NEPS
National Educational Panel Study

NEPS Working Papers

Katrin Schöps & Steffani Saß

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

NEPS Working Paper No. 23

Bamberg, April 2013

SPONSORED BY THE
Federal Ministry
of Education
and Research

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

Katrin Schöps & Steffani Saß
Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the lead author:

schoeps@ipn.uni-kiel.de

Bibliographic data:

Schöps, K. & Sass, S. (2013): *NEPS Technical Report for Science – Scaling results of Starting Cohort 4 in ninth grade* (NEPS Working Paper No. 23). Bamberg: University of Bamberg, National Educational Panel Study.

We 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.

NEPS Technical Report for Science – 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 scientific literacy for starting cohort 4 in grade 9. 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 science test in grade 9 consisted of 28 multiple choice and complex multiple choice items and covers two knowledge domains as well as three different contexts. The test was administered to 14,475 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 illustrate 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 are very high in a multidimensional model, the assumption of unidimensionality seems adequate. Among the challenges of this test is the lack of very difficult items. But overall, the results emphasize the good psychometric properties of the science test, thus supporting the estimation of reliable scientific literacy scores. In this paper, the data available in the Scientific Use File, are described and the ConQuest-Syntax for scaling the data is provided.

Key words:

scientific literacy, 9th grade, 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. Tests have been developed for different domains including scientific literacy. 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 were developed solely for implementation in NEPS, several analyses have been 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). In this paper the results of these analyses are presented for scientific literacy in the starting cohort 4.

The present report has been modeled along the technical reports of Pohl, Haberkorn, Hardt, & Wiegand (2012) and Haberkorn, Pohl, Hardt, & 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. We do, however, not expect severe changes in results.

2. Testing Scientific Literacy

The science test aims at assessing 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 specified as knowledge of basic scientific concepts and facts, whereas knowledge about science (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 (see Hahn, Schöps et al. (in prep.) and Weinert et al., 2011, for the description of the framework). 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 find the correct answer out of four response options. In CMC items the test taker has to decide at each answer option whether the answer is correct or not.

3. Data

3.1 The design of the study

There were two testing groups which differ in the order of the tests they received. Some subjects received first the science test before completing the other tests, while other subjects received the science test after having completed the computer literacy test. The test time for the scientific literacy test was 29 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 students got the same test items in the same order.

The scientific literacy test in grade nine consists of 28 items. The characteristics of these 28 items are depicted in table 1. Table 2 is concerned with the response format whereas table 3 shows how the items cover the different contents and components of the science framework (see Hahn et al. 2012).

Table 1: Classification of the science test items for grade 9

Knowledge domains	**Frequency**
Knowledge of Science (KOS)	21
Knowledge about Science (KAS)	7
Total number of items	28

Table 2: Response formats of the science test items for grade 9

Response format	**Frequency**
Simple Multiple-Choice	19
Complex Multiple-Choice (True false items)	9
Total number of items	28

Table 3: Number of items for the different contexts of the science test for grade 9

Context	**Frequency**
Health	6
Environment	10
Technology	9
Total number of items	28

3.2 Sample

Overall 16,425 students are part of the sample. 14,475 of these students took the science tests. There were two testing groups which differ in the order of the tests they received. 7,208 persons received first the science test and then the ICT test while 7,267 persons received the science test after completing the ICT test.

All 14475 persons who took part in the science test are included in the descriptive analyses. The results are presented in the following sections.

The analyses for item fit, item parameters and differential item functioning however have been carried out with a preliminary data set that included only 14295 persons. These results will be presented in section 5.6.

4. Analyses

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 in an 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 occur, for example, when two response options are selected in simple MC items where just one is required, or when numbers or letters that are not within the range of valid responses are given as a response. Missing responses due to omitted items occur when test persons skip items. Due to time limits, it might happen that not every person finishes the test within the given time. Consequently, missing responses occur due to the fact items that items are not reached. As complex multiple choice items are aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses may be found in these items. A CMC item is coded as missing if at least one subtask contained a missing response. When just one kind of missing response occurs, the item is coded according to the corresponding missing response. When the subtasks contain different kinds of missing responses, the item is labeled as a not-determinable missing response.

Missing responses provide information on how well a test works (e.g., time limits, understanding of instructions, handling of different response formats) and they need to be accounted for in the estimation of item and person parameters. We, therefore, thoroughly investigated the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This gave an indication of 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

For estimating item and person parameters for scientific literacy, a Partial Credit Model (Masters, 1982) 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).

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[1]. If at least one of the subtasks contains a missing response, the whole CMC item was scored as missing. When categories of the polytomous variables had less than N=200 the categories 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 to one category. However, for the kindergarten test this did not apply (cell frequency was always >2%).

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 Pohl & Carstensen, 2012b, for studies on the scoring of different response formats).

Ability estimates for scientific literacy will be estimated as weighted maximum likelihood estimates (WLEs, Warm, 1989) and later also in 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 was done in order to judge how well the item difficulties are targeted to the ability of the persons. The test targeting gives some information about the precision of the ability estimates at the different levels of ability.

