NEPS SURVEY PAPERS

Linda Haschke, Jana Kähler, and Inga Hahn

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 6 FOR ADULTS

NEPS Survey Paper No. 19
Bamberg, March 2017

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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

NEPS Technical Report for Science:

Scaling Results of Starting Cohort 6 for Adults

Linda Haschke, Jana Kähler & Inga Hahn,
Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the lead author:

haschke@ipn.uni-kiel.de

Bibliographic data:

Haschke, L. I., Kähler, J. & Hahn, I. (2017): NEPS Technical Report for Science: Scaling Results of Starting Cohort 6 for adults (NEPS Survey Paper No. 19). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/ NEPS:SP19:1.0

NEPS Technical Report for Science: Scaling Results of Starting Cohort 6 for Adults

Abstract

The National Educational Panel Study (NEPS) investigates the development of competences across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the scientific literacy test in starting cohort 6 (adults). The science test for adults contained 24 multiple choice and two complex multiple choice items that covered two knowledge domains as well as three different contexts. The test was administered to 6,665 adults. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test's dimensionality were evaluated to ensure the quality of the test. The results illustrated good item fit values and measurement invariance across various subgroups. Moreover, the test showed a high reliability. As the correlations between the two knowledge domains were very high in a multidimensional model, the assumption of unidimensionality seemed adequate. Limitations of the test pertained to the lack of very difficult items. However, the results emphasized the good psychometric properties of the science test, thus supporting the estimation of reliable scientific literacy scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the ConQuest syntax for scaling the data.

Key words

scientific literacy, adults, differential item functioning, item response theory, scaling, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, and information and communication technologies literacy. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper the results of these analyses are presented for scientific literacy in starting cohort 6 (adults). First, the main concepts of the scientific literacy test are introduced. Then, the scientific literacy data of starting cohort 6 and the analyses performed on the data to estimate literacy scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is presented.

The present report has been modelled along the technical reports of Pohl, Haberkorn, Hardt and Wiegand (2012) and Haberkorn, Pohl, Hardt and Wiegand (2012). Note that the analyses of this report are based on preliminary data releases. 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 the analyses in this paper. However, we do not expect fundamental changes in the presented results.

2 Testing Scientific Literacy

The framework and test development for the scientific literacy test are described by Weinert and colleagues (2011) as well as Hahn and colleagues (2013). In the following, we briefly describe specific aspects of the scientific literacy test that are necessary for understanding the scaling results presented in this paper.

The science test assesses two types of scientific sub-competencies. These are a) knowledge of science (KOS) and b) knowledge about science (KAS). Using the definition by PISA (OECD, 2007; Prenzel et al., 2007) KOS is defined as the knowledge of basic scientific concepts and facts, whereas KAS can be regarded as the understanding of scientific processes.

KOS is divided into content-related components: matter, system, development and interaction. KAS is divided in the process-related components scientific enquiry and scientific reasoning. KAS and KOS are implemented in three contexts: health, environment, and technology. The test items are organized in units (testlets). Thus, one unit consists of two or three items. Each unit refers to one context-component combination.

There are two types of response formats. These are simple multiple choice (MC) and complex multiple choice (CMC) in the special form of true false items. In MC items the test taker has to identify the correct answer out of four response options. In CMC items, the test taker has to decide for each response option whether the answer is correct or not.

sca40420_c (t-value >10), sca40120_c sca40140_c (both too easy in an already too easy test, almost 100% answering the items right, no), sca40930_c (trennschärfe < .20)

The scientific literacy test that was administered in the present study included 26 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses identified a poor fit for four items: item sca50420_c had a t-value >10.0, items sca50120_c and sca50140_c were both too easy (-3.58 logits and -2.72) for the sample and the discrimination of item sca50930 was <.20. Therefore, these items were removed from the final scaling procedure. Thus, the analyses presented in the following sections and the literacy scores derived for the respondents are based on the remaining 22 items.

3 Data

3.1 The design of the study

The study assessed different competence domains including, among others, scientific literacy and computer literacy. The competence tests for these domains were always presented first within the test battery. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the science test was either administered as the first or the second test (i.e., after the computer literacy test). The test time for the scientific literacy test was 25 minutes, with one additional minute for the procedural metacognition item. There was no multi-matrix design regarding the choice and order of the items within a test. All adults received the same test items in the same order.

The scientific literacy test for adults originally consisted of 26 items. As mentioned above, only 22 items met the quality standards. The characteristics of these 22 items are depicted in Table 1. Table 2 shows how the items cover the different contents and components of the science framework (see Hahn et al., 2013) whereas Table 3 refers to the different response formats.

Table 1: Number of items for the different contents of the science test for adults

Knowledge domains	**Frequency**
Knowledge of Science (KOS)	14
Knowledge about Science (KAS)	8
Total number of items	**22**

Table 2: Number of items for the different contexts of the science test for adults

Context	**Frequency**
Health	9
Environment	4
Technology	9
Total number of items	**22**

Table 3: Number of Response formats for the different contexts of the science test for adults

Response format	**Frequency**
Simple Multiple-Choice	20
Multiple True False	2
Total number of items	**22**

3.2 Sample

The science test was administered to 6,665 participants. Six persons had to be excluded from the analyses because no valid person identifiers could be assigned to them. Moreover, three persons were excluded because they had less than three valid answers in the science test. Because no reliable ability scores can be estimated based on such few valid responses, these cases were exclude from the analyses (see Pohl & Carstensen, 2012). Thus, the scaling analyses were carried out with a data set that included 6,656 persons (see section 5). About half of the sample (3,301 persons) received the science test first, whereas 3,355 persons received the science test after completing the computer literacy test.

