid responses to the test items. Since no reliable mathematical competence score may be estimated on the basis of such few responses, this case was excluded from further analyses. Another 174 cases were not included in the analyses due to a separate data delivery for these cases.

The results of the remaining 14,349 test takers are presented in the following sections. For the Scientific Use File, however, the WLE scores were estimated with the data of 14,523 subjects, including these 174 cases.

4. Analyses

In order to carry out first analyses, the SCR item was scored, rating each answer either as correct or wrong or some kind of missing.

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 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.

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

[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.

The CMC items consist 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 whole CMC item was scored as missing.

In the following analyses each category of the polytomous items was scored with 0.5 points, while simple MC items and the SCR item were scored as 1 (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2012b, for studies on the scoring of different response formats).

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 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 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 responses to the subtasks of the CMC items were aggregated to a polytomous variable for each CMC item. 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 the MC 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 polytomous CMC item variables.

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 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, thereby treating all subtasks of CMC items as single items. 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 for the MC items and the polytomous CMC items 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 (t-value > |6|) were considered as having a noticeable items 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 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 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 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, school types (high school vs. non-high school) 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 were estimated. Differences in the estimated item difficulties between the subgroups were evaluated. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties that were greater than 1 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 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) (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 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 (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

As can be seen in Figure 1, the number of invalid responses per person is very small. In fact, 93.2% of test persons gave no invalid response. Less than 1% of the subjects have more than one invalid response. Considering the response format, this kind of missing response occurs particularly in the short SCR answer item (mag9r061_c).

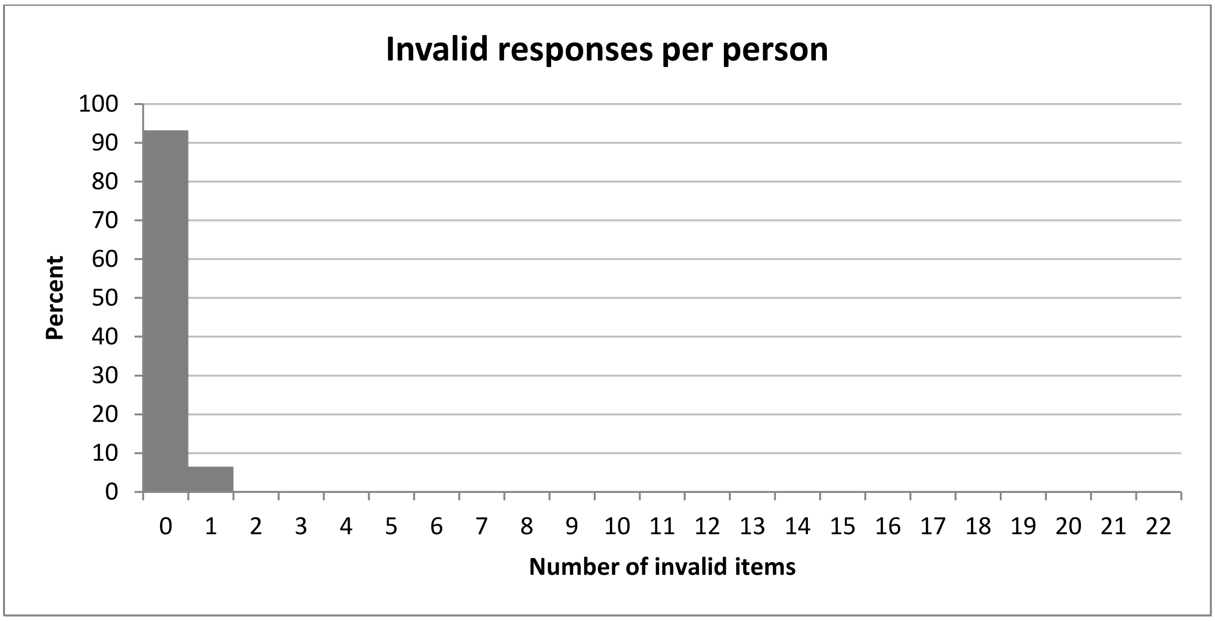

Figure 1: Number of invalid responses

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 66.3% of the subjects omit no item at all. 1.2% of the subjects omit more than five items.