4.3 Checking the quality of the scale

The grade 9 science test was specifically constructed to be implemented in 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.

The responses on the subtasks of CMC items are aggregated to a polytomous variable for each CMC item. In order to justify such an aggregation, the fit of the single subtasks is checked in analyses. For this the single subtasks are separately included in a Rasch model together with the MC items and the fit of the subtasks is evaluated based on the weighted mean square error (WMNSQ), the respective t-value, point biserial correlations of the responses with total correct score and the Item Characteristic Curve. Only if the subtasks have a satisfactory item fit, they were used to construct polytomous CMC item variables.

In MC and CMC items consisted of one correct response and a number of distractors (incorrect response options). We investigated whether the distractors worked well, that is, whether they are chosen by the students with a lower general ability in science more often

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

than by those with a higher general ability in science. For this 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. 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, point biserial correlations of the correct responses 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 further investigated. Point biserial correlations of the correct responses with the total score 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 science literacy test that measures the same construct for all students. If there are items that favor certain subgroups (e.g., that are easier for boys than for girls), measurement invariance would be violated and a comparison of literacy scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. Test fairness was investigated for the variables test position, gender, the number of books at home (as a proxy for socio-economic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). In order to test for measurement invariance, differential item functioning is estimated using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty are estimated. Differences in the estimated item difficulties between the subgroups are evaluated. Based on experiences with preliminary data, we consider absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF (differential item functioning), absolute differences between .6 and 1 noteworthy to further investigate, and differences smaller than .4 as no considerable DIF. Additionally model fit was investigated by comparing a model including differential item functioning to a model that only includes 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 with the generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005), and by comparing model fit indices of the 2PL model to those obtained when applying the partial credit model.

The science test is constructed to measure a unidimensional science literacy score (Hahn et al., 2012). The assumption of unidimensionality was, nevertheless, tested in the data by specifying a two dimensional model with KAS items representing one and KOS 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 Exclusion of cases from the analyses

The original data file included 16425 persons. In an initial step for calculating item parameters all persons who took part in the test were included (n= 14475). For further analyses only persons with more than two valid responses were taken into account (n=14474). The results are presented in the following sections.

5.2 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 percent of persons correctly responding to an item (relative to all valid responses) varies over items from 18.6% to 86.3% for the MC items. For the CMC items the percent of persons who correctly answered all subtasks varies from 17.5% to 45.6%. From a descriptive point of view, the items cover a relatively wide range of difficulties. However, there are no very difficult items as the majority of items show a low difficulty.

5.3 Missing responses

5.3.1 Missing responses per person

The number of non-valid responses per person is shown in Figure 1. The number of non-valid responses is very small. For 80.8 % of the persons all answers were valid.

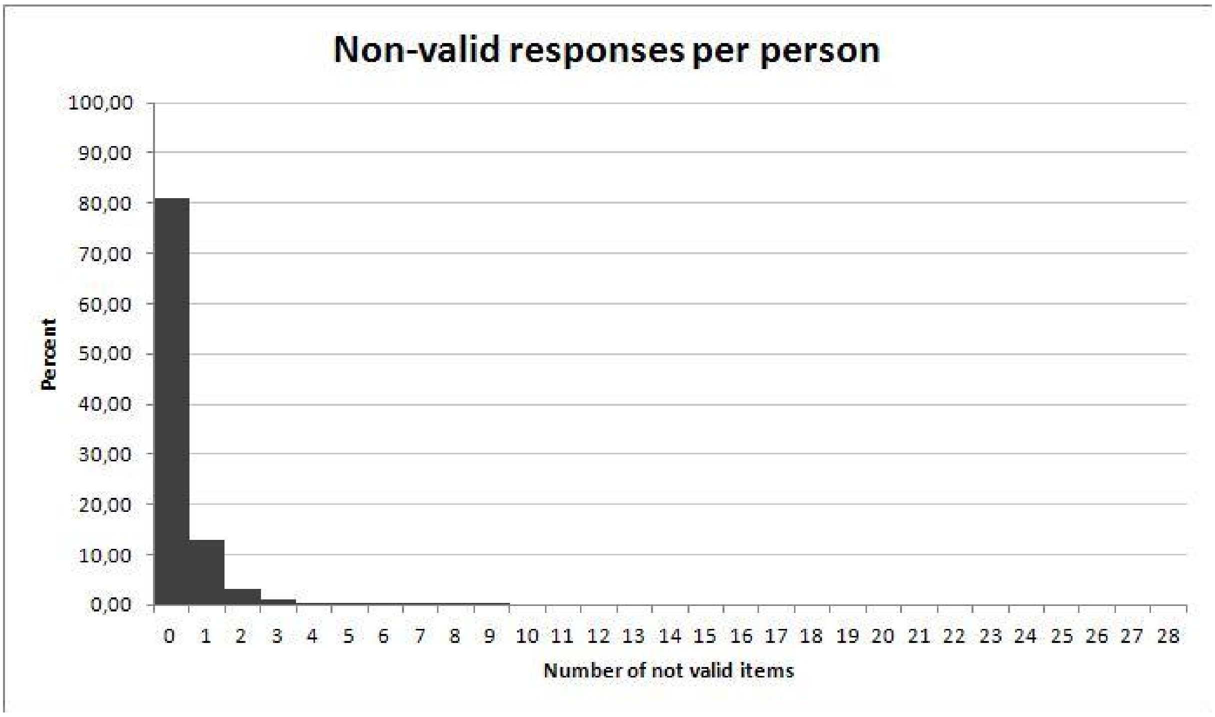

Figure 1: Number of non-valid responses

The number of omitted responses per person is depicted in Figure 2. 79.3 percent of the persons did not omit a single item. Only 3.2% omitted 3 or more than 3 items.

