NEPS SURVEY PAPERS

Linda Funke, Jana Kähler, and Inga Hahn

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 3 IN 6TH GRADE

NEPS Survey Paper No. 5
Bamberg, July 2016

LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

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

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

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

NEPS Technical Report for Science:

Scaling Results of Starting Cohort 3 in 6th Grade

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

E-mail address of the lead author:

funke@ipn.uni-kiel.de

Bibliographic data:

Funke, L., Kähler, J., & Hahn, I. (2016): NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 in 6th Grade (NEPS Survey Paper No. 5). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 in 6th 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 3 in grade 6. 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 6 originally consisted of 27 multiple choice and complex multiple choice items and covered two knowledge domains as well as three different contexts. The test was administered to 4,871 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. Two items had to be eliminated due to insufficient item discrimination. The results of the remaining 25 items 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, 6th 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 3.

The present report has been modeled 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. We do, however, not expect severe changes in the 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 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 et al., 2013, 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 differed in the order of the tests they received. Some subjects received 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 six consisted of 27 items. Two items had to be eliminated from the test due to insufficient item discrimination. The characteristics of the remaining 25 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., 2013).

Table 1: Classification of the science test items for grade 6

Knowledge domains	Frequency
Knowledge of Science (KOS)	18
Knowledge about Science (KAS)	9
Total number of items	27

Table 2: Response formats of the science test items for grade 6

Response format	Frequency
Simple Multiple-Choice	17
Complex Multiple-Choice (True false items)	10
Total number of items	27

Table 3: Number of items for the different contexts of the science test for grade 6

Context	Frequency
Health	8
Environment	8
Technology	11
Total number of items	27

3.2 Sample

Overall, 5,465 students are part of the sample. 4,871 of these students took the science test. There were two testing groups which differed in the order of the tests they received. 2,448 persons received the science test before the ICT test while 2,423 persons received the science test after completing the ICT test.

All 4,871 persons who took part in the science test are included in the descriptive analyses. The results are presented in the following sections.

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 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, in order to avoid possible estimation problems, the categories were collapsed. 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.

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). The technical report includes two WLE-estimates, one WLE (scg6_sc1) correcting for the rotation (Science-ICT / ICT-Science) which should be used for cross-sectional analyses and one WLE (scg6_sc1u) not correcting for the rotation which should be used for the longitudinal analyses.

Person parameter estimation in NEPS is described in Pohl and Carstensen (2012a) while the data available in the SUF are described in section 7. The item parameters were plotted to the ability estimates of the persons 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 6 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 visual inspection of the item characteristic curves. Only if the subtasks had 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 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

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

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 correlation of the correct responses with the total score, and the item characteristic curve were evaluated for each item. Items with a WMNSQ > 1.15 (t-value > |6|) were considered having a noticeable item misfit and items with a WMNSQ > 1.2 (t-value > |8|) were judged as a considerable item misfit and their performance was 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 worth further investigation, 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., 2013). 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 5,465 persons. In an initial step, for calculating item parameters, all persons who took part in the test were included (n=4,871). For further analyses, only persons with more than two valid responses were taken into account. In this case, all of the persons who took the test had more than two valid responses, so the results presented in the following sections are based on all 4,871 persons.

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, we evaluated the relative frequency of the responses given before performing IRT-analyses. The percentage of persons correctly responding to an item (relative to all valid responses) varies over items from 41.1% to 86.3% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varies from 23.9% to 74.9%. 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 or medium 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 79.8 % of the persons all answers were valid.