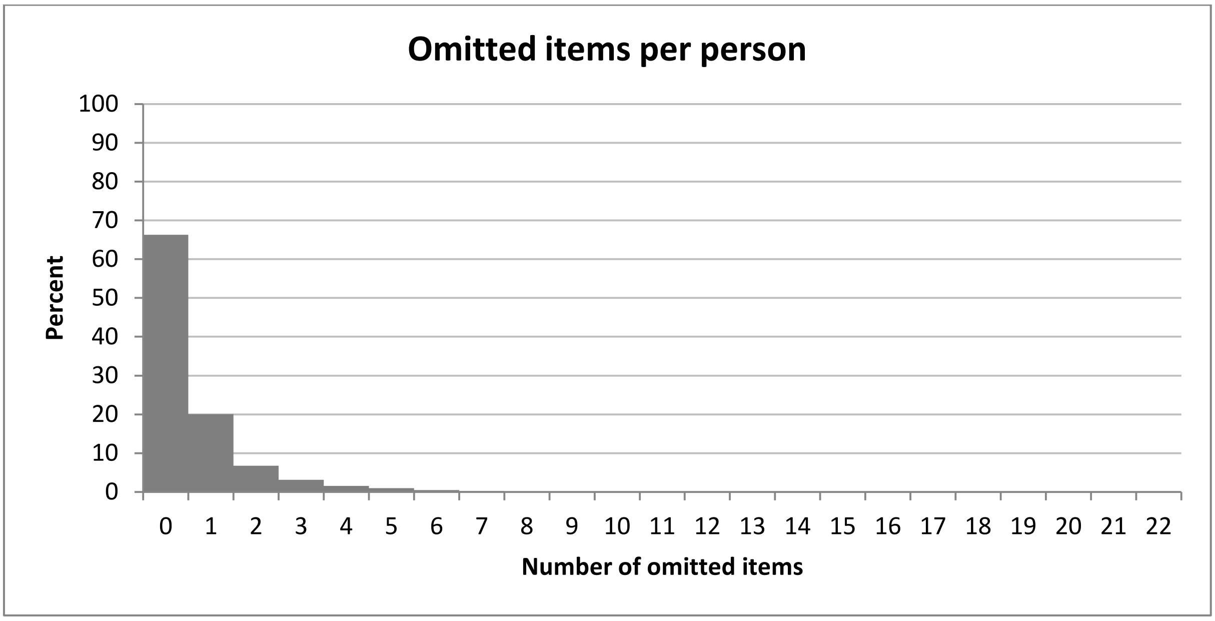

Figure 2: Number of omitted items

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, items that were not reached are quite rare and pose no problem for this test. Only about 4.6% of the subjects did not reach the end of the test.