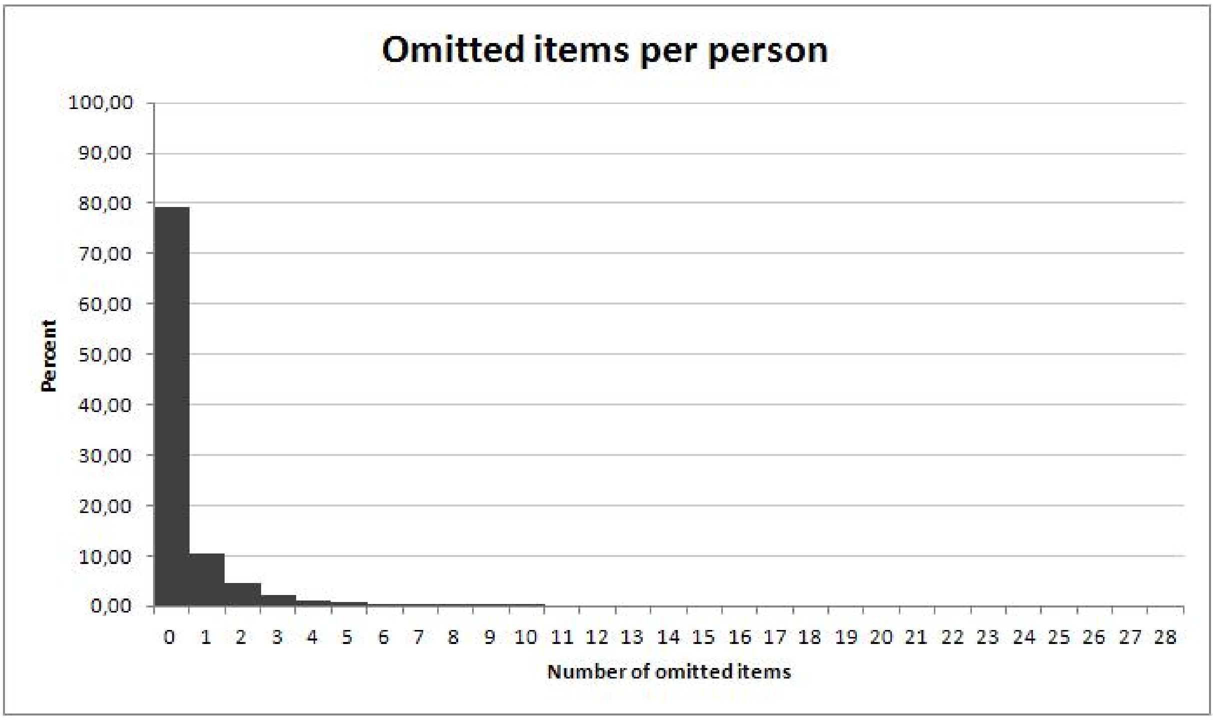

Figure 2: Number of omitted items

Most students reached the end of the test (68.5 %) and only a small proportion did not manage to finish at least two thirds of the test.