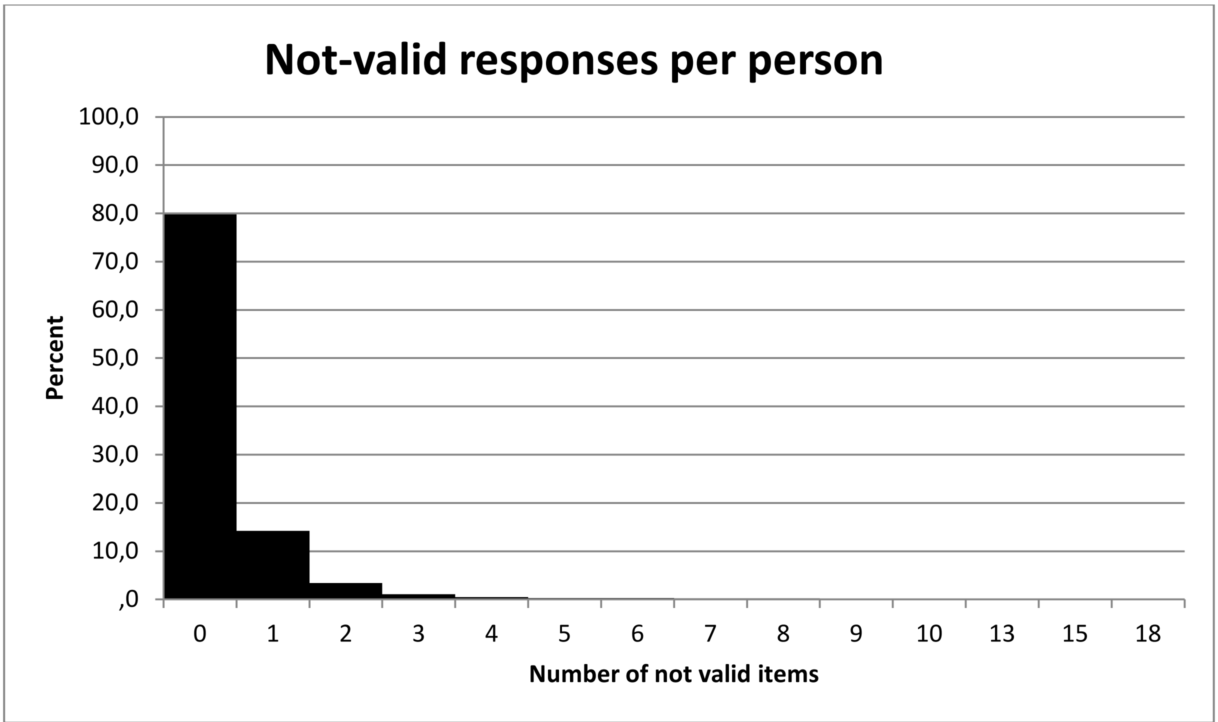

Figure 1: Number of non-valid responses

The number of omitted responses per person is depicted in Figure 2. 80.2 percent of the persons did not omit a single item. Only 4.8% omitted 3 or more than 3 items.