4 Analyses

4.1 Missing responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and, finally, e) multiple kinds of missing responses that occur in an item and are not determined. In this study, all persons received the same set of items. As a consequence, there were no items that were not administered to a person.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labelled as a not-determinable missing response.

Missing responses provide information on how well a test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

4.2 Scaling model

Item and person parameters were estimated in ConQuest (Wu, Adams & Wilson, 1997) using a partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item[1]. Categories of the polytomous variables with less than 200 responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items, especially when the item consisted of many subtasks. In these cases, the lower categories were collapsed into one category. For the two CMC items sca5652s_c and sca5091s_c the two lowest categories were collapsed. To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Haberkorn, Pohl, Carstensen & Wiegand, 2012; and see Pohl & Carstensen, 2013, for studies on the scoring of different response formats and the handling of missing values).

Ability estimates for scientific literacy were derived as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF are described in section 7.

4.3 Checking the quality of the scale

The adults' science test was specifically constructed to be implemented in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was evaluated in pilot studies but also checked in several analyses for the data from the main study.

[1] As described later, due to collapsing of categories, this interpretation does not necessarily hold for the variables in the SUF.

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

MC and CMC items consisted of one correct response and a number of distractors (incorrect response options). We investigated whether these distractors worked well, that is, whether they were chosen by the adults with a lower general ability in science more often than by those with a higher general ability in science. Thus, we evaluated the point-biserial correlation of giving a certain incorrect response and the total number correct score estimated in the analysis treating all subtasks of CMC items as single items. Negative correlations indicated good distractors, whereas correlations between .00 and .05 were considered acceptable and correlations above .05 were viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit and items with a WMNSQ > 1.2 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the total correct score (equal to the discrimination as computed in ConQuest) above .30 were considered as good, above .20 as acceptable, and below .20 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The science literacy test should measure the same construct for all adults. If some items favored certain subgroups (e.g., they were easier for men than for women), measurement invariance would be violated and a comparison of literacy scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. Test fairness was investigated for the variables test position, gender, age, the number of books at home (as a proxy for socio-economic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). In order to test for measurement invariance, differential item functioning (DIF) 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 logit as very strong DIF, absolute differences between 0.6 and 1 noteworthy of further investigation, and differences smaller than 0.4 as negligible DIF. Additionally, model fit was investigated by comparing a model including differential item functioning to a model that only includes main effects and no DIF.

The science test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data using the software mdltm (von Davier, 2005). Model fit indices of the PCM and GPCM were compared to evaluate the two models.

The science test was constructed to measure a unidimensional science literacy score (Hahn et al., 2013). The assumption of unidimensionality was, nevertheless, tested by specifying a two-dimensional model with KAS items representing one and KOS items the other dimension. The correlation between the subdimensions as well as differences in model fit between the unidimensional model and the two dimensional model were used to evaluate the unidimensionality of the scale.

5 Results

5.1 Descriptive statistics of the responses

In order to a) get a first rough descriptive measure of item difficulty and b) check for possible estimation problems, before performing IRT analyses we evaluated the relative frequency of the responses given. The percentage of persons correctly responding to an item (relative to all valid responses) ranged from 27.0% to 85.5% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 43.1% and 47.5%. From a descriptive point of view, the items covered a rather wide range of difficulties. However, there were no very difficult items as the majority of items showed low or medium difficulties. The mean item difficulty of -0.81 (*SD* = 0.03) showed that the test was a bit too easy for the sample as compared to the mean person ability (fixed at zero).

5.2 Missing responses

5.2.1 Missing responses per person

The number of invalid responses per person is shown in Figure 1. The number of not-valid responses was quite small. For 74.7 % of the persons, all answers were valid.