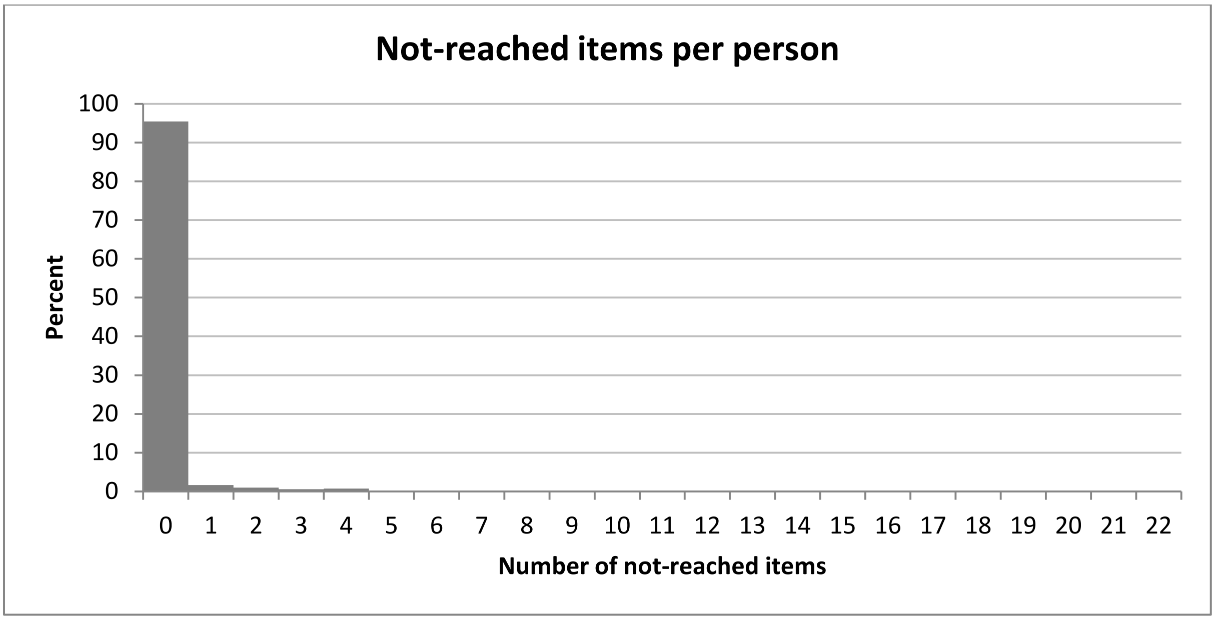

Figure 3: Number of not-reached items

Complex multiple-choice items 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 – with the exception that the CMC items were 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 nine times.

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, 60% of the subjects show no missing response at all. Only 2.45% show more than five missing responses.

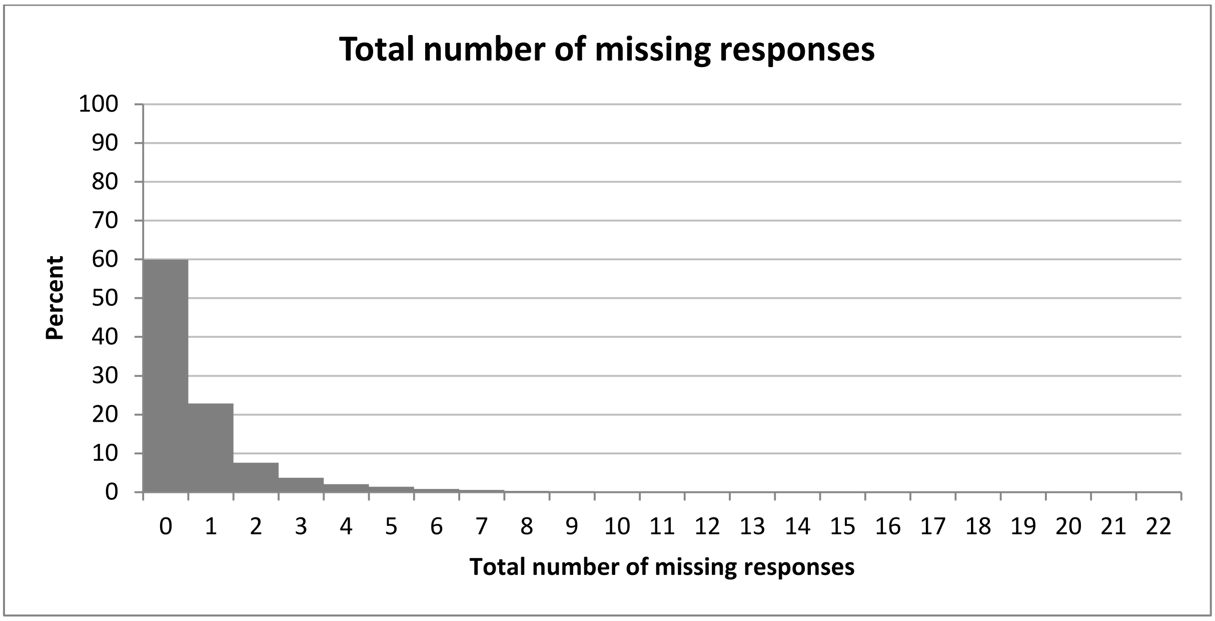

Figure 4: Total number of missing responses

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

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.

Overall, the number of invalid responses per item is very small. The only exception is item mag9r061_c, the only SCR item, to which 5.69% of the participants responded invalidly. This might be due to the response format.