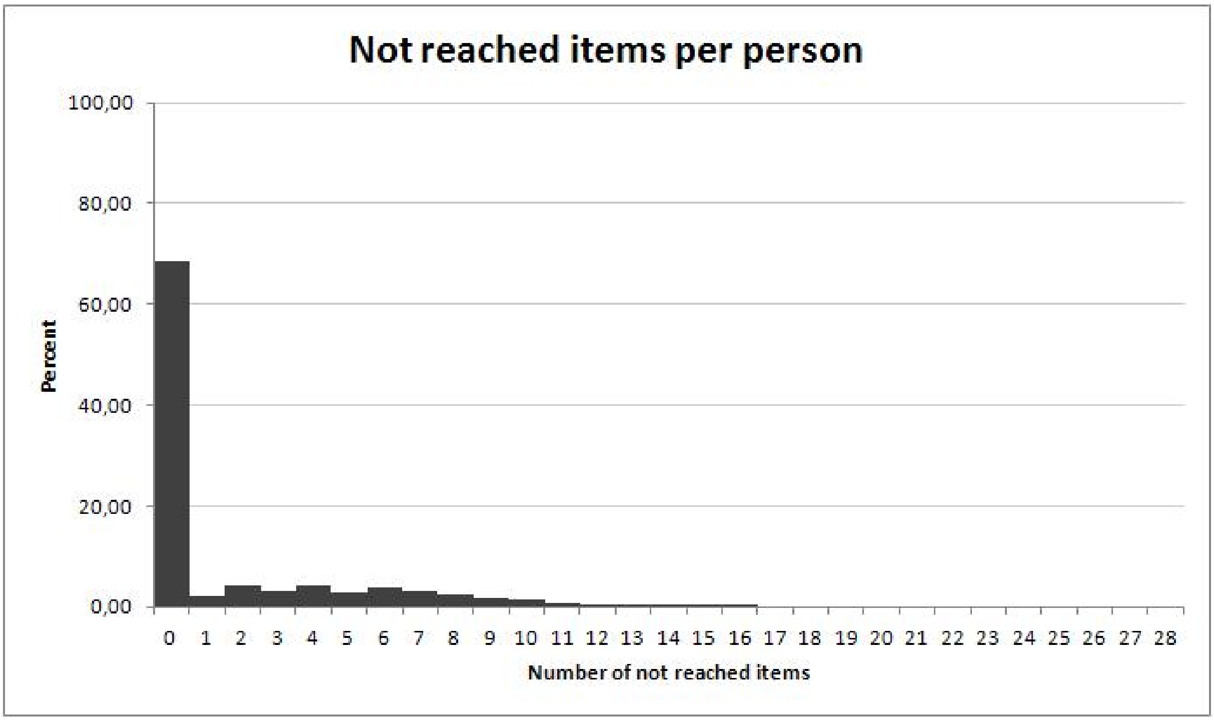

Figure 3: Number of not reached items

Figure 4 shows the total number of missing responses per person. The total number of missing responses is the sum of not valid, omitted, and not reached missing responses. 48% of the students answered all questions and consequently had no missing responses. Only 0.9% of the students have missing responses on more than half of the items. Hence the amount of missing responses per person can be classified as very small.

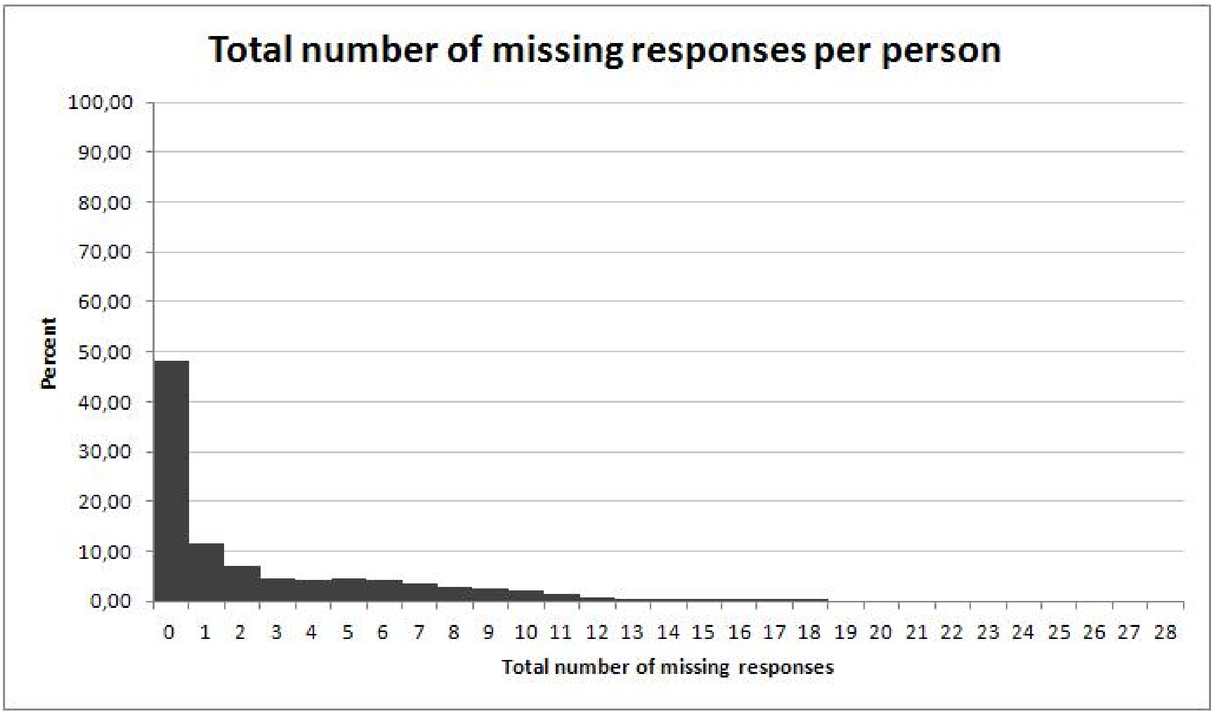

Figure 4: Total number of missing responses

5.3.2 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 omit an item is small. There is no item with an omission rate above 5%. The number of missing responses is correlated to .59 with the difficulty of the item. This result indicates that the test takers tend to omit items that are more difficult. The number of invalid responses per item is small. The highest number is 4.7% for item SCG9052s_c. The relative frequency of not reached items increases towards the end of the test. Eventually 31.5 % of the students did not reach the last item and thus did not complete the test. The total number of missing responses per item varies between 1.0% and 31.7%.