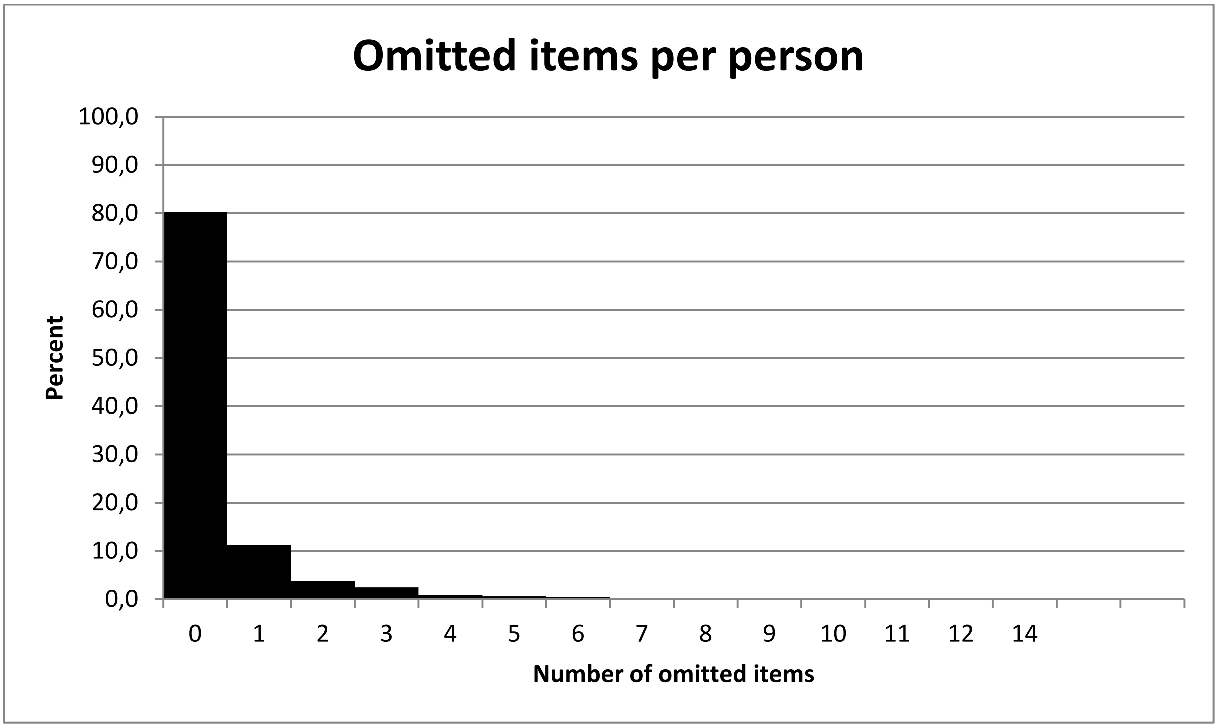

Figure 2: Number of omitted items

Most students reached the end of the test (88.6%) and only a small proportion did not manage to finish at least two thirds of the test.

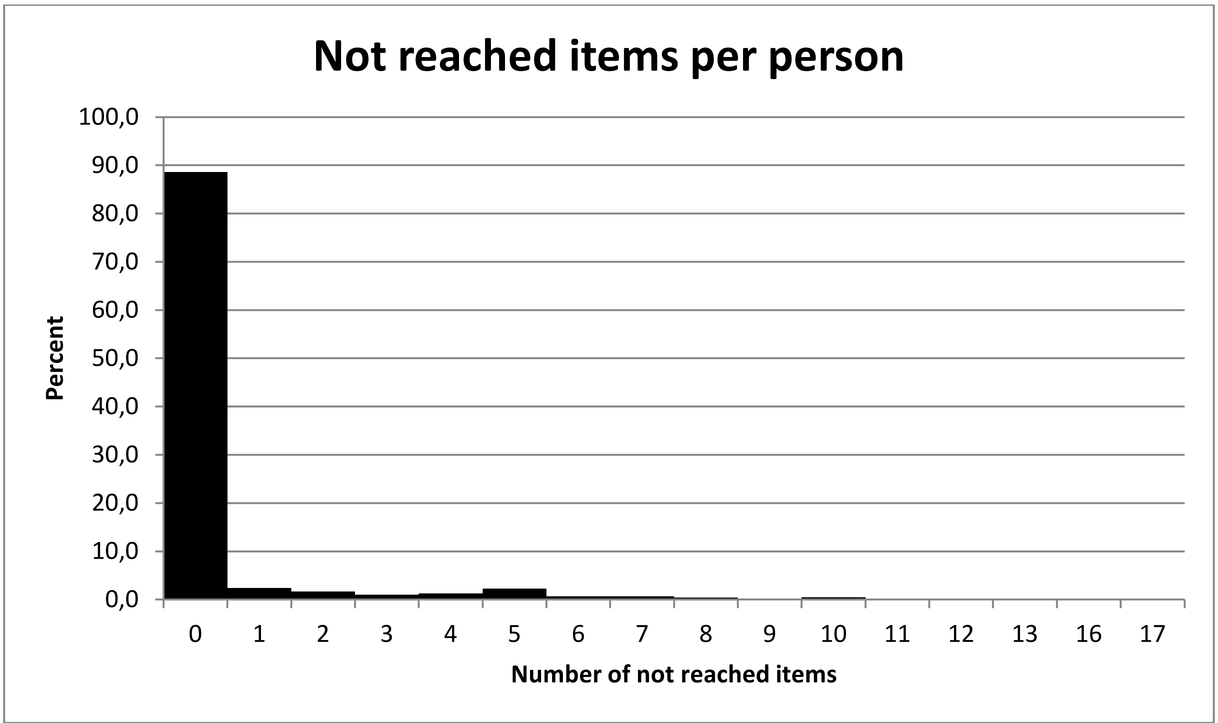

Figure 3: Number of not reached items

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. 59.5% of the students answered all questions and consequently had no missing responses. Only 0.5% 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.