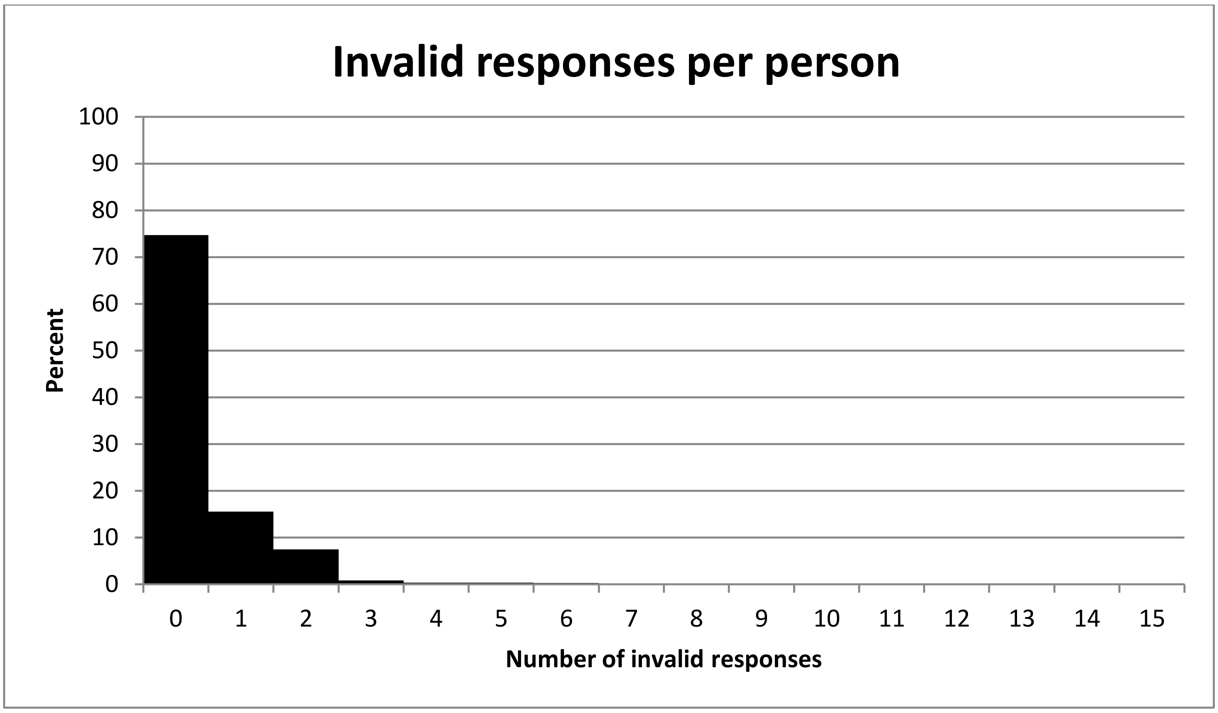

Figure 1: Number of invalid responses per person

The number of omitted responses per person is depicted in Figure 2. 71.3 percent of the persons did not omit a single item. Only 10.4% omitted three or more than three items.

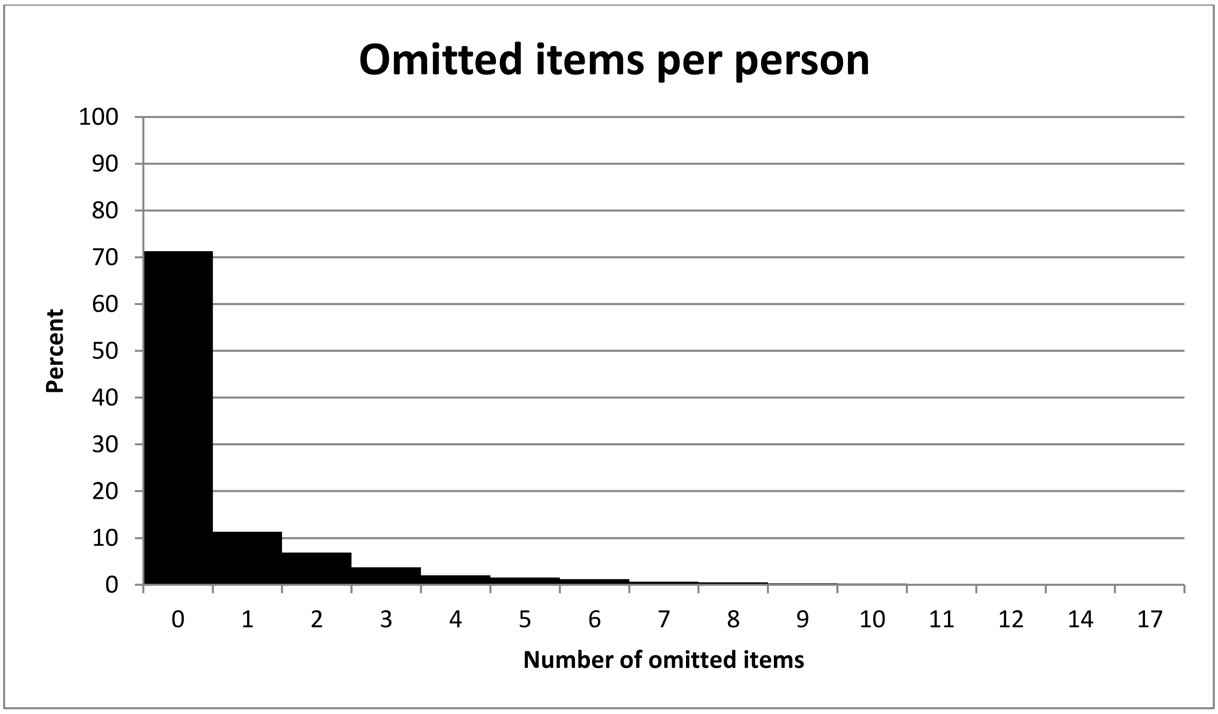

Figure 2: Number of omitted items per person

Most adults reached the end of the test (67.9%) and only a small proportion did not manage to finish at least two thirds of the test (see figure 3).

