NEPS *SURVEY PAPERS*

Jana Kähler

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 7

NEPS *Survey Paper* No. 78
Bamberg, November 2020

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).

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 edited by a review board consisting of the scientific management of LIfBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section "Publications") and at www.lifbi.de/publications.

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 2 for Grade 7

Jana Kähler
Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany

Email address of the lead author:

jkaehler@leibniz-ipn.de

Bibliographic data:

Kähler, J. (2020). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 2 for Grade 7* (NEPS *Survey Paper* No. 78). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP78:1.0

NEPS Technical Report for Science: Scaling Results of Starting Cohort 2 for Grade 7

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span and develops tests for the assessment of different competence domains. To evaluate the quality of these competence tests various analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the scientific literacy test that was administered in Grade 7 of starting cohort 2. The scientific literacy test contained 26 items with different response formats representing different contexts as well as different areas of knowledge. The test was administered to 2,969 students. Their responses were scaled using a partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited a good reliability and that all items satisfactorily fitted the model. Furthermore, test fairness could be confirmed for different subgroups. As the correlations between the two knowledge domains were very high, the assumption of unidimensionality seems adequate. A limitation of the test was the lack of very difficult items. However, the results revealed good psychometric properties of the scientific literacy test, thus, supporting the estimation of a reliable scientific literacy score. Besides the scaling results, this paper also describes the data available in the scientific use file and provides the ConQuest syntax for scaling the data.

Keywords: scientific literacy, 7th Grade, differential item functioning, item response theory, scaling, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the lifespan (Blossfeld, Roßbach, & Maurice, 2011). These include, among others, reading competence, mathematical competence, scientific literacy, information and communication literacy, metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert et al. (2011) and by Fuß, Gnambs, Lockl, and Attig (2019).

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

In this paper, the results of these analyses are presented for a scientific literacy test that was administered in Grade 7 of starting cohort 2. First, the main concepts of the scientific literacy test are introduced. Then, the scientific literacy data of starting cohort 2 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is presented.

Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the SUF may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamental changes in the presented results.

2 Testing Scientific Literacy

The framework and test development for the scientific literacy test are described by Weinert et al. (2011) and by Hahn et al. (2013). In the following, we point out specific aspects of the scientific literacy test that are necessary for understanding the scaling results presented in this paper.

Scientific literacy is conceptualized as a unidimensional construct comprising two sub-dimensions. These are a) the knowledge of science (KOS) and b) the knowledge about science (KAS). KOS is specified as the knowledge of basic scientific concepts and facts whereas KAS can be regarded as the understanding of scientific processes.

KOS is divided into the content-related components of matter, system, development, and interaction. KAS is divided into the process-related components of scientific enquiry and scientific reasoning. KAS and KOS are implemented in three contexts: health, environment, and technology (see Figure 1). The test items are organized as single items or as units (testlets). One unit consists of two items. Each item or unit refers to one context-component-combination.

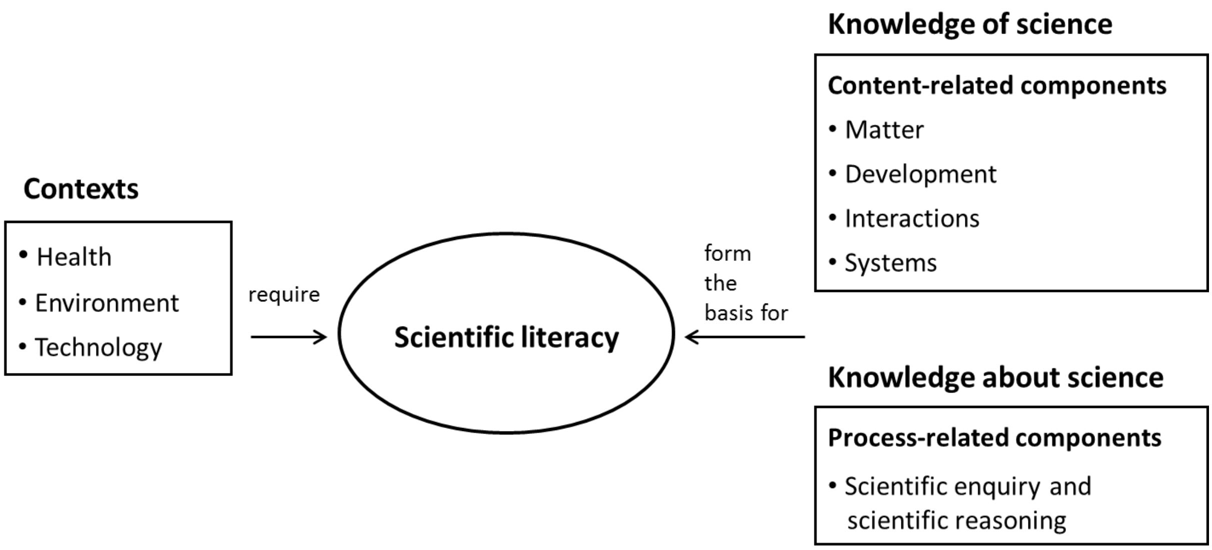

Figure 1. Assessment framework for scientific literacy (Hahn et al., 2013).