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 %
SCG90110_c	14474	1	0.00	1.50	0.00
SCG9012s_c	13910	2	0.00	0.20	3.90
SCG90510_c	14460	3	0.00	1.10	0.10
SCG9052s_c	13794	4	0.00	1.10	4.70
SCG90920_c	14387	5	0.00	1.30	0.60
SCG90930_c	14431	6	0.00	1.40	0.30
SCG9611s_c	14069	7	0.00	0.40	2.80
SCG96120_c	14431	8	0.00	1.40	0.30
SCG96410_c	14416	9	0.00	0.60	0.40
SCG96420_c	14445	10	0.00	3.50	0.20
SCG9061s_c	14271	11	0.10	4.30	1.40
SCG90630_c	14315	12	0.10	1.20	1.10
SCG90810_c	14445	13	0.20	0.80	0.20
SCG9083s_c	14069	14	0.30	1.40	2.80
SCG91030_c	14445	15	0.70	3.10	0.20
SCG91040_c	14445	16	0.90	0.90	0.20
SCG91050_c	14445	17	1.20	1.50	0.20
SCG9042s_c	13924	18	1.90	0.60	3.80
SCG9043s_c	13996	19	3.30	1.60	3.30
SCG9651s_c	14112	20	5.20	1.30	2.50
SCG96530_c	14416	21	7.60	1.50	0.40
SCG90320_c	14431	22	10.80	2.80	0.30
SCG90330_c	14445	23	14.70	5.00	0.20
SCG9621s_c	13996	24	17.70	2.10	3.30
SCG96220_c	14358	25	21.80	2.40	0.80
SCG91110_c	14431	26	24.90	1.20	0.30
SCG91120_c	14387	27	29.20	1.20	0.60
SCG91130_c	14445	28	31.50	0.00	0.20

Table 5: Item parameters

Item	Difficulty/location parameter	SE (difficulty/ location parameter)	Weighted MNSQ	Weight t-value	Pt.bis of correct response	Discrimination (2PL)
SCG90110_c	-0.608	0.019	1.06	7.5	0.35	0.81
SCG9012s_c	-2.103	0.020	0.96	-2.4	0.43	0.97
SCG90510_c	-0.641	0.019	1.03	4.1	0.37	0.60
SCG9052s_c	-0.743	0.018	1.02	1.3	0.41	0.99
SCG90920_c	-0.099	0.018	1.11	16.6	0.29	1.11
SCG90930_c	-1.027	0.020	1.02	2.1	0.37	1.17
SCG9611s_c	-1.673	0.018	1.03	2.5	0.37	1.46
SCG96120_c	-0.912	0.019	1.00	0.6	0.39	1.25
SCG96410_c	-1.752	0.023	0.99	-0.4	0.35	2.24
SCG96420_c	-0.236	0.018	0.95	-7.7	0.49	1.13
SCG9061s_c	0.346	0.019	0.95	-8.1	0.43	1.00
SCG90630_c	-0.789	0.019	0.99	-1.8	0.44	1.01
SCG90810_c	-2.199	0.026	0.92	-4.6	0.43	1.15
SCG9083s_c	-1.558	0.017	0.89	-8.9	0.56	1.92
SCG91030_c	0.030	0.018	1.00	0.1	0.43	1.12
SCG91040_c	-1.369	0.021	1.02	1.8	0.35	0.90
SCG91050_c	-0.913	0.020	1.02	2.1	0.38	0.69
SCG9042s_c	-2.478	0.018	0.99	-0.9	0.45	0.95
SCG9043s_c	-1.264	0.017	0.98	-1.7	0.48	0.64
SCG9651s_c	-1.150	0.015	1.04	3.8	0.47	0.72
SCG96530_c	-0.319	0.019	1.00	-0.3	0.43	0.74
SCG90320_c	-0.553	0.020	0.90	14.2	0.55	0.49
SCG90330_c	0.401	0.020	0.99	-0.8	0.42	1.46
SCG9621s_c	-1.215	0.018	0.94	-4.3	0.54	0.96
SCG96220_c	-0.320	0.021	1.04	5.5	0.37	0.61
SCG91110_c	0.149	0.021	1.09	12	0.31	0.66
SCG91120_c	1.160	0.024	1.02	1.8	0.35	0.49
SCG91130_c	0.835	0.023	1.09	8.8	0.28	0.78

5.4 Parameter estimates

5.4.1 Item parameters

All 28 items (including all subtasks for the polytomous items) were included in the analyses. The estimated item difficulties for polytomous variables (CMC items) and location parameters for dichotomous variables (MC items) are listed in table 5. The step parameters (for polytomous variables) are depicted in table 6. For two of the eight CMC items (items SCG9012s_c and SCG9052s_c) categories were collapsed. As these items were CMC items with a maximum score of 2, these items were scaled using the following intervals 0, 0.66, 1.33 and 2. For estimating item difficulties the mean of the ability distribution was constrained to be zero. The estimated item difficulties (or location parameters for polytomous variables) vary between -2.199 (SCG90810_c) and 1.160 (SCG91120_c) with a mean of -0.75. Due to the large sample size, the standard error of the estimated item difficulties is very small (SE(ß) ≤ 0.06). Overall, the item difficulties are low and the test includes no items with a high difficulty (above 2 logits).