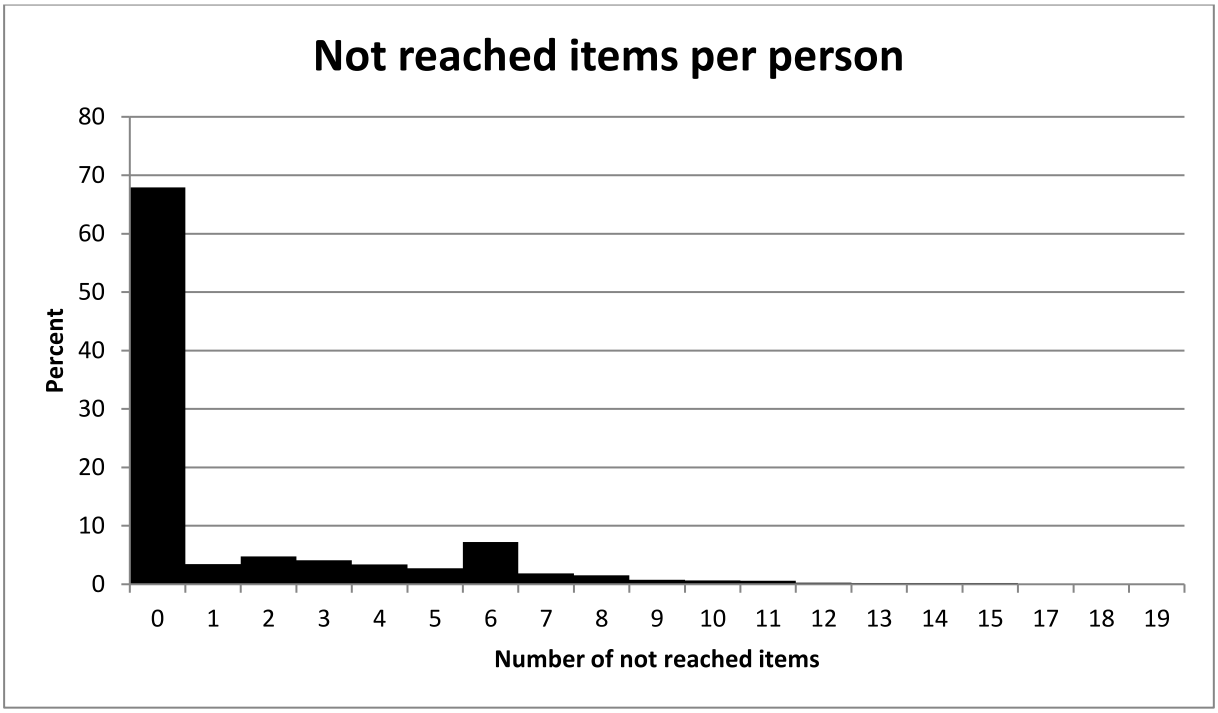

Figure 3: Number of not reached items per person

Figure 4 shows the total number of missing responses per person. The total number of missing responses is the sum of invalid, omitted and not reached missing responses. 41.0% of the adults answered all questions and, consequently, had no missing responses. Only 2.8% of the adults had missing responses on more than half of the items. The amount of missing responses per person can be classified as very small.

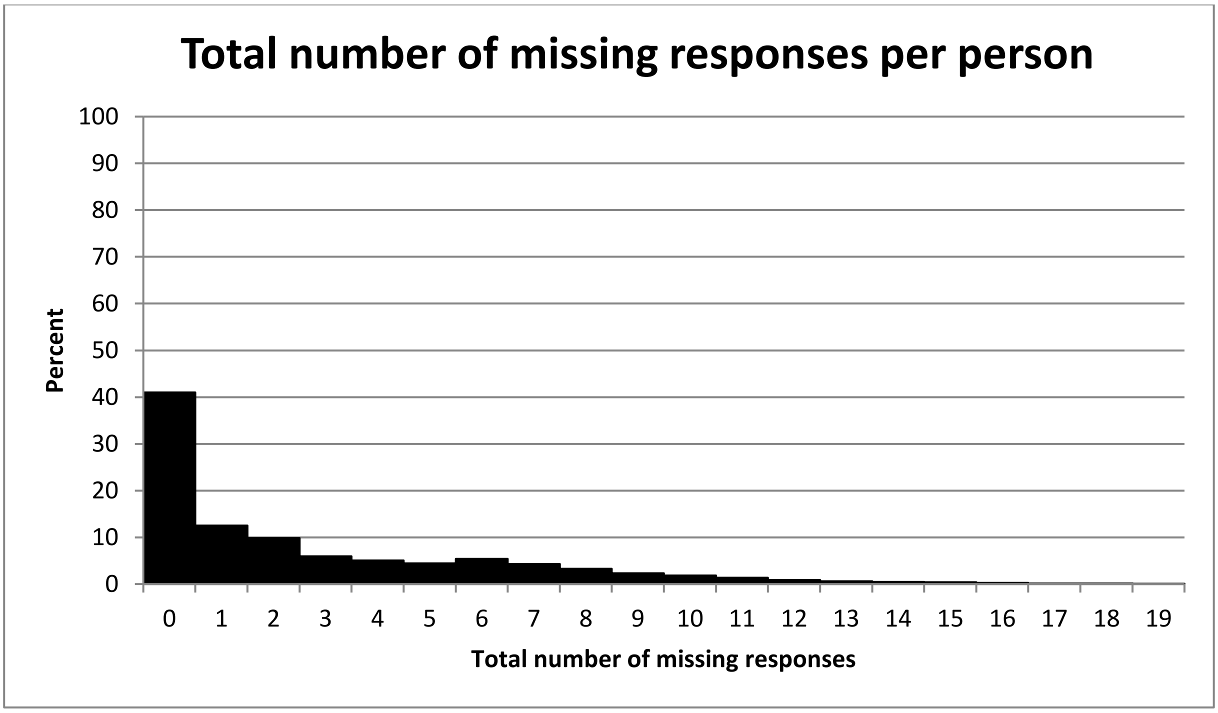

Figure 4: Total number of missing responses per person

5.2.2 Missing responses per item

Table 4 shows the number of valid responses for each item, as well as the number and percentage of missing responses. Overall, the number of persons that omitted an item was small. There was only one item with an omission rate above 10% (item sca56020_c). The number of missing responses was correlated at *r* = 0.10 (*p* = .65) with the difficulty of the item. Because the correlation was rather small, this result indicates that the test takers did not omit items that were more difficult. The number of invalid responses per item was small. The largest number was 14.1% for item sca5652s_c. The relative frequency of not reached items increased towards the end of the test. Eventually, 32.1% of the adults did not reach the last item and thus did not complete the test. The total number of missing responses per item varied between 1.5% (sca41110_c) and 32.3% (sca41030_c).