In the scientific literacy test for Grade 7 of starting cohort 2 (Kindergarten), there were two types of response formats. These were simple multiple-choice (MC) and complex multiple-choice (CMC) in the special form of true-false items. In MC items the test taker had to identify the correct answer out of four response options. The three incorrect response options functioned as distractors. In CMC items four subtasks with two response options each (e.g., yes/ no) were presented.

3 Data

3.1 The design of the study

The study assessed different competence domains including scientific literacy, reading, and mathematical competences. The scientific literacy test was administered either before or after the other competence test (reading and math). Therefore, one testing group first completed the science test, followed by the mathematics test or the reading test, while the other group completed the tests in the opposite order. Note that there was no multi-matrix design regarding the choice and the order of the items within a specific test. All children received the same science items in the same order. The testing time for the scientific literacy test was 29 minutes.

The allocation of the 26 items to the content areas (KOS and KAS) is summarized in Table 1. Table 2 shows how the items cover the different contexts of the scientific literacy framework (Hahn et al., 2013), whereas Table 3 gives an overview of the response formats.

Table 1: Classification of Items into Knowledge Domains

Knowledge domains	Number of Items
Knowledge of Science (KOS)	18
Knowledge about Science (KAS)	8
Total number of items	26

Table 2: Number of Items by Different Contexts

Context	Number of Items
Health	6
Environment	9
Technology	11
Total number of items	26

Table 3: Number of Items by Response Formats

Response format	Number of Items
Simple Multiple-Choice	16
Complex Multiple-Choice (True-false items)	10
Total number of items	26

3.2 Sample

A total of 2,969 individuals received the scientific literacy test. For two participants less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, these cases were excluded from further analyses (Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 2,967 individuals (50.7% girls). A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

4 Analyses

4.1 Missing responses

There are different kinds of missing responses. These are a) invalid responses, b) omitted items, c) items that test-takers did not reach, d) items that have not been administered, and

e) multiple kinds of missing responses within CMC items that are not determined. In this study, all subjects received the same set of items so there are no missing responses due to items not being administered.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response, or when less than four answers were given in a CMC item (which consists of four subtasks). Omitted items occurred when test-takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response were coded as not-reached. As CMC 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 was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats) and need to be accounted for in the estimation of item and person parameters. We, therefore, thoroughly investigated the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This indicated how well the persons were coping with the test. We then looked at the occurrence of missing responses per item to obtain some information on how well the items worked.

4.2 Scaling model

To estimate item and person parameters for scientific literacy, a partial credit model was used (PCM; Masters, 1982) that estimates item difficulties for dichotomous variables and location parameters for polytomous variables. Ability estimates for scientific literacy were estimated as weighted maximum likelihood estimates (WLEs). Item and person parameter estimation in NEPS is described in Pohl and Carstensen (2012), whereas the data available in the SUF are described in Section 7.

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item was scored as missing. Categories of polytomous variables with less than N = 200 responses were collapsed to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; especially when the item consisted of many subtasks. In these cases, the lower categories were collapsed into one category. For all of the four CMC items categories were collapsed (see Appendix A). 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 as 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

4.3 Checking the quality of the test

The scientific literacy test was specifically constructed to be implemented in the NEPS. To ensure appropriate psychometric properties, the quality of the test was evaluated in several pretests and analyses.

Before aggregating the subtasks of CMC items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1980). The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC variables that were included in the final scaling model.

The MC items consisted of one correct response and one or more distractors (i.e., incorrect response options). The quality of the distractors within MC items was examined using the point-biserial correlation between an incorrect response and the total score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the corrected discrimination as computed in ConQuest) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. The overall judgment of the fit of an item was based on all fit indicators.

Scientific literacy should measure the same construct for all children. If any items favored certain subgroups (e.g., if they were easier for boys than for girls), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., boys and girls) would be biased and thus unfair. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) analyses were estimated using a multigroup IRT model, in which the main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy of further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The scientific literacy test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The science test was constructed to measure a unidimensional scientific literacy score (Hahn et al., 2013). The assumption of unidimensionality was, nevertheless, tested by specifying a two-dimensional model with process-related items (KAS) representing one and content related items (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 test.

Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's Q3 (Yen, 1984). Because in the case of locally independent items, the Q3 statistic tends to be slightly negative, we report the corrected Q3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q3 falling below .20 indicate that the assumption of local item dependence (LID) is essentially met.

4.4 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

5 Results

All 26 items (including all subtasks for the polytomous items) were included in the analyses.

5.1 Descriptive statistics of the responses

To a) get a first rough descriptive measure of the item difficulties and b) check for possible estimation problems, before performing IRT analyses we evaluated the relative frequency of the responses given. The percentage of persons correctly responding to an item (relative to all valid responses) ranged from 13.2% to 76.3% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 25.3% and 64.9%.