Table 6: Step parameters for the CMC items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)	Step 4 (SE)
SCG9012s_c	-0.831 (0.018)	0.587 (0.020)	0.244	
SCG9052s_c	-0.425 (0.019)	-1.060 (0.018)	1.485	
SCG9611s_c	-0.542 (0.019)	-1.621 (0.018)	1.611 (0.021)	0.553
SCG9083s_c	0.060 (0.018)	-0.898 (0.017)	0.581 (0.020)	0.257
SCG9042s_c	-1.506 (0.018)	0.185 (0.018)	1.153 (0.023)	0.168
SCG9043s_c	-0.979 (0.019)	-0.263 (0.018)	0.760 (0.021)	0.482
SCG9651s_c	0.895 (0.018)	-1.493 (0.018)	1.529 (0.023)	-0.931
SCG9621s_c	-0.143 (0.020)	-0.520 (0.019)	0.210 (0.021)	0.453

5.4.2 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 is given in Pohl and Carstensen (2012a).

5.4.3 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 the analyses, the mean of ability is constrained to be zero. The variance was estimated to be 0.780, indicating that the test has good potential to differentiate between subjects. The reliability of the test (WLE reliability = .777) is good. The amount to which the item difficulties and location parameters are targeted to the ability of the persons is shown in Figure 4. The Figure shows that the items cover a great range of the ability distribution of the persons. However, only few items cover medium person ability and there are no items available for persons with high science ability. Instead the majority of items are 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 students with high science ability will have a larger standard error.

Scale in logits	Person ability	Item difficulty
		Each ‚X‘ represents 81.1 cases
3		
	X	
	X	
	X	
	X	
2	X	
	XX	
	XX	
	X	
	XXX	
	XXX	
	XXXX	
	XXXX	
1	XXXXX	
	XXXXX	
	XXXXXX	
	XXXXXXXX	19
	XXXXXX	
	XXXXXXX	
	XXXXXXXXX	
	XXXXXX	15
0	XXXXXXXXXX	23
	XXXXXXX	17
	XXXXXXXX	10
	XXXXXXX	3
	XXXXXXXXX	7
	XXXXXX	13 16
	XXXXXXXX	
	XXXXXXX	14
	XXXXXX	1 2
-1	XXXXXX	8 21
	XXXXXX	5 12
	XXXX	4
	XXXX	27
	XXX	26 28
	XXX	11
	XX	
	X	24
-2	X	6 22
	X	
		20
		9
		25
-3		

Figure 4: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 81.1 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see table 1).

5.5 Quality of the test

5.5.1 Fit of the subtasks of complex multiple-choice items

The following analyses have been carried out with the preliminary data set.

Before the responses on the subtasks of CMC items are aggregated and analyzed via a partial credit model, the fit of the subtasks is checked by analyzing the single subtasks together with the simple MC and SCR items in a Rasch model.

No estimation problems occurred and all subtasks showed a satisfactory item fit. The WMNSQ ranged from 0.90 to 1.15, the respective t-value from -9.6 to 20.7. 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 students' total score. All distractors had a point biserial correlation with the total score below zero. The results indicate that the distractors work well.

5.5.2 Item fit

Regarding the MC and the aggregated CMC items the fit is very good. WMNSQs are close to 1 with the lowest value being 0.89 (item SCG9083s_c) and the highest being 1.11 (item SCG90920_c). Overall, there are no items with a WMNSQ above 1.2. However, there were three items with a t-value above 8 (items SCG90920_c, SCG90320_c and SCG91130_c) but the item characteristic curves of these items showed a reasonable or 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.

5.5.3 Differential item functioning

We checked for test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables test position, gender, the number of books at home (as a proxy for socio-economic status), migration background, and school type (see Pohl & Carstensen, 2012a, 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). 7,175 students received the mathematics test first, then the computer literacy test, and at last the science test (position 2), while 7,119 subjects received the scientific literacy test before completing the mathematics and computer literacy test (position 1). The students were randomly assigned to either of the two design groups. Differential item functioning of the position of the test may, for example, occur if the different certain parts or items of the test are more or less tiring for the students. Regarding the items the results show a small average effect of test position (see Table 7). There is only a small DIF due to the position of the test in the booklet. The highest difference in difficulty between the two design groups is 0.124 logits.

DIF was also investigated for gender. 7,136 (49.9%) of the test takers were female and 7,158 (50.1%) were male. On average, male students have slightly higher scores in scientific literacy than female students (main effect = 0.160 logits, Cohen's d = 0.180). There is no item with a considerable gender DIF. The highest difference in difficulties between the two groups is -0.380 logits.