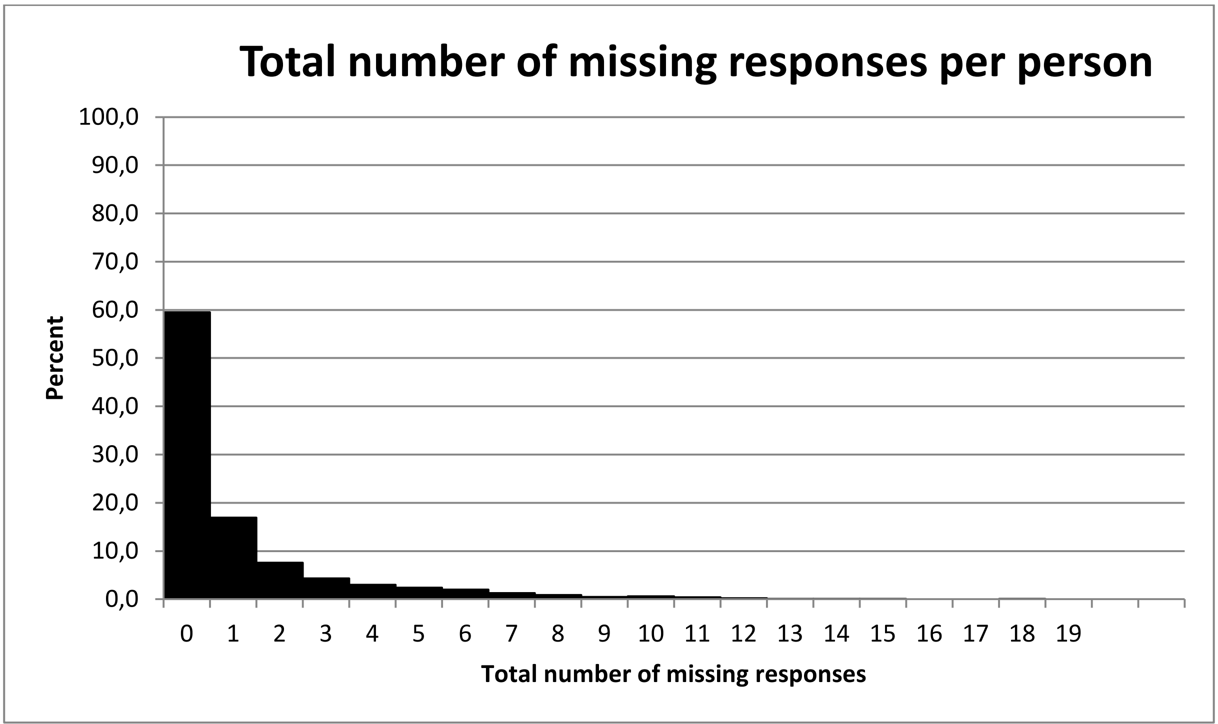

Figure 4: Total number of missing responses

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 4.1%. The number of missing responses is correlated to .49 with the difficulty of the item ($p < .05$). Since the correlation is significant, 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 5.5% for item scg6103s_c. The relative frequency of not reached items increases towards the end of the test. Eventually 11.4% 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 0.5% and 12.9%.