5.2 Missing Responses

5.2.1 Missing responses per person

Figure 2 shows the number of invalid responses per person. Overall, there were very few invalid responses. 96.8% of the respondents did not have any invalid response at all; overall, about 0.8% of the respondents had more than one invalid response.

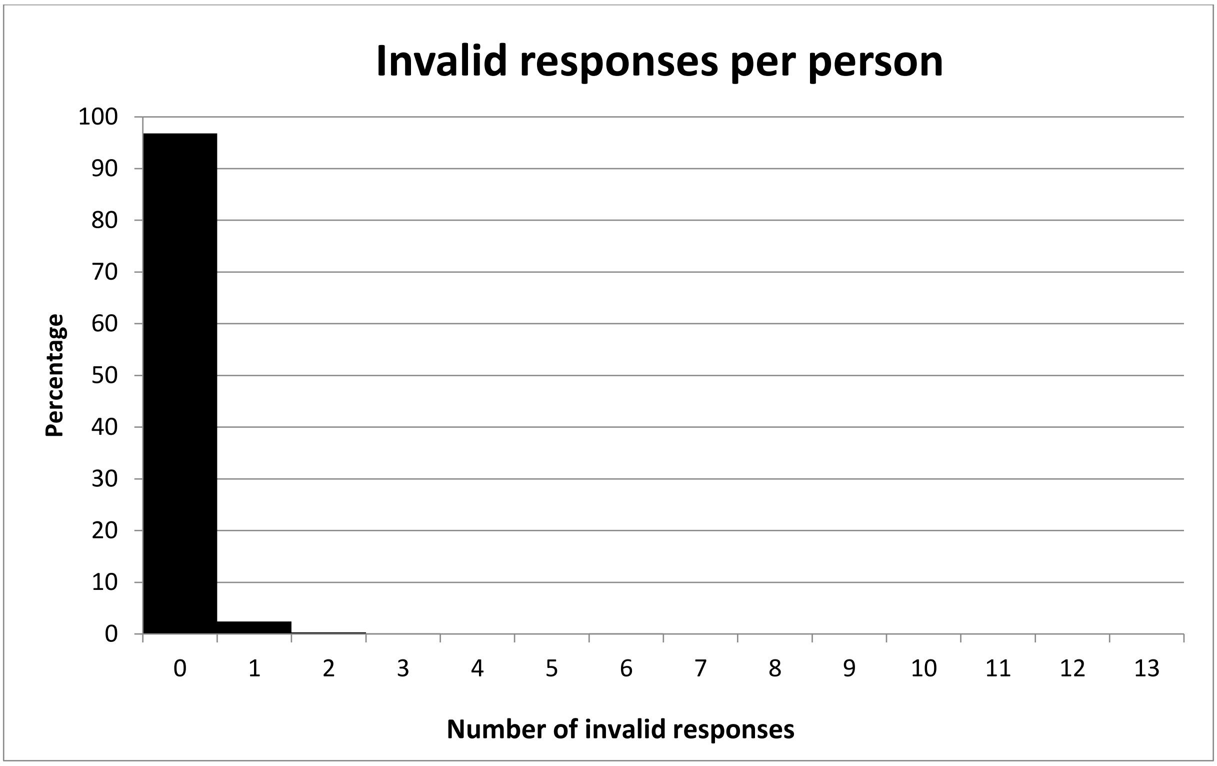

Figure 2. Number of invalid responses per person.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 65.0%, did not skip any item, and less than 6.9% omitted more than three items.