The omission rates are acceptable. There are three items with are omitted by more than 5% of the participants. The by far highest omission rate (15.71%) occurs for item mag9r061_c. As this item is the only SCR item, the subjects might have preferred to skip this item rather than to guess. The number of omitted responses is correlated to .156 with the difficulty of the item. When excluding the one item with the highest omission rate (mag9r061_c), the correlation decreases to -0.08. Hence, there does not seem to be any systematic relationship between the difficulty of an item and its omission rate.

The number of persons that did not reach an item increases with the position of the item in the test to up to 4.59%. This is a rather small and acceptable amount.

The total number of missing responses per item varies between 0.50% (item mag9q181_c) and 21.92% (item mag9r061_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
mag9q071_c	1	14182	0.10	1.07	0.00
mag9v131_c	2	14206	0.05	0.95	0.00
mag9v13s_c	3	13219	0.06	7.82	0.00
mag9r261_c	4	13887	0.07	3.15	0.00
mag9r111_c	5	14009	0.03	2.33	0.00
mag9d171_c	6	14138	0.06	1.41	0.00
mag9d151_c	7	14226	0.15	0.70	0.01
mag9r051_c	8	14165	0.06	1.21	0.01
mag9v011_c	9	14171	0.03	1.19	0.02
mag9v012_c	10	14005	0.04	2.33	0.03
mag9q161_c	11	13946	0.06	2.70	0.06
mag9d201_c	12	14194	0.04	0.98	0.06
mag9r191_c	13	14198	0.03	0.94	0.08
mag9v121_c	14	14145	0.03	1.30	0.09
mag9q181_c	15	14277	0.02	0.32	0.16
mag9r25s_c	16	13039	0.29	8.51	0.33
mag9r061_c	17	11204	5.69	15.71	0.52
mag9q081_c	18	13930	0.04	2.18	0.70
mag9q101_c	19	13662	0.06	3.32	1.41
mag9q021_c	20	13529	0.13	3.59	2.00
mag9v091_c	21	13681	0.10	1.60	2.95
mag9q211_c	22	13668	0.16	0.00	4.59

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 the responses given before performing IRT analyses. Regarding each subtask of a CMC item as a single variable, the percentage of persons correctly responding to an item (relative to all valid responses) varies between 10.5% and 85.6% across all items. On average, the rate of correct responses is 54.3% (SD = 16.9%). 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 polytomous variables) 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 items are depicted in Table 4b. The estimated item difficulties (or location parameters for polytomous variables) vary between -2.144 (item mag9q181_c) and 2.585 (item mag9r261_c) with a mean of -0.131. Due to the large sample size, the standard error of the estimated item difficulties (column 4) is very small (SE(ß) ≤ 0.03).

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 / 2PL
mag9q071_c	1	-0.61	0.019	1.02	2.7	0.45	0.88
mag9v131_c	2	-0.01	0.019	1.02	2.5	0.47	0.89
mag9v13s_c	3	-0.986	0.02	0.99	-0.8	0.49	1.14
mag9r261_c	4	2.585	0.03	0.97	-1.7	0.39	1.02
mag9r111_c	5	-0.761	0.019	1	0.1	0.46	1.03
mag9d171_c	6	0.029	0.019	1.1	13.5	0.39	0.63
mag9d151_c	7	-1.394	0.021	0.95	-5.2	0.47	1.44
mag9r051_c	8	0.439	0.019	1.05	6.4	0.43	0.75
mag9v011_c	9	-0.897	0.02	0.95	-5.6	0.50	1.26
mag9v012_c	10	-0.17	0.019	0.96	-4.9	0.52	1.09
mag9q161_c	11	0.971	0.02	1.1	10.4	0.36	0.58
mag9d201_c	12	0.222	0.019	1.01	0.9	0.48	0.91
mag9r191_c	13	-0.836	0.019	1.06	7.4	0.40	0.75
mag9v121_c	14	1.289	0.021	1.02	1.5	0.42	0.78
mag9q181_c	15	-2.144	0.025	0.99	-0.9	0.35	1.19
mag9r25s_c	16	-1.219	0.019	1.06	5.9	0.43	0.80
mag9r061_c	17	1.053	0.023	0.93	-7.4	0.54	1.10
mag9q081_c	18	0.304	0.019	1.01	1.4	0.48	0.86
mag9q101_c	19	-0.724	0.02	0.91	-11.7	0.55	1.54
mag9q021_c	20	0.222	0.019	0.97	-3.7	0.52	1.04
mag9v091_c	21	-0.314	0.019	0.93	-9	0.55	1.29
mag9q211_c	22	0.08	0.019	0.99	-1.8	0.50	0.99