5.3 Parameter estimates

5.3.1 Item parameters

Column 2 in table 5 shows the percentage of correct responses in relation to all valid responses for each item. Note that since there is a non-negligible amount of missing responses, this probability cannot be interpreted as an index for item difficulty. The percentage of correct responses within items varied between 27.0% and 85.5% with an average of 63.8% correct responses.

The estimated item difficulties (for dichotomous items, MC items) and location parameters (for polytomous variables, CMC items) are given in Table 5. The step parameters (for polytomous variables) are depicted in Table 6. Because for the two CMC items sca5652s_c and sca5091s_c the two lowest categories were collapsed, these items were scaled using a scoring of 0, 0.5, 1, and 1.5. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -2.01 (sca56030_c) to 1.28 (sca51020_c). In total the estimated item difficulties had a mean of -0.81. Due to the large sample size, the standard errors of the estimated item difficulties was very small (*SE*(ß) ≤ 0.042). Overall, the item difficulties were rather low; the test did not include items with a high difficulty (above 2.5 logits).

Table 4: Valid Responses and Missing Values

Variable name	Number of valid responses	Position in the test	Relative frequency of not reached items %	Relative frequency of omitted items %	Relative frequency of invalid responses %
sca56120_c	6,458	1	0.0	1.4	1.6
sca56130_c	6,339	2	0.0	0.6	4.2
sca51110_c	6,553	3	0.0	0.7	0.8
sca51140_c	6,360	4	0.0	3.7	0.7
sca50410_c	6,533	5	0.0	1.3	0.5
sca5652s_c	5,078	7	0.0	9.6	14.1
sca56540_c	6,405	8	0.0	2.9	0.8
sca51430_c	6,495	11	0.3	1.7	0.4
sca51440_c	6,223	12	0.5	5.1	0.9
sca50210_c	6,048	13	0.7	8.3	0.1
sca50220_c	6,236	14	1.0	4.7	0.6
sca50710_c	6,298	15	1.6	3.5	0.3
sca50720_c	5,977	16	2.3	7.5	0.4
sca56310_c	6,334	17	3.0	1.1	0.7
sca56320_c	6,058	18	4.6	3.8	0.7
sca5091s_c	5,250	19	6.4	0.8	13.9
sca56020_c	5,057	21	13.6	10.2	0.2
sca56030_c	5,169	22	16.3	5.9	0.1
sca50520_c	5,251	23	19.7	0.7	0.7
sca50530_c	4,925	24	23.9	1.7	0.5
sca51020_c	4,706	25	28.6	0.4	0.3
sca51030_c	4,508	26	32.1	0.0	0.2

Note. The items on position 6, 9, 10 and 20 were excluded from the analyses due to insufficient item quality (see section 5.1).

Table 5: Item parameters

Variable name	Percentage correct	Difficulty/location parameter	*SE* (difficulty/ location parameter)	Weighted MNSQ	*t*-value	Pt.bis of correct response	Discrimination (2PL)
sca56120_c	37.3	0.633	0.028	0.97	-2.9	0.49	1.13
sca56130_c	79.5	-1.601	0.033	1.05	2.9	0.32	0.68
sca51110_c	81.8	-1.774	0.034	0.92	-4.0	0.49	1.83
sca51140_c	76.7	-1.407	0.032	1.06	3.9	0.35	0.73
sca50410_c	77.4	-1.462	0.032	0.97	-1.8	0.46	1.25
sca5652s_c	n.a.	-1.634	0.034	0.93	-3.9	0.43	0.75
sca56540_c	52.3	-0.100	0.028	1.01	1.4	0.46	0.95
sca51430_c	81.6	-1.758	0.034	1.04	2.3	0.33	0.75
sca51440_c	63.2	-0.628	0.029	1.09	7.7	0.37	0.63
sca50210_c	62.5	-0.585	0.029	0.95	-4.5	0.52	1.32
sca50220_c	71.5	-1.073	0.030	1.01	0.7	0.43	1.03
sca50710_c	76.4	-1.377	0.032	1.01	0.6	0.41	1.06
sca50720_c	35.7	0.757	0.030	1.00	0.0	0.44	0.97
sca56310_c	75.2	-1.306	0.031	1.03	1.7	0.38	0.80
sca56320_c	35.4	0.769	0.029	0.96	-3.0	0.49	1.18
sca5091s_c	n.a.	-1.697	0.035	1.00	0.1	0.29	0.45
sca56020_c	71.4	-0.984	0.034	1.03	1.8	0.41	0.88
sca56030_c	85.5	-2.008	0.042	0.99	-0.5	0.38	1.13
sca50520_c	61.4	-0.501	0.031	0.98	-2.0	0.50	1.13
sca50530_c	43.0	0.416	0.032	0.94	-4.9	0.51	1.22
sca51020_c	27.0	1.276	0.036	1.04	2.3	0.37	0.80
sca51030_c	82.0	-1.721	0.041	0.97	-1.4	0.44	1.29