The number of books at home was used as a proxy for socio-economic status. There were 6,008 (42.0%) test takers with 0 to 100 books at home, 7,967 (55.7%) test takers with more than 100 books at home, and 319 (2.2%) test takers did not give a valid response.

DIF was investigated using these three groups. There are considerable average differences between the three groups. Participants with 100 or less books at home perform on average 0.716 logits (Cohen's d =0.882) lower score in scientific literacy than participants with more than 100 books. Participants without a valid response on the variable 'books at home' performed 0.12 logits (Cohen's d =0.152) or 0.806 logits (Cohen's d =0.972) worse than participants with up to 100 and more than 100 books, respectively.

There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.245). Comparing the group without valid responses to the two groups with valid responses, DIF occurs up to 0.54 logits. This is a rather large difference, which may, however, also be the result of the uncertainty in estimation due to the small number of persons with missing responses.

There were 10,009 (70.0%) participants without a migration background, 3,160 (22.1%) of the participants with a migration background (for 3.8% students neither their mother father or themselves were born in Germany, for 10.2% only the participants were born in Germany and both of their parents were born abroad, for the rest of the students only one of their parents was born abroad). 1,119 (7.8%) students could not be allocated to either group. The first three groups were used for investigating DIF of migration. There is a considerable difference in the average performance of participants with or without migration background (main effect = 0.574 logits, Cohen's d = 0.678). Participants without a migration background have a higher scientific literacy than participants with a migration background. Also students without a migration background differ from those with an unknown background on migration (main effect = 0.576 logits, Cohen's d = 0.683). However, there was no considerable difference between students with a migration background and those with an unknown background on migration (main effect = 0.006 logits, Cohen's d = 0.007).

DIF was also investigated for school type. 5,056 (35.4%) of the test takers were high-school students and 9238 (64.6%) were non high-school students. On average, high-school students have a higher scientific literacy score than students who do not attend a high school (main effect = 1.024 logits, Cohen's d = 1.380).

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 estimating 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 most DIF variables (position, books, migration background and country of origin). Only for the DIF variables gender and school type the more complex DIF model have slightly better information criterions.

Table 7: Differential item functioning (absolute differences between difficulties)

Item	Booklet	Gender	School type	Books			Migration status		
	Position 1 vs. Position 2	Male vs. female	high school vs. others	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing
SCG9011_c	0.052	-0.08	0.236	-0.128	-0.045	0.07	0.067	0.075	0.01
SCG9012s_c	0.016	0.05	-0.039	0.093	0.044	-0.037	0.019	0.103	0.086
SCG9051_c	-0.015	-0.17	0.181	-0.101	0.043	0.131	0.091	0.125	0.036
SCG9052s_c	-0.023	0.01	-0.053	0.059	-0.019	-0.066	0.006	0.038	0.033
SCG9092_c	-0.036	-0.05	0.273	-0.117	0.076	0.18	0.004	-0.065	-0.067
SCG9093_c	0.058	-0.17	0.207	-0.06	0.066	0.113	-0.001	0.035	0.038
SCG9611s_c	-0.117	0.04	0.063	-0.063	0.068	0.121	-0.062	-0.102	-0.038
SCG9612_c	0.099	-0.03	-0.020	-0.023	-0.059	-0.049	0.012	-0.019	-0.029
SCG9641_c	0.102	-0.08	0.111	-0.001	-0.128	-0.14	-0.155	-0.214	-0.058
SCG9642_c	-0.073	0.13	-0.203	0.076	0.003	-0.087	0.023	0.042	0.02
SCG9061s_c	-0.097	0.12	-0.136	0.068	0.151	0.07	0.05	-0.025	-0.073
SCG9063_c	0.124	0.37	-0.065	0.052	0.026	-0.04	-0.096	-0.072	0.025
SCG9081_c	0.037	-0.29	-0.331	0.177	0.01	-0.181	0.008	-0.026	-0.032
SCG9083s_c	-0.025	-0.38	-0.198	0.164	-0.011	-0.167	-0.098	-0.104	-0.005
SCG9103_c	0.007	-0.05	-0.013	-0.006	-0.025	-0.032	0.066	0.106	0.042
SCG9104_c	0.013	0.01	0.168	-0.039	-0.012	0.014	0.126	0.09	-0.035
SCG9105_c	0.019	-0.15	0.209	-0.084	-0.021	0.049	0.139	0.132	-0.006
SCG9042s_c	0.052	-0.01	-0.045	-0.009	0.032	0.031	0.001	0.025	0.026
SCG9043s_c	-0.006	0.15	-0.377	0.109	0.101	0.001	0.089	0.086	-0.001
SCG9651s_c	-0.006	0.10	0.070	-0.051	-0.033	0.008	-0.076	-0.105	-0.038
SCG9653_c	0.029	0.12	-0.029	0.005	-0.077	-0.094	0.017	0.118	0.102
SCG9032_c	-0.006	-0.06	-0.108	0.118	-0.139	-0.27	0.037	0.054	0.016
SCG9033_c	0.007	-0.03	-0.069	0.025	0.061	0.025	-0.004	-0.092	-0.087
SCG9621s_c	0.012	0.05	-0.141	0.072	0.069	-0.001	-0.11	-0.027	0.088
SCG9622_c	0.005	0.16	0.002	-0.019	-0.02	-0.014	0.027	0.032	0.005