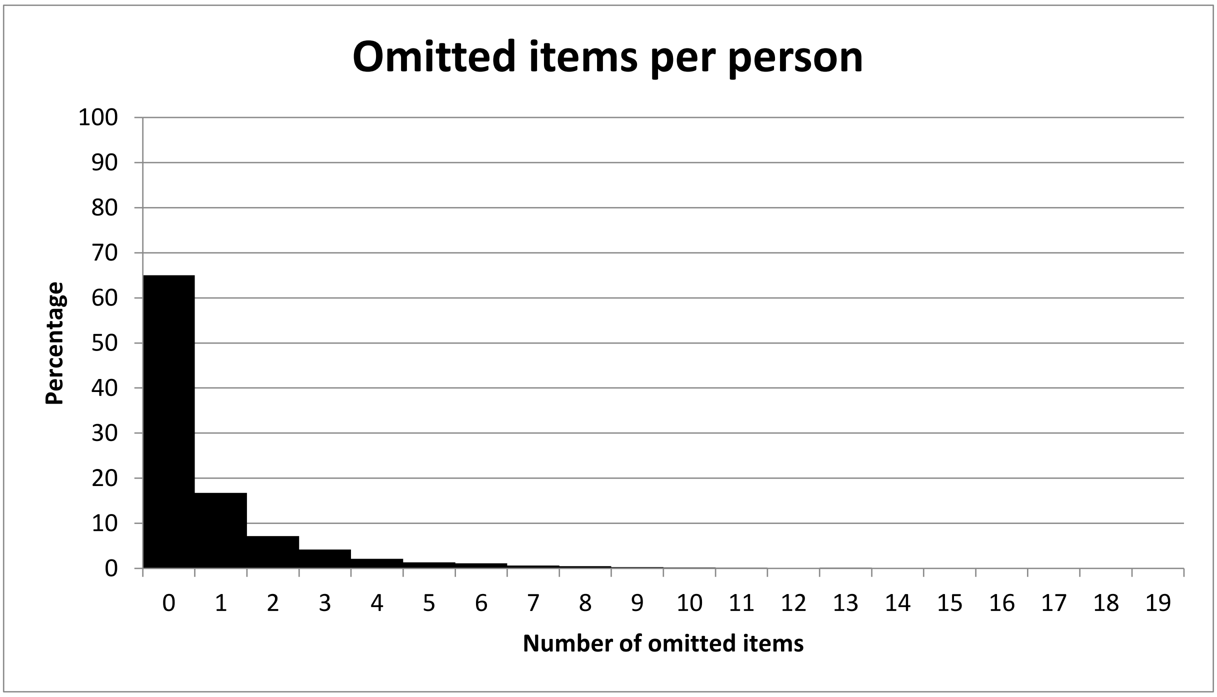

Figure 3. Number of omitted responses per person.

Another source of missing responses are items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was medium, about 49.1% of the respondents were able to finish the test within the allocated time limit (Figure 4). About 1.4% did not finish more than half of the items.

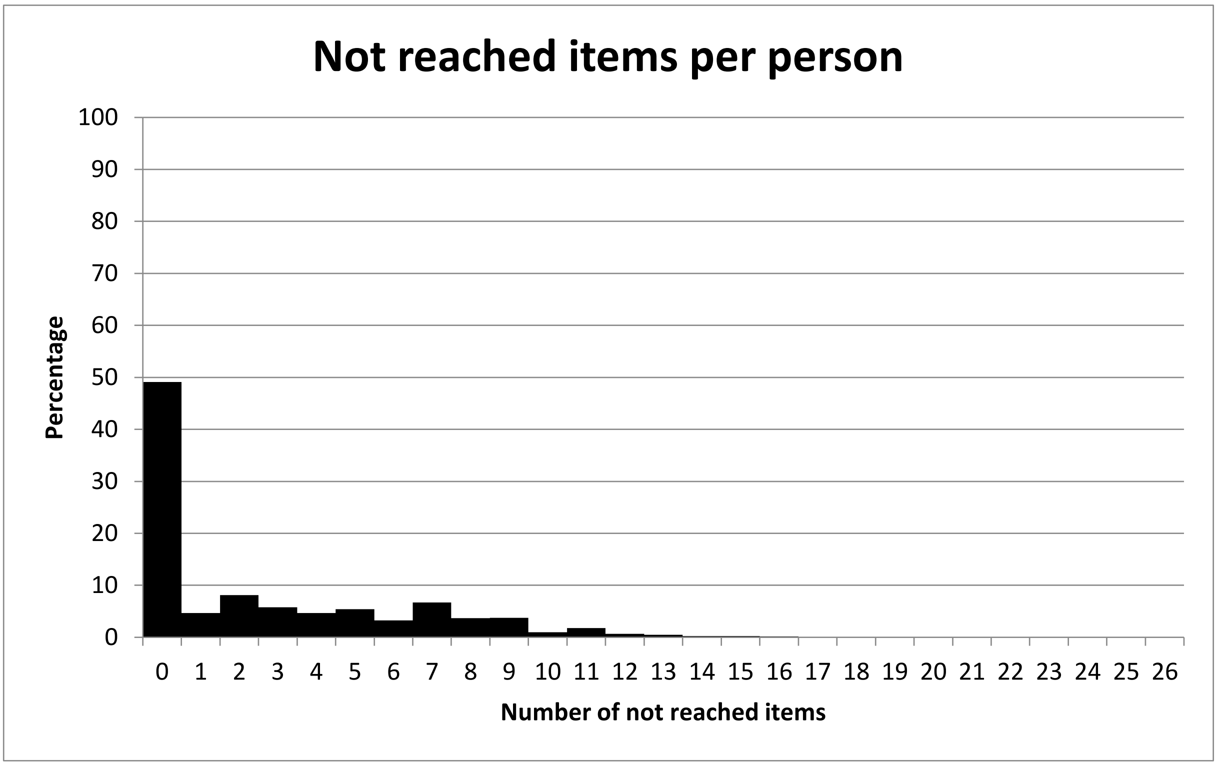

Figure 4. Number of not reached items per person.

The total number of missing responses, aggregated over invalid, omitted and not-reached missing responses, is illustrated in Figure 5. 33.4% of the students answered all questions and, consequently, had no missing responses. Only 3.4% of the students had missing responses on more than half of the items. Hence, the number of missing responses per person can be classified as small.