Table 4: Valid Responses and Missing Values

Item No.	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 %
1	scg6103s_c	4,564	1	0.0	0.8	5.5
2	scg61050_c	4,693	2	0.0	2.1	1.6
3	scg60120_c	4,845	3	0.0	0.4	0.1
4	scg60410_c	4,826	4	0.0	0.4	0.5
5	scg60430_c	4,815	5	0.0	0.7	0.4
6	scg66310_c	4,783	6	0.0	1.4	0.4
7	scg66320_c	4,694	7	0.0	3.4	0.2
8	scg66340_c	4,731	8	0.0	2.1	0.7
9	scg61410_c	4,818	9	0.0	0.9	0.1
10	scg6142s_c	4,557	10	0.1	3.0	3.4
11	scg61430_c	4,683	11	0.1	3.5	0.3
12	scg6144s_c	4,586	12	0.1	2.1	3.7
13	scg60510_c	4,819	13	0.1	0.6	0.3
14	scg60530_c	4,785	14	0.2	0.8	0.8
15	scg6661s_c	4,753	15	0.3	0.5	1.7
16	scg66620_c	4,628	16	0.7	4.1	0.2
17	scg66630_c	4,724	17	0.9	1.9	0.2
18	scg6664s_c	4,565	18	1.3	2.1	2.9
19	scg6111s_c	4,709	19	2.0	0.4	0.9
20	scg6113s_c	4,471	20	2.7	1.5	4.0
21	scg66040_c	4,519	21	5.0	1.9	0.3
22	scg61310_c	4,464	22	6.3	1.6	0.5
23	scg61330_c	4,357	23	7.3	2.7	0.5
24	scg6061s_c	4,245	24	9.0	0.3	3.6
25	scg60620_c	4,272	25	11.4	0.0	0.9

Table 5: Item parameters

Item No.	Item	Difficulty/location parameter	SE (difficulty/ location parameter)	Weighted MNSQ	Weight t-value	Pt.bis of correct response	Discrimination (2PL)
1	scg6103s_c	0.103	0.033	1.04	3.2	0.44	0.89
2	scg61050_c	-0.736	0.034	1.00	-0.3	0.47	1.04
3	scg60120_c	-1.065	0.035	1.04	2.4	0.41	0.86
4	scg60410_c	-0.945	0.034	1.03	1.6	0.43	0.92
5	scg60430_c	-1.237	0.036	1.07	3.8	0.37	0.79
6	scg66310_c	-0.140	0.032	1.07	6.0	0.39	0.73
7	scg66320_c	-1.059	0.035	0.95	-3.0	0.50	1.28
8	scg66340_c	-0.780	0.034	1.01	0.7	0.46	1.00
9	scg61410_c	-1.923	0.041	0.95	-2.2	0.46	1.50
10	scg6142s_c	0.069	0.033	1.04	3.2	0.42	0.83
11	scg61430_c	-1.135	0.036	0.92	-4.8	0.53	1.47
12	scg6144s_c	-0.299	0.033	1.05	4.1	0.42	0.81
13	scg60510_c	-2.185	0.045	1.00	-0.1	0.35	0.98
14	scg60530_c	-1.989	0.042	0.97	-1.0	0.41	1.21
15	scg6661s_c	-0.721	0.034	0.93	-3.9	0.40	0.73
16	scg66620_c	-0.906	0.035	1.05	3.1	0.41	0.87
17	scg66630_c	-0.710	0.033	0.94	-4.0	0.52	1.27
18	scg6664s_c	-0.190	0.033	1.01	0.7	0.47	1.00
19	scg6111s_c	-1.311	0.037	1.06	3.1	0.37	0.76
20	scg6113s_c	-1.448	0.043	0.92	-4.4	0.46	0.77
21	scg66040_c	-1.377	0.038	0.94	-3.2	0.51	1.48
22	scg61310_c	-0.268	0.033	1.11	8.2	0.36	0.65
23	scg61330_c	0.460	0.034	0.99	-0.6	0.45	0.98
24	scg6061s_c	-0.283	0.034	1.00	0.0	0.47	1.00
25	scg60620_c	-0.931	0.036	0.96	-2.4	0.50	1.18

5.4 Parameter estimates

5.4.1 Item parameters

In the end, 25 of the original 27 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 the two remaining polytomous variables) are depicted in Table 6. For item scg6661s_c the two lowest categories were collapsed and for item scg6113s_c the three lowest categories were collapsed. As these items were CMC items with a maximum score of 2, they were scaled using the following intervals: 0, 0.5, 1 and 1.5 (for scg6661s_c) and 0, 0.5 and 1 (for scg6113s_c). The CMC items scg6103s_c, scg6142s_c, scg6144s_c, scg6664s_c, scg6111s_c and scg6061s_c were reduced to a 0 and 1 scoring since they showed a decrease in one or two of their step parameters instead of an increase.