Table 4b: Step Parameters of Polytomous Items

Item	Position in the test	location parameter	step 1 (SE)	step 2 (SE)	step 3
mag9v13s_c	3	-0.986	-0.061 (0.018)	-0.629 (0.018)	0.691
mag9r25s_c	16	-1.219	-1.083 (0.018)	0.895 (0.021)	0.188

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 1.178, indicating that the test differentiated well between subjects. The reliability of the test (EAP/PV reliability =.811, WLE reliability = .794) 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 is only one very difficult item (4). 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	4
	X	
	XX	
	XX	
2	XX	
	XX	
	XX	
	XXXX	
	XXXX	
	XXXX	14
	XXXXXX	17
1	XXXXX	11
	XXXXX	
	XXXXXX	
	XXXXXXX	8
	XXXXXXXX	18
	XXXXXXXXX	12 2
	XXXXXXXX	6 22
0	XXXXXXXX	2
	XXXXXXXX	10
	XXXXXXXXX	21
	XXXXXXXXXX	
	XXXXXXXXXX	1
	XXXXXXXX	5 13
-1	XXXXXXXX	3 9
	XXXXX	
	XXXXX	16
	XXXXX	7
	XXX	
	XX	
	XX	
-2	X	
	X	15
	X	
	X	
-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 85.2 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4).

5.3 Quality of the test

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

5.3.1 Fit of the subtasks of complex multiple-choice items

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

The rates of correct responses given to the subtasks of the CMC items varied from 56.9% to 76.1%. With one exception, the subtasks showed a good item fit – WMNSQ ranging between 0.93 and 1.09 and the respective t-values between -6.7 and 11.8. The only subtask exhibiting unsatisfactory item fit – WMNSQ of 1.21 and a respective t-value of 27 – was excluded from further analysis. The good model fit of the other subtasks was considered to justify their aggregation to polytomous variables for each item (mag9v13s_c and mag9r25s_c).

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 – for the MC items – the point biserial correlation between each incorrect response (distractor) and the students' total score. This 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.

Two distractors of the MC item mag9r261_c had a positive point biserial correlation (0.08 and 0.1), 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.

One distractor for another MC item (mag9q161_c) had a point biserial correlation of 0. However, this finding was not considered problematic. All other distractors exhibited a point biserial correlation with the total score below zero. These results indicate that the distractors worked reasonably well.

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 subject 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	0.444	-0.183
Minimum	0.340	-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 0.91 (item mag9q101_c) and the highest being 1.1 (items mag9d171_c and mag9q161_c). The correlation of the item score with the total score varies between .35 (item mag9q181_c) and .55 (items mag9q101_c and mag9v091_c) with an average correlation of .46. Almost all item characteristic curves (ICC) showed a good or very good fit of the items. The two items with the highest WMNSQs (items mag9d171_c and mag9q161_c) showed an acceptable, slightly flat ICC. The item with the lowest WMNSQ (mag9q101_c) showed an acceptable but slightly steep 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 gender, the number of books at home (as a proxy for socioeconomic status), school type, 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.