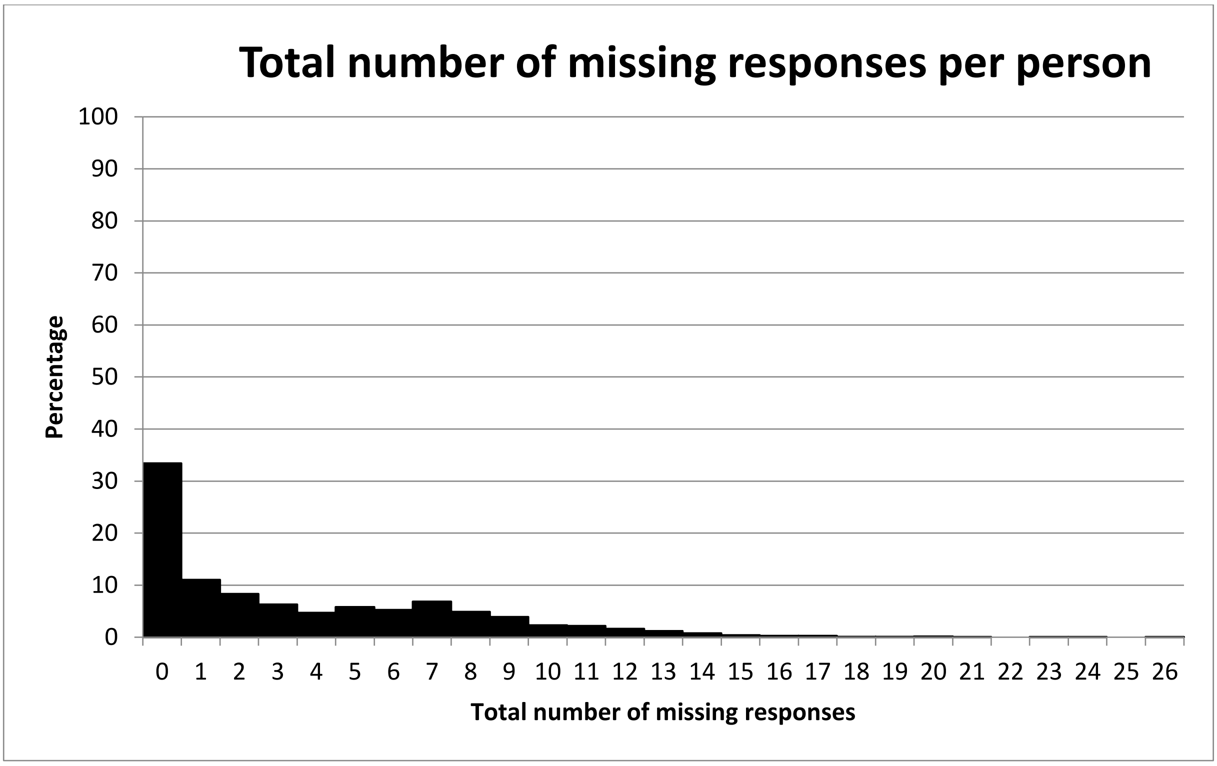

Figure 5. Total number of missing responses per person.

5.2.2 Missing responses per item

Table 5 shows the number of valid responses for each item as well as the percentage of missing responses. Overall, omission rates were rather low, varying across items between 0.4% and 7.3%. Thus, there was no item with an omission rate exceeding 10.0%. The number of missing responses was uncorrelated (r = .067, p = .745) with the difficulty of the item. This result indicates that the test-takers did not omit more difficult items. Generally, the percentage of invalid responses per item was rather low with the maximum rate being 0.7% (item scg90630_c). The relative frequency of not reached items increased towards the end of the test. Eventually, 50.9% of the students did not reach the last item and, thus, did not complete the test. The total number of missing responses per item varied between 1.4% and 51.3%.

Table 4: Valid Responses and Missing Values

Item	Position in the test	Number of valid responses	Not reached items (%)	Omitted items (%)	Invalid responses (%)
scg9611s_c	1	2793	0.0	5.8	0.1
scg96120_c	2	2900	0.0	1.8	0.5
scg91030_c	3	2790	0.1	5.7	0.2
scg91040_c	4	2927	0.1	1.0	0.3
scg91050_c	5	2907	0.1	1.9	0.2
scg96420_c	6	2815	0.1	4.9	0.2
scg9042s_c	7	2878	0.1	2.8	0.1
scg9043s_c	8	2840	0.1	4.2	0.0
scg90110_c	9	2896	0.2	2.1	0.2
scg9012s_c	10	2851	0.2	3.6	0.1
scg90510_c	11	2883	0.4	2.3	0.2
scg9052s_c	12	2751	0.7	6.7	0.0
scg91110_c	13	2876	0.9	1.9	0.3
scg91120_c	14	2797	1.4	3.7	0.6
scg97410_c	15	2807	2.1	3.2	0.2
scg6142s_c	16	2717	3.9	4.5	0.1
scg6144s_c	17	2692	4.9	4.4	0.1
scg90320_c	18	2578	8.6	4.3	0.2
scg90330_c	19	2383	12.3	7.3	0.2
scg9061s_c	20	2199	19.0	6.7	0.2
scg90630_c	21	2205	22.3	2.8	0.7
scg9651s_c	22	2072	27.7	2.4	0.1
scg96530_c	23	1972	32.3	1.1	0.1
scg90930_c	24	1804	38.1	0.7	0.4
scg9621s_c	25	1548	46.2	1.6	0.0
scg96220_c	26	1446	50.9	0.4	0.0