Table 6: Step parameters for the CMC items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)
scg6661s_c	-1.001 (0.031)	-0.008 (0.030)	1.009
scg6113s_c	-0.011 (0.034)	0.011	

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 -2.19 (scg60510_c) and 0.46 (scg61330_c) with a mean of -0.84. Due to the large sample size, the standard error of the estimated item difficulties is very small, *SE*(ß) ≤ 0.05. Overall, the item difficulties are low and the test lacks items with a high difficulty (above 2 logits).

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 ability score is constrained to zero. The variance was estimated to be 1.082, indicating that the test has good potential to differentiate between subjects. The reliability of the test (WLE reliability = .771) is good. The amount to which the item difficulties and location parameters are targeted to the ability of the persons is shown in Figure 5. The Figure shows that the items cover a great range of the ability distribution of the persons. However, only few items cover medium person abilities and there are no items available for persons with high science abilities. 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 29.3 cases
3		
	X	
	XX	
2	XX	
	X	
	XX	
	XXX	
	XXXX	
	XXXX	
	XXXX	
1	XXXXXX	
	XXXXXX	
	XXXXXX	
	XXXXXXX	
	XXXXXXXXX	23
	XXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXX	1 10
0	XXXXXXX	
	XXXXXXX	6 18 22
	XXXXXXXX	12 24
	XXXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXX	2 8 15 17
	XXXXXXX	4 16 25
-1	XXXX	3 7
	XXXXX	5 11
	XXXXX	19 21
	XXX	20
	XXX	
	XX	
	XX	9
-2	X	14
	X	13
	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 29.3 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item.

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 whole 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 items in a Rasch model.

No estimation problems occurred and all subtasks showed a satisfactory item fit. The WMNSQ ranged from 0.92 to 1.11, the respective t-value from -4.8 to 8.2. 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. For two items one or more distractors showed a positive or almost zero point-biserial correlation. These items were removed from further analyses (e.g. computation of WLE-scores, differential item functioning). The distractors of all the other items had a point-biserial correlation with the total score below zero. The results indicate that these 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 0.92 (item scg61430_c and scg6113s_c) and the highest being 1.11 (item scg61310_c). Overall, there are no items with a WMNSQ above 1.2. However, there was one item with a *t*-value above 8 (item scg61310_c) but the item characteristic curve of this item showed a reasonable or good fit. Hence, no indications for a heavy misfit of the item could be detected and therefore, it was 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). 2,423 students received the mathematics test first, then the computer literacy test, and at last the science test (position 2) while 2,448 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 specific 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 only amounts to 0.128 logits.

DIF was also investigated for gender. 2,363 (48.5%) of the test takers were female and 2,504 (51.4%) were male. (Four test takers did not specify their gender.) On average, female students show slightly lower scores in scientific literacy than male students (main effect = 0.218 logits, Cohen's *d* = -0.211). There is no item with a considerable gender DIF. The highest difference in difficulties between the two groups is 0.273 logits.