Item	Booklet	Gender	School type	Books			Migration status		
	Position 1 vs. Position 2	Male vs. female	high school vs. others	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing
SCG9111_c	-0.08	0.06	0.135	-0.146	0.115	0.249	0.009	-0.039	-0.055
SCG9112_c	-0.004	-0.05	0.094	-0.043	-0.009	0.023	-0.039	0.014	0.056
SCG9113_c	-0.044	0.03	0.109	-0.158	0.043	0.19	-0.036	-0.046	-0.01

Table 8: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	556814.07	52	557311.59	556918.07
	DIF	556590.70	80	556923.11	556750.70
Gender	main effect	556771.34	52	556987.40	556875.34
	DIF	555222.82	80	555555.23	555382.82
Books	main effect	542344.60	52	542560.16	542448.60
	DIF	541811.04	80	542142.67	541971.04
Migration	main effect	430109.37	52	430319.79	430213.37
	DIF	429970.09	80	430293.80	430130.09
School type	main effect	552933.01	52	553149.08	553037.01
	DIF	551356.69	80	551689.10	551516.69

5.5.4 Rasch-homogeneity

In order to test for the assumption of Rasch-homogeneity all 28 items entered the analysis with the generalized partial credit model (2PL) to test for Rasch-homogeneity. The estimated discrimination parameters are depicted in the last column in table 4. They range from 0.40 (items SCG90320_c and SCG91120_c) to 2.4 (item SCG96410_c). The discriminations differ considerably among the items and the 2PL model (BIC = 555125.24, number of parameters = 92) fits the data slightly better than the partial credit model (1PL) (BIC = 557076.80, number of parameters = 51). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework, and thus the partial credit model was used to model the data and to estimate competence scores.

5.5.5 Unidimensionality of the test

The unidimensionality of the test was investigated by specifying a one- and a two-dimensional model.

The first model is based on the assumption that scientific literacy is a one-dimensional construct that measures one distinct competence whereas the second model distinguishes between the two sub-competencies knowledge about science (KAS) and knowledge of science (KOS) (for more details see Hahn et al., 2012). For estimating a two-dimensional model based on the Gauss Hermite quadrature estimation implemented in ConQuest (nodes were chosen in such a way that stable parameter estimation was obtained) was used. The two-dimensional model (BIC= 558228.91, number of parameters = 55) fits the data less well than the unidimensional model (BIC= 557076.80, number of parameters = 51; correlations of the two dimensions: 0.960). 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 in grade 9 and at describing how the scientific literacy score is estimated.

The amount of invalid responses and not-reached items is low. However, some items show higher omission rates, although, in general, the amount of omitted items is acceptable.

The test has a good reliability (WLE reliability = .777) and distinguishes well between test takers of average and low scientific literacy, but not as well for high performers. Very difficult items are missing; hence, test targeting is somewhat suboptimal and the test measures scientific literacy of high-performing students less accurately. This is depicted by the test’s variance (=0.780) which, ideally, should be higher.

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

A unidimensional partial credit model 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 has good psychometric properties that facilitate the estimation of a unidimensional scientific literacy score.

7. Data in the Scientific Use file

There are 28 items in the data set that are either scored as dichotomous variables (MC or SCR items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '0_c' at the end of the variable name, the CMC items are marked with a 's_c' at the end of the variable name. 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 is scored with 0.5 points. Manifest scale scores are provided in form of WLE estimates (sc_wle) including the respective standard error (sc_wle_se). Please note that when categories of the polytomous variables had less than N=200, the categories were collapsed. For the science test this concerned the two lowest categories of two polytomous items (see section 5.4.) on the aggregation of the CMC items. In the scaling model, the collapsed polytomous items are scored in steps of 0, 0.66, 1.33, and 2.0 (denoting the highest). Note than 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 from the items is provided in Appendix A. Students that did not take part in the test or those that do not have enough valid responses to estimate a scale score will have 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. User 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 indentification. *IEEE Transactions on Automatic Control*, 19, 716-722.

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.

Hahn, I. Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I.M., & Prenzel, M. (2012). Assessing scientific literacy over the lifespan - A description of the NEPS science framework and the test development. Manuscript submitted for publication.

Masters (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. (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.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C., & Hamann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In PISA-Konsortium Deutschland (Hrsg.), 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. v. Maurice & (Eds.). Education as a Lifelong Process: The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft, Sonderheft 14 .* 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, SCIENCE: Partial Credit Model;

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

labels << filename_with_labels.txt;

codes 0,1,2,3,4;

recode (0,1,2,3,4)      (0,0,1,2,3)             !item (20,21);
score (0,1)             (0,1)                   !item (1-19,23);
score (0,1,2,3,4)       (0,0.5,1,1.5,2)         !item (22,24-28);
score (0,1,2,3)         (0,0.66,1.33,2)         !item (20,21);

set constraint=cases;

model item + item*step + position;
estimate;

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