Table 5: Item parameters

No.	Item	Percentage correct	Difficulty/location parameter	*SE* (difficulty/ location parameter)	WMNSQ	*t*-value for WMNSQ	Pt.-bis. Corr. of correct response	Discrimination (GPCM)	Yens Q3
1	scg9611s_c	n.a.	−0.811	0.041	1.04	1.90	0.35	0.92	0.12
2	scg96120_c	71.4	−1.047	0.044	1.02	0.80	0.35	0.72	0.04
3	scg91030_c	47.7	0.097	0.041	0.98	−1.50	0.43	0.86	0.05
4	scg91040_c	76.3	−1.328	0.046	1.04	1.90	0.31	0.64	0.06
5	scg91050_c	55.0	−0.237	0.040	1.01	0.70	0.41	0.78	0.05
6	scg96420_c	47.7	0.103	0.040	0.97	-2.50	0.46	0.96	0.07
7	scg9042s_c	n.a.	−1.270	0.038	1.02	1.10	0.40	1.06	0.05
8	scg9043s_c	n.a.	−0.193	0.038	0.96	−1.90	0.48	1.49	0.05
9	scg90110_c	60.2	−0.473	0.041	1.03	1.90	0.36	0.65	0.05
10	scg9012s_c	n.a.	−1.570	0.043	0.98	−0.90	0.41	1.37	0.07
11	scg90510_c	54.9	−0.228	0.040	1.04	3.10	0.36	0.66	0.05
12	scg9052s_c	n.a.	−0.956	0.048	0.97	−1.00	0.43	1.67	0.06
13	scg91110_c	46.3	0.173	0.040	1.11	7.70	0.28	0.40	0.05
14	scg91120_c	25.1	1.250	0.046	1.02	1.00	0.33	0.69	0.05
15	scg97410_c	13.2	2.116	0.058	1.00	−0.10	0.29	0.78	0.07
16	scg6142s_c	n.a.	-2.503	0.051	0.94	−1.90	0.43	1.94	0.07
17	scg6144s_c	n.a.	−1.723	0.057	0.97	−1.30	0.37	1.09	0.06
18	scg90320_c	59.8	−0.431	0.043	0.92	-5.60	0.53	1.36	0.04
19	scg90330_c	37.4	0.611	0.045	1.02	1.20	0.38	0.72	0.04
20	scg9061s_c	37.3	0.610	0.047	0.97	−1.90	0.46	1.03	0.04
21	scg90630_c	48.3	0.097	0.046	0.96	-2.50	0.47	1.01	0.06
22	scg9651s_c	n.a.	−1.016	0.043	0.98	−0.70	0.47	1.32	0.12
23	scg96530_c	52.1	−0.086	0.049	1.00	0.00	0.43	0.87	0.09
24	scg90930_c	60.8	−0.500	0.052	1.06	3.10	0.34	0.60	0.05
25	scg9621s_c	n.a.	−0.806	0.052	0.91	-3.30	0.55	2.02	0.06
26	scg96220_c	52.4	−0.129	0.057	1.06	2.90	0.36	0.64	0.04

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

5.3 Parameter estimates

5.3.1 Item parameters

Column 3 in Table 5 shows the percentage of correct responses in relation to all valid responses for each item. Note that since there was a non-negligible amount of missing responses, this probability cannot be interpreted as an index for item difficulty. The percentage of correct responses within items varied between 13.2% and 76.3% with an average of 47.2% (*SD* = 15.4) correct responses.

The estimated item difficulties (for dichotomous items, MC items) and location parameters (for polytomous variables, CMC items) are also given in Table 5. The step parameters (for polytomous variables) are depicted in Table 6. All CMC items showed less than *N* = 200 participants in the lowest category, thus the two lowest categories were collapsed. These items were scaled using a scoring of 0, 0.5, 1, and 1.5. Additionally, for one of the CMC items (scg6144_c) one of the subtasks was eliminated due to negative discrimination. Thus, this item was scaled using a scoring of 0, 0.5, and 1. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged between −2.50 (scg6142s_c) and 2.12 (scg97410_c). In total, the estimated item difficulties had a mean of −0.39 (*SD* = 0.96). Due to the large sample size, the standard errors of the estimated item difficulties were very small (*SE*(ß) ≤ 0.058). Overall, the item difficulties were rather low; the test did not include many items with high difficulty.