The number of books at home was used as a proxy for socio-economic status. There were 1,764 (36.2%) test takers with 0 to 100 books at home, 2,431 (49.9%) test takers with more than 100 books at home, and 676 (13.9%) 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 on average show a 0.686 logits (Cohen's *d* = -0.718) lower scientific literacy score than participants with more than 100 books. Participants without a valid response on the variable 'books at home' performed 0.150 logits (Cohen's *d* = 0.154) lower than participants with up to 100 and 0.836 logits (Cohen's *d* = 0.829) lower than participants with more than 100 books at home, respectively. There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.179). Comparing the group without valid responses to the two groups with valid responses, DIF occurs up to -0.271 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 3,160 (64.9%) participants without a migration background, 894 (18.4%) of the participants with a migration background (for 1.9% students neither their mother, father nor they, themselves, were born in Germany, for 7.3% only the participants were born in Germany and both of their parents were born abroad, for 8.0% of the students only one of their parents was born abroad, 0.8% of the children were born abroad while their parents had no migration background and in 0.4% of the cases the child and only one parent had migration background). 817 (16.8%) students could not be allocated to either group. All children who could either be clearly allocated to the group with or to the group without migration background 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 = 0.566 logits, Cohen's d = 0.574). 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 = 0.666 logits, Cohen's *d* = 0.662). However, there was no considerable difference between students with a migration background and those with an unknown background on migration (main effect = 0.100 logits, Cohen's *d* = 0.098).

DIF was also investigated for school type. 2,253 (46.3%) of the test takers were high-school students and 2,618 (53.7%) 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 = 0.974 logits, Cohen's *d* = -1.069).

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
scg6103s_c	0.007	-0.019	-0.026	-0.019	0.004	0.018	0.032	0.062	0.032
scg61050_c	-0.070	0.018	0.012	0.112	0.081	-0.036	-0.080	-0.057	0.025
scg60120_c	-0.100	0.020	-0.020	-0.107	0.091	0.194	0.154	0.156	0.002
scg60410_c	-0.025	0.092	-0.076	-0.113	-0.068	0.041	-0.058	-0.014	0.046
scg60430_c	0.014	0.170	-0.172	-0.122	-0.065	0.054	0.041	0.030	-0.010
scg66310_c	-0.080	-0.034	-0.123	-0.027	-0.007	0.016	0.066	-0.009	-0.074
scg66320_c	-0.062	-0.034	0.136	0.077	-0.028	-0.110	-0.090	-0.100	-0.009
scg66340_c	-0.036	0.273	-0.058	0.003	0.003	-0.004	0.044	0.033	-0.010
scg61410_c	0.040	0.231	0.113	0.037	-0.061	-0.102	-0.074	-0.125	-0.050
scg6142s_c	0.025	-0.040	0.052	-0.068	-0.003	0.060	0.059	0.000	-0.058
scg61430_c	0.128	0.011	0.060	0.077	0.049	-0.032	-0.091	-0.006	0.087
scg6144s_c	-0.038	0.261	-0.232	-0.119	-0.043	0.072	-0.033	0.019	0.052
scg60510_c	0.075	-0.151	-0.137	-0.102	-0.098	0.001	0.122	0.075	-0.046
scg60530_c	0.053	0.020	-0.016	-0.058	-0.099	-0.046	0.069	-0.117	-0.185
scg6661s_c	-0.041	-0.078	0.357	0.179	-0.017	-0.179	-0.060	-0.141	-0.082
scg66620_c	0.013	-0.093	-0.099	-0.145	-0.049	0.092	0.006	0.103	0.099
scg66630_c	0.043	-0.133	0.155	0.106	-0.005	-0.116	-0.030	-0.061	-0.030
scg6664s_c	0.051	-0.246	0.037	0.002	0.041	0.035	-0.022	0.026	0.050
scg6111s_c	0.049	-0.156	-0.186	-0.101	0.130	0.228	0.101	0.205	0.105
scg6113s_c	0.072	-0.011	0.147	0.172	-0.099	-0.271	-0.314	-0.210	0.104
scg66040_c	-0.007	-0.169	0.221	0.099	-0.012	-0.115	0.008	-0.079	-0.086
scg61310_c	-0.038	0.101	-0.094	0.013	0.067	0.050	0.106	0.088	-0.018
scg61330_c	0.020	-0.037	-0.030	-0.034	0.074	0.105	0.042	0.122	0.081
scg6061s_c	0.068	-0.162	0.004	0.033	0.004	-0.034	0.055	-0.013	-0.067
scg60620_c	-0.055	0.170	0.070	0.158	0.014	-0.149	-0.170	-0.161	0.011