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

	Gender	**Books**			**School type**	**Migration status**		
Item	female vs male	>100 vs <100	>100 vs missing	<100 vs missing	non-high school vs high school	without vs with	without vs missing	with vs missing
mag9q071_c	0.33	-0.017	-0.187	-0.17	-0.204	0.062	0.001	-0.061
mag9v131_c	0.006	-0.032	0.157	0.189	-0.08	0.053	-0.014	-0.067
mag9v13s_c	-0.174	0.133	0.034	-0.099	0.47	0	-0.033	-0.033
mag9r261_c	0.012	0.222	0.335	0.113	0.134	0.023	0.236	0.213
mag9r111_c	-0.476	0.054	-0.36	-0.414	0.094	-0.189	-0.042	0.147
mag9d171_c	-0.228	-0.273	0.245	0.518	-0.382	0.372	0.264	-0.108
mag9d151_c	-0.034	0.297	-0.177	-0.474	0.362	-0.333	-0.327	0.006
mag9r051_c	-0.028	0.007	0.083	0.076	-0.20	-0.018	0.09	0.108
mag9v011_c	0.252	0.157	0.05	-0.107	0.04	-0.346	-0.215	0.131
mag9v012_c	0.164	0.076	0.065	-0.011	0.11	-0.093	-0.183	-0.09
mag9q161_c	0.108	-0.275	0.191	0.466	-0.49	0.215	0.226	0.011
mag9d201_c	0.078	0.02	0.28	0.26	0.014	0.105	0.162	0.057
mag9r191_c	-0.074	-0.114	-0.117	-0.003	-0.244	0.09	0.021	-0.069
mag9v121_c	0.19	-0.082	-0.071	0.011	-0.298	0.085	0.203	0.118
mag9q181_c	-0.212	0.145	0.028	-0.117	0.178	-0.059	-0.238	-0.179
mag9r25s_c	-0.328	-0.048	0.369	0.417	-0.104	0.028	0.029	0.001
mag9r061_c	-0.048	0.177	0.243	0.066	0.094	-0.105	-0.279	-0.174
mag9q081_c	0.394	-0.102	-0.049	0.053	-0.356	-0.014	0.039	0.053
mag9q101_c	-0.072	0.266	0.043	-0.223	0.402	-0.197	-0.112	0.085
mag9q021_c	-0.132	0.111	-0.097	-0.208	0.234	0.021	-0.138	-0.159
mag9v091_c	0.196	0.28	0.188	-0.092	0.254	-0.14	-0.115	0.025
mag9q211_c	0.030	-0.014	-0.01	0.004	0.262	0.058	-0.078	-0.136
Main effect	**0.374**	**0.779**	**-0.053**	**-0.832**	**1.382**	**-0.529**	**-0.554**	**-0.025**

Overall, 7,154 (49.9%) of the test takers were female and 7,188 (50.1%) were male. Seven missing responses were given in relation to the variable gender. These cases were excluded from the DIF analysis. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.374 logits, Cohen's d = 0.350). 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 mag9r111_c (0.476 logits).

The number of books at home was used as a proxy for socioeconomic status. There were 6,048 (42.1%) test takers with 0 to 100 books at home, 7,989 (55.7%) test takers with more than 100 books at home, and 312 (2.2%) test takers without any valid response. 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 0.779 logits (Cohen's d = 0.770) lower in mathematics than participants with more than 100 books. Participants without a valid response in relation to the variable books at home performed 0.053 logits (Cohen's d = 0.052) or 0.832 logits (Cohen's d = 0.822) worse than participants with up to 100 and more than 100 books, respectively. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.297). Comparing the group without valid responses to the two groups with valid responses, DIF exceeding 0.4 logits occurred in five items (mag9r111_c, mag9d171_c, mag9d151_c, mag9q161_c and mag9r25s_c), the maximum being 0.518 logits.

5,070 (35.3%) of the participants were high-school students, whereas 9,279 (64.7%) attended different types of school. On average, high-school students showed a considerably higher mathematical competence than the other students (main effect = 1.382, Cohen's d = 1.619). There was no item with a considerable DIF. Differences in item difficulties exceeding 0.4 logits were observed in the items mag9v13s_c, mag9q161_c, mag9q101_c, the maximum being 0.49.