Table 6: Step parameters for the CMC items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3 (*SE*)
scg9611s_c	−1.226 (0.040)	1.017 (0.046)	0.210
scg9042s_c	−0.504 (0.039)	0.858 (0.049)	−0.354
scg9043s_c	−0.378 (0.038)	0.159 (0.043)	0.219
scg9012s_c	−1.191 (0.039)	0.629 (0.041)	0.561
scg9052s_c	−0.492 (0.041)	−0.799 (0.038)	1.290
scg6142s_c	−0.616 (0.042)	0.459 (0.047)	0.157
scg6144s_c	−0.007 (0.044)	0.007	−
scg9651s_c	−0.409 (0.046)	1.076 (0.064)	−0.667
scg9621s_c	−0.322 (0.051)	0.113 (0.057)	0.209

Note. The last step parameters are not estimated and have, thus, no standard error because they are constrained parameters for model identification.

5.3.2 Person parameters

Person parameters are estimated as WLEs (Pohl & Carstensen, 2012). 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 (2012).

5.3.3 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 6, the difficulties of the scientific literacy items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties.

The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.677, indicating a somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .747; WLE reliability = .717) was acceptable. Although the items covered a wide range of the ability distribution, no items were covering the upper peripheral ability areas. As a consequence, person abilities in low and medium ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

Scale in logits	Person ability	Item difficulty
		15
2		
	X	
	XX	
	XX	
	XX	
	XX	
	XXX	14
	XXX	
1	XXXX	
	XXXX	
	XXXXXX	
	XXXXXX	
	XXXXXX	19 20
	XXXXXXX	
	XXXXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	13
	XXXXXXXX	3 6 21
0	XXXXXXX	23
	XXXXXXX	8 26
	XXXXXXXX	5 11
	XXXXXXXX	18
	XXXXXXXXXX	9 24
	XXXXXXXX	
	XXXXXXX	
	XXXX	1 25
	XXXXX	12
−1	XXXXX	2 22
	XXX	
	XXX	7
	XX	4
	XX	
	X	10
	X	17
	X	
	X	
−2		
		16

Figure 6. Test targeting. The distribution of person abilities in the sample is depicted on the left side of the graph. Each ‘X’ represents 18.1 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 5).

5.4 Quality of the test

5.4.1 Fit of the subtasks of complex multiple-choice items

Before the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of the CMC item separately, there were 58 items. The percentage of a correct response ranged from 8.5% to 92.0% across all items (*Mdn* = 71.3%). Thus, the number of correct and incorrect responses was reasonably large. All but one subtask of the CMC items showed a satisfactory item fit. One subtask (scg6144a) showed a negative discrimination, and was, therefore, excluded from further analysis. The remaining 57 items showed a good WMNSQ, ranging from 0.90 to 1.13. The respective *t*-value ranged from −6.7 to 9.8, and there were no noticeable deviations of the empirically estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to a polytomous variable seemed justified.

5.4.2 Distractor analyses

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 five of the ten CMC items, there was one distractor with a point-biserial correlation with the total scores over zero: scg9611s_c (0.05), scg9043s_c (0.07), scg9012s_c (0.05), scg9651s_c (0.06), and scg9621s_c (0.03). All of the other items only had distractors with a point-biserial correlation with the total scores below zero. Besides these five deviations, the results indicate that the distractors worked well.

5.4.3 Item fit

The evaluation of the item fit was performed based on the final scaling model, the partial credit model, using the MC items and the CMC items. Altogether, the item fit can be considered to be very good (see Table 5). Values of the WMNSQ ranged from 0.91 (item scg9621s_c) to 1.11 (item scg91110_c). Only one item exhibited a *t*-value of the WMNSQ greater than 6 (item scg91110_c). Thus, there was no indication of a severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .28 (items scg91110_c) to .55 (items scg9621s_c) and had a mean of .40. All item characteristic curves showed a good fit of the items to the PCM.

5.4.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables rotation (test order), gender, the number of books at home (as a proxy for socioeconomic status), migration background, and school type (see Pohl & Carstensen, 2012, for a description of these variables). Table 7 shows the absolute 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. Also, Table 8 shows the main effect for the examined subgroups (inclusive Cohen's d).