Note. *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ. Percent correct scores are not informative for polytomous CMC and MA item scores. These are denoted by n.a. For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

Table 6: Step parameters for the CMC items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3
sca5652s_c	-0.540 (0.029)	0.015 (0.030)	0.526
sca5091s_c	-0.582 (0.028)	-0.253 (0.029)	0.835

5.3.2 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. For these analyses, the mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.003, indicating that the test had good potential to differentiate between persons. The reliability of the test (WLE reliability = .720) was acceptable. The mean of the item distribution was about 0.81 logits below the mean person ability distribution. The amount to which the item difficulties and location parameters were targeted to the ability of the persons is shown in Figure 5. In the right panel, the estimated item difficulties are given. Subjects with an ability corresponding to the difficulty of an item have a probability of 50% of correctly responding to this item. As a consequence the item information is highest for subjects with an ability that corresponds to the difficulty of the item. Figure 5 shows that the items covered a wide range of the persons' ability distribution. However, only few items covered medium person abilities and there were no items available for persons with high science ability. Instead, the majority of items were easy or of medium difficulty. As a consequence, persons with a medium and low ability will be measured relatively precisely with a low standard error while ability estimates for adults with higher science ability will have a larger standard error.

Scale in logits	Person ability	Item difficulty
		Each ‚X‘ represents 41.9 cases
3		
	X	
	X	
	XX	
2	XX	
	X	
	XX	
	XXX	
	XXX	
	XXX	21
	XXXX	
1	XXXXXX	
	XXXXX	
	XXXXX	13 15
	XXXXXXX	1
	XXXXXXXX	20
	XXXXXXXX	
	XXXXXXXXX	
	XXXXXXXX	
0	XXXXXX	7
	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXX	19
	XXXXXXXXXX	9 10
	XXXXXXXX	
	XXXXXXX	
-1	XXXX	17
	XXXX	11
	XXXXX	14
	XXX	4 5 12
	XXX	2
	XX	6 16 22
	X	3 8
-2	X	
	X	18
	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 41.9 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see table 4).

5.4 Quality of the test

5.4.1 Fit of the subtasks of complex multiple-choice items

Before the responses on the subtasks of CMC items were aggregated and analyzed via a PCM, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC in a Rasch model. No estimation problems occurred and all subtasks showed a satisfactory item fit. The WMNSQ ranged from 0.92 to 1.09, the respective *t*-value from -4.9 to 7.7. There were no unacceptable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Hence, an aggregation of polytomous variables seemed to be justified. In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the adults' total correct score. All distractors had point-biserial correlations with the total score below zero. These results indicate that the distractors worked well.

5.4.2 Item fit

Regarding the MC and the aggregated CMC items the fit was very good. WMNSQs were close to 1 with the lowest value being 0.92 (item sca51110_c) and the highest being 1.09 (item sca51440_c). Overall, there were no items with a WMNSQ above 1.1. None of the items showed a *t*-value above 8 and the item characteristic curves of these items showed a good fit. Hence, no indications for a heavy misfit of these items could be detected and, therefore, they were kept in the analysis for estimating the scientific literacy scores.

5.4.3 Differential item functioning

We checked for test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). DIF was investigated for the variables test position, gender, age, the number of books at home (as a proxy for socio-economic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 7 shows the difference between the estimated item difficulties in different groups. Male vs. female, for example, indicates the difference in difficulty ß(male) – ß(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females.

The scientific literacy test was administered in two different positions (see section 3.1 for the design of the study). 3,355 adults (50.4%) received the computer literacy test first and then the science test, while 3,301 persons (49.6%) received the scientific literacy test before completing the computer literacy test. The adults were randomly assigned to either of the two design groups. The results showed a small average effect of test position (see Table 7). There was small DIF due to the position of the test in the booklet. Item sca56120_c exhibited the highest DIF with an absolute difference in difficulty of 0.478 logits. The differences between the two design groups were small (main effect = 0.081, Cohen's d = 0.082).

DIF was also investigated for gender. 3,365 (50.6%) of the test takers were female and 3,290 (49.4%) were male. On average, male adults had slightly higher scores in scientific literacy than female adults (main effect = 0.486 logits, Cohen's d = 0.494). There were four items showing considerable DIF (items sca5091s_c, sca50520_c, sca51020_c and sca51020_c). However since these items only showed DIF in the gender category and since there were some items with DIF in favor of men and some items with DIF in favor of women none of the items had to be removed.

The number of books at home was used as a proxy for socio-economic status. There were 2,179 (32.7%) test takers with 0 to 100 books at home, 4,031 (60.6%) test takers with more than 100 books at home and 446 (6.7%) test takers that did not give a valid response. DIF was investigated using these three groups. There were considerable mean differences between the three groups. Participants with 100 or less books at home on average had a 0.624 logits (Cohen's d =-0.658) lower scientific literacy score than participants with more than 100 books. Participants without a valid response on the variable 'books at home' performed 0.194 logits (Cohen's d =-0.201) better than participants with up to 100 and 0.430 logits (Cohen's d =0.442) worse than participants with more than 100 books, respectively. There was no considerable DIF comparing participants with many or fewer books (largest absolute difference = -0.444). Comparing the group without valid responses to the two groups with valid responses, absolute DIF occurred up to 0.474 logits (item sca5091s_c) which is still no considerable DIF.