There were 10,046 (70.0%) participants without migration background, 3,669 (25.6%) participants with migration background, and 634 (4.4%) 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 = 0.529 logits, Cohen's d = 0.501). Also, subjects with missing values for migration performed 0.554 logits (Cohen's d = 0.524) or 0.025 logits (Cohen's d = 0.024) worse than those without migration background and with migration background, respectively. There is no considerable DIF comparing the three groups. The highest difference in item difficulties between groups was 0.372 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 number of books and migration status, 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 school type, respectively, to the models including only the respective main effect. (Note that the analyses including gender contain fewer cases and, 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
Gender	main effect	380033.503	28	380089.503	380301.490
	DIF	379424.762	50	379524.762	379903.309
Books	main effect	378975.210	29	379033.210	379252.782
	DIF	378610.990	73	378756.990	379309.705
Migration	main effect	379963.789	29	380021.789	380241.361
	DIF	379651.565	73	379797.565	380350.280
School type	main effect	375296.159	28	375352.159	375564.159
	DIF	374489.730	50	374589.730	374968.302

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 between 0.58 (item mag9q161_c) and 1.54 (item mag9q101_c). The 2PL model (AIC = 378100.243, BIC = 378546.958, number of parameters = 59) fits the data better than the partial credit model (1PL) (AIC = 379816.919, BIC = 380119.776, number of parameters = 40). 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.

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 to one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Gauss-Hermite quadrature estimation implemented in ConQuest was used. The number of nodes per dimension was chosen in such a way that a stable parameter estimation was obtained. 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 .906 and .967.

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

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
Quantity (7 items)	1.220			
Space and shape (6 items)	0.925	1.180		
Change and relationships (6 items)	0.965	0.942	1.466	
Data and chance (3 items)	0.967	0.906	0.946	1.109

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	380564.947	27	380618.947	380823.376
Four-dimensional	380384.102	36	380456.102	380728.673

6. Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 4 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 = .811, WLE reliability = .794). It distinguishes well between test takers, indicated by the test's variance (= 1.178). However, very difficult items are rare, 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 the 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, very high correlations of over 0.9 between the four dimensions indicate that the 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 22 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 name, the polytomous variables are marked with a 's_c' behind their variable names. In the scaling model the polytomous variables are scored in steps of 0.5 – 0 for the lowest category, 1.5 for the highest.

As mentioned in section 3.2, 174 cases have failed to be considered in the analyses presented so far. For the Scientific Use File the item parameters from these analyses (i.e., those reported above) were fixed in order to estimate ability scores for all test takers. 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 Appendix A, the fixed item parameters are provided in Appendix B. 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 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*.*

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. 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 reading – 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, & v. Maurice, J. (Eds.). *Education as a lifelong process: The German National Educational*

Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14 (pp. 67-86). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Appendix

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

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


data filename.dat;
format pid 4-10 responses 12-33; /* insert number of columns with data*/


labels << filename_with_labels.txt;
import anchor_parameters << fixed_parameters.par;


codes 0,1,2,3;


score (0,1) (0,1)                        !items (1-2,4-15,17-22);
score (0,1,2,3) (0,0.5,1,1.5)            !items (3,16);


set constraint=cases;


model item + item*step + booklet;
estimate;


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

Appendix B: Fixed Item Parameters

```
 1      -0.60999        /* item mag9q071_c */
 2      -0.00987        /* item mag9v131_c */
 3      -0.98617        /* item mag9v13s_c */
 4       2.58498        /* item mag9r261_c */
 5      -0.76074        /* item mag9r111_c */
 6       0.02884        /* item mag9d171_c */
 7      -1.39374        /* item mag9d151_c */
 8       0.43851        /* item mag9r051_c */
 9      -0.89714        /* item mag9v011_c */
10      -0.17001        /* item mag9v012_c */
11       0.97149        /* item mag9q161_c */
12       0.22235        /* item mag9d201_c */
13      -0.83579        /* item mag9r191_c */
14       1.28854        /* item mag9v121_c */
15      -2.14397        /* item mag9q181_c */
16      -1.21869        /* item mag9r25s_c */
17       1.05281        /* item mag9r061_c */
18       0.30432        /* item mag9q081_c */
19      -0.72396        /* item mag9q101_c */
20       0.22162        /* item mag9q021_c */
21      -0.31432        /* item mag9v091_c */
22       0.07985        /* item mag9q211_c */
23      -0.06128        /* item mag9v13s_c step 1 */
24      -0.62938        /* item mag9v13s_c step 2 */
25      -1.08305        /* item mag9r25s_c step 1 */
26       0.89501        /* item mag9r25s_c step 2 */
```

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