Table 8: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	139575.45	30	139635.45	139830.18
	DIF	139513.36	55	139623.36	139980.37
Gender	main effect	139443.80	30	139503.80	139698.50
	DIF	139065.32	55	139175.32	139532.29
Books	main effect	139083.52	31	139145.52	139346.74
	DIF	138873.75	81	139035.75	139561.53
Migration	main effect	139273.28	31	139335.28	139536.50
	DIF	139094.18	81	139256.18	139781.96
School type	main effect	138646.63	30	138706.63	138901.36
	DIF	138315.41	55	138425.41	138782.42

5.5.4 Rasch-homogeneity

In order to test for the assumption of Rasch-homogeneity all 25 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 5. They range from 0.65 (item scg61310_c) to 1.50 (item scg61410_c). The discriminations do not differ much among the items. However, the 2PL model (BIC = 139435.57, number of parameters = 64) fits the data better than the 1PL model (1PL, BIC = 139602.75, number of parameters = 29). This result fits the theoretical aim of constructing a test that equally represents the different aspects of the framework.

5.5.5 Unidimensionality of the test

The unidimensionality of the test was investigated by specifying an onedimensional and a twodimensional model.

The first model is based on the assumption that scientific literacy is a onedimensional construct that measures one distinct competence whereas the second model distinguishes between the two sub-competencies, knowledge about science and knowledge of science (for more details see Hahn et al., 2013). For estimating a twodimensional 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 twodimensional model (BIC= 139559.69, number of parameters = 31) slightly fits the data better than the unidimensional model (BIC= 139602.75, number of parameters = 29). However, the two subdimensions show a high correlation, r = .957, and thus, scientific literacy as measured by this test is regarded as unidimensional.

6. Discussion

The analyses in the previous sections aimed at providing information on the quality of the science test in grade 6 and at describing how the scientific literacy score is 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 is acceptable.

The test had a good reliability (WLE reliability = .771) and distinguished well between test takers of average and low scientific literacy, but not as well for high performers. Very difficult items were missing; hence, test targeting was somewhat suboptimal and the test measured scientific literacy of high-performing students less accurately.

Indicated by various fit criteria – WMNSQ, t-value of the WMNSQ, ICC – the items exhibited 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) were good. Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test was fair to the considered subgroups.

A twodimensional partial credit model yielded a better model fit than a unidimensional partial credit model (between-item-multidimensionality, the dimensions being the content areas). Due to the high correlations between the subdimensions, 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 25 items in the data set that are either scored as dichotomous variables (MC 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 (scg6_sc1u) including the respective standard error (scg6_sc2u). 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 and three 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, 0.5, 1.0 and 1.5 or 0, 0.5 and 1 (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 Appendices A and B. 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. (2013). Assessing scientific literacy over the lifespan - A description of the NEPS science framework and the test development. *Journal for Educational Research Online, 5*(2),110-138.

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. (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 (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

<u>Appendix A: ConQuest-Syntax for estimating WLE estimates in starting cohort 3 (controlled for the "rotation"-variable"</u>

```
Title Starting Cohort 3, SCIENCE: Partial Credit Model;

data filename.dat;
format pid 1-7 responses * /* insert number of columns with data*/

labels << filename_with_labels.txt;

codes 0,1,2,3;

score (0,1)         (0,1)             !item (1-14,16-19,21-25);
score (0,1,2,3)     (0,0.5,1,1.5)     !item (15);
score (0,1,2)       (0,0.5,1)         !item (20);


set constraint=cases;

model item + item*step + rotation;
estimate;

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

Appendix B: ConQuest-Syntax for estimating WLE estimates in starting cohort 3 (not controlled for the "rotation"-variable")

```
Title Starting Cohort 3, SCIENCE: Partial Credit Model;

data filename.dat;
format pid 1-7 responses * /* insert number of columns with data*/

labels << filename_with_labels.txt;

codes 0,1,2,3;

score (0,1)          (0,1)            !item (1-14,16-19,21-25);
score (0,1,2,3)      (0,0.5,1,1.5)    !item (15);
score (0,1,2)        (0,0.5,1)        !item (20);

set constraint=cases;

model item + item*step;
estimate;

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