There were 5,563 (83.6%) participants without a migration background and 1,092 (16.4%) participants with a migration background. These two groups were used for investigating DIF for migration. There was a considerable difference in the mean performance of participants with or without migration background (main effect = 0.246 logits, Cohen's d = 0.246). Participants without a migration background had a higher scientific literacy than participants with a migration background. None of the items showed considerable DIF. The highest absolute DIF value amounted to 0.256 logits (item sca50410_c).

DIF was also investigated for age. There were 1,316 participants aged from 27 to 40 years (19.8%), 4,452 participants aged from 41 to 62 years (66.9%), and 887 participants, which were older than 63 years (13.3%). DIF was investigated using these three groups. There were considerable differences in the average performance between the three groups. Participants aged from 27 to 40 years, on average, showed a 0.340 logits (Cohen's d = 0.343) higher scientific literacy score than participants aged from 41 to 62 years or a 0.898 logits (Cohen's d = 0.927) higher score than participants older than 63 years. Participants aged from 41 to 62 years reach a scientific literacy score of 0.562 logits (Cohen's *d* =0.585). There were some items showing considerable DIF >0.6, one item (sca51440_c) in the < 40 vs. 40-62 analysis and six items (sca51440_c, sca56120_c, sca51140_c, sca50710_c, sca50720_c, sca56320_c) in the < 40 vs. > 62 analysis. These DIFs might be due to very different learning environments and contents the age groups encountered throughout their lives which might have resulted in single items being easier or more difficult for persons from extremely different age groups although they have the same person ability.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF with those that allow only for main effects. In Table 8, the models including only main effects are compared with those that additionally estimate DIF. Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC, Schwarz, 1978) were used for assessing the models. Using the AIC the models considering DIF are favored for all four DIF variables. The BIC takes the number of estimated parameters into account and, thus, prevents from overparameterization of models. Using BIC, the more parsimonious model including only the main effect is preferred over the more complex DIF model for the DIF variables *position* of the test, *books* and *migration background*. For the DIF variables *gender* and *age* the more complex DIF models have slightly better BIC values.

Table 7: Differential item functioning (differences between difficulties)

Item	**Booklet**	**Gender**	**Books**			**Migration status**	**Age**		
	Position 1 vs. Position 2	Male vs. Female	<100 vs. >100	<100 vs. Missing	>100 vs. Missing	Without vs. With	<40 vs. 40-62	<40 vs. >62	40-62 vs. >62
sca56120_c	0.239	0.327	0.106	0.116	0.007	-0.014	-0.187	-0.427	-0.237
sca56130_c	0.119	0.281	-0.067	-0.088	-0.024	-0.123	0.035	0.213	0.181
sca51110_c	0.089	-0.120	0.069	-0.076	-0.148	0.018	-0.083	-0.257	-0.171
sca51140_c	-0.065	0.149	-0.004	0.012	0.013	-0.007	0.065	0.318	0.256
sca50410_c	0.070	0.002	-0.014	0.037	0.048	0.128	-0.140	-0.217	-0.075
sca5652s_c	0.051	0.061	0.103	0.070	-0.041	-0.015	-0.046	-0.220	-0.179
sca56540_c	-0.145	-0.200	-0.071	-0.051	0.017	0.036	-0.155	-0.117	0.040
sca51430_c	-0.024	-0.076	-0.100	-0.008	0.089	-0.078	0.207	0.281	0.077
sca51440_c	-0.002	-0.140	-0.082	-0.043	0.037	-0.148	0.336	0.590	0.258
sca50210_c	-0.079	-0.065	0.034	-0.014	-0.050	-0.010	-0.011	-0.142	-0.128
sca50220_c	-0.025	0.047	-0.051	0.020	0.069	0.046	-0.085	-0.053	0.035
sca50710_c	0.025	-0.099	0.113	0.004	-0.113	0.055	0.258	0.525	0.270
sca50720_c	-0.026	-0.222	0.009	0.045	0.033	0.030	0.201	0.304	0.106
sca56310_c	0.022	-0.046	-0.075	-0.069	0.003	-0.004	-0.124	-0.191	-0.064
sca56320_c	0.033	-0.094	0.025	0.025	-0.002	0.075	-0.136	-0.421	-0.282
sca5091s_c	-0.147	0.408	0.119	0.237	0.116	-0.066	-0.020	-0.060	-0.040
sca56020_c	-0.016	0.072	-0.005	-0.043	-0.040	0.069	0.011	-0.009	-0.018
sca56030_c	-0.032	-0.093	0.000	-0.062	-0.065	-0.033	-0.042	0.083	0.127
sca50520_c	-0.037	0.538	0.092	0.081	-0.015	0.033	-0.030	-0.194	-0.162
sca50530_c	-0.017	-0.030	-0.031	-0.013	0.042	0.040	-0.071	-0.193	-0.120
sca51020_c	-0.017	-0.498	-0.222	-0.182	0.038	-0.002	-0.046	0.074	0.122
sca51030_c	-0.040	-0.462	0.032	-0.067	-0.101	-0.087	0.111	0.303	0.194
Main effect	0.082	0.486	0.624	0.194	0.430	0.247	0.340	0.898	0.562

Table 8: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	151,056.586	28	151,112.586	151,303.078
	DIF	150,895.927	50	150,995.927	151,336.091
Gender	main effect	150,737.954	28	150,793.954	150,984.441
	DIF	149,611.392	50	149,711.392	150,051.548
Age	main effect	150,707.098	29	150,765.098	150,962.389
	DIF	150,091.203	73	150,237.203	150,733.831
Books	main effect	150,618.935	29	150,676.935	150,874.230
	DIF	150,486.909	73	150,632.909	151,129.548
Migration	main effect	150,996.968	28	151,052.968	151,243.455
	DIF	150,945.984	50	151,045.984	151,386.140

5.4.4 Rasch-homogeneity

In order to test for the assumption of Rasch-homogeneity the 22 items were scaled with the GPCM. The estimated discrimination parameters are depicted in the last column in table 5. They ranged from 0.45 (item sca5091s_c) to 1.83 (item sca51110_c). The discriminations differed considerably among the items. The average discrimination parameter fell at 0.43. The GPCM (BIC = 150,885.97, number of parameters = 59) fitted the data slightly better than the PCM (BIC = 151,064.99, number of parameters = 27). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.4.5 Unidimensionality of the test

The unidimensionality of the test was investigated by specifying a one- and a two-dimensional model. The first model was based on the assumption that scientific literacy is a one-dimensional construct reflecting one distinct competence whereas the second model distinguished between the two sub-competencies *knowledge about science* and *knowledge of science* (for more details see Hahn et al., 2013). For estimating a two-dimensional model based on the Gauss Hermite quadrature estimation implemented in ConQuest was used (n=30 nodes were chosen so that stable parameter estimations could be obtained). The two-dimensional model (BIC= 151,317.07, number of parameters = 29) fitted the data worse than the unidimensional model (BIC = 151,302.67, number of parameters = 27; correlations of the two dimensions: 0.972). Consequently, scientific literacy as measured by this test can be regarded as unidimensional and therefore this simpler model was used for estimating competence scores.

6 Discussion

The analyses in the previous sections aimed at providing information on the quality of the science test for adults and at describing how the scientific literacy score was estimated. The amount of invalid responses and not-reached items was low. However, some items showed higher omission rates, although, in general, the amount of omitted items was acceptable. The test had an acceptable reliability and distinguished well between test takers of average and low scientific literacy, but not so well for high performers. There was a lack of very difficult items; hence, test targeting was somewhat suboptimal and the test measured scientific literacy of high-performing adults less accurately. Various criteria indicated a good fit of the items to the PCM. Also, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with total score) were acceptable. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test was fair for the considered subgroups. A unidimensional PCM yielded a better model fit than a two-dimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas). Hence, the unidimensional model was used for estimating scientific literacy scores. Summarizing the results, the test had good psychometric properties that facilitate the estimation of a unidimensional scientific literacy score.

7 Data in the Scientific Use file

The SUF contains all 26 items of the test of which 24 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. For the two polytomous variables (CMC items) scores indicate the (partial) credit. The MC items are marked with a '0_c' at the end of the variable name, whereas the CMC items end in 's_c'. For further details on the naming conventions of the variables see Fuß and colleagues (2016). Note that the value of the polytomous variable does not necessarily indicate the number of correctly responded subtasks (see section 4.2 aggregation of CMC items). In the scaling model each category of CMC items was scored with 0.5 points.

In the SUF manifest scale scores are provided in the form of WLE estimates (sc_1) including the respective standard error (sc_2).

For the estimation of the WLE scores, the effect of test position in the booklet is controlled for. The ConQuest Syntax for estimating the WLE scores is provided in Appendix A. Adults that did not take part in the test or those that do not have enough valid responses show a non-determinable missing value on the WLE score for scientific literacy.

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

References

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

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2016). Incorporating different response formats in the IRT-scaling model for competence data. In H.-P. Blossfeld, J. Skopek & J. Maurice (Eds.). Methodological issues of longitudinal surveys. Wiesbaden: Springer VS.

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.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., …& Prenzel, M. (2013). Assessing scientific literacy over the lifespan: A description of the NEPS science framework and the test development. Journal of Educational Research Online, 5(2), 110–138.

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

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

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

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*. PISA: OECD Publishing.

Pohl, S. & Carstensen, C. H. (2012). NEPS technical report – Scaling the data of the competence tests. (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S. & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. Journal for Educational Research Online/Journal für Bildungsforschung Online, 5(2), 189-216.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). Technical Report of Reading– Scaling Results of Starting Cohort 3 in Fifth Grade. (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C., & Hamann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In PISA-Konsortium Deutschland (Ed.), PISA 2006 - Die Ergebnisse der dritten internationalen Vergleichsstudie (S. 63-105). Münster: Waxmann.

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

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

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika, 54, 427-450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of Competencies Across the Life Span. In H. P. Blossfeld, H. G. Roßbach & J. von Maurice (Eds.). *Zeitschrift für Erziehungswissenschaften, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in starting cohort VI

```
Title Starting Cohort VI, SCIENCE: Partial Credit Model;

datafile B69_C_A_S_C2_suf.dat;

format id 1-7 responses 8-29

labels << Variablenname.txt;
codes 0,1,2,3;

score (0,1)       (0,1)              !item (1-5,7-15,17-22);
score (0,1,2,3)   (0,0.5,1,1.5)      !item (6,16);

set constraint=cases;
model item + item*step-position;
estimate;

show cases !estimates=wle >> B69_C_A_S_C2_suf.wle;
show cases !estimates=latent >> B69_C_A_S_C2_suf.pls;
show ! estimates=latent >> B69_C_A_S_C2_suf.shw;
itanal! estimates=latent >> B69_C_A_S_C2_suf.ita;
```