Table 7: Differential item functioning (differences between difficulties)

Item	Rotation	Gender	Books			Migration status			School type		
	First vs. Second	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing	Other vs. Gym.	Other vs. Missing	Gym. vs. Missing
scg9611s_c	0.100	0.008	−0.352	0.248	0.600	0.010	0.118	0.108	−0.328	−0.208	0.126
scg96120_c	0.202	−0.216	−0.084	0.110	0.198	0.198	−0.018	−0.218	0.246	0.382	0.138
scg91030_c	0.084	−0.136	0.090	0.260	0.174	−0.110	−0.030	0.080	0.112	0.306	0.198
scg91040_c	−0.014	−0.212	−0.068	0.342	0.412	0.042	0.166	0.124	−0.286	0.132	0.422
scg91050_c	0.084	−0.228	0.020	−0.318	−0.336	−0.338	−0.204	0.136	−0.104	−0.132	−0.026
scg96420_c	0.214	0.412	0.262	−0.172	−0.432	−0.292	−0.336	−0.046	0.126	−0.018	−0.142
scg9042s_c	0.016	−0.192	−0.020	−0.088	−0.070	0.350	0.096	−0.254	−0.198	−0.174	0.020
scg9043s_c	−0.070	0.058	0.062	−0.138	−0.192	0.066	−0.066	−0.138	0.580	0.200	−0.348
scg90110_c	0.038	0.072	−0.050	0.006	0.060	0.014	0.050	0.034	−0.120	−0.018	0.104
scg9012s_c	−0.042	0.378	0.204	0.244	0.042	−0.148	0.092	0.238	0.132	0.064	−0.054
scg90510_c	0.004	−0.294	−0.004	0.218	0.224	0.104	0.098	−0.006	−0.202	−0.010	0.194
scg9052s_c	−0.076	−0.044	0.300	0.290	−0.010	0.214	−0.030	−0.258	0.200	0.080	−0.136
scg91110_c	−0.014	−0.040	−0.392	0.240	0.636	0.290	0.394	0.102	−0.334	0.050	0.388
scg91120_c	−0.196	−0.082	−0.034	−0.094	−0.058	0.300	−0.044	−0.346	−0.228	−0.224	0.006
scg97410_c	−0.052	−0.010	−0.130	−0.320	−0.186	0.254	−0.024	−0.280	−0.124	0.216	0.344
scg6142s_c	−0.054	−0.116	0.082	−0.742	−0.796	−0.128	−0.178	−0.044	0.500	0.056	−0.458
scg6144s_c	−0.070	−0.604	0.042	0.134	0.080	−0.192	−0.142	0.056	0.034	−0.028	−0.062
scg90320_c	−0.110	−0.030	0.202	−0.010	−0.212	−0.098	−0.002	0.094	0.186	−0.046	−0.230
scg90330_c	0.028	−0.090	−0.188	−0.056	0.136	−0.028	0.140	0.168	0.110	0.206	0.098
scg9061s_c	0.000	0.316	0.022	−0.588	−0.608	−0.098	−0.086	0.010	0.050	−0.100	−0.148
scg90630_c	−0.008	0.762	0.022	0.266	0.248	0.018	−0.048	−0.068	0.010	−0.038	−0.046
scg9651s_c	0.022	−0.014	−0.120	−0.138	−0.026	−0.112	0.050	0.158	−0.090	−0.272	−0.208
scg96530_c	−0.064	0.092	0.042	0.308	0.270	−0.040	0.004	0.042	−0.120	−0.074	0.046
scg90930_c	−0.100	−0.282	−0.018	0.226	0.246	−0.010	−0.038	−0.030	−0.180	−0.152	0.030
scg9621s_c	0.006	0.116	0.220	−0.638	−0.776	−0.260	−0.196	0.060	0.192	−0.184	−0.336
scg96220_c	−0.242	0.282	0.100	0.114	0.016	−0.082	0.092	0.172	0.012	0.044	0.034

Rotation

The scientific literacy test was administered in two different positions (see section 3.1 for the design of the study). A total of 1,467 (49.4%) of the test takers received the scientific literacy test first and then the mathematical or reading test (coded 0), while 1,500 (50.6%) received the mathematical literacy or reading test before completing the scientific literacy test (coded 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 participants. There was only a small difference between the first test position and the second test position (main effect = −0.024 logits, Cohen's d = −0.029), indicating a lower difficulty for test-takers with the second test position. Also, the highest difference in difficulties between the two groups is −0.242 logits.

Gender

The sample included 1,462 (49.3%) male test-takers (coded 0) and 1,505 (50.7%) female test-takers (coded 1). On average, male students had slightly higher scores in scientific literacy than female students (main effect = 0.048 logits, Cohen's d = 0.058). There was also one item with a considerable gender DIF (highest DIF = 0.762 logits).

Books

The number of books at home was used as a proxy for socioeconomic status. There were 789 (26.6%) test takers with 0 to 100 books at home (coded 0), 2,039 (68.7%) test takers with more than 100 books at home (coded 1), and 139 (4.7%) test-takers did not give a valid response (coded 9). DIF was investigated using these three groups. There were considerable average differences between these three groups. Participants with 100 or fewer books at home on average showed lower scientific literacy scores than participants with more than 100 books (main effect = −0.480 logits, Cohen's d = −0.613). Participants without a valid response on the variable 'books at home' performed lower than participants with up to 100 (main effect = 0.130 logits, Cohen's d = 0.165) and lower than participants with more than 100 books at home, respectively (main effect = 0.612 logits, Cohen's d = 0.769). There was no considerable DIF comparing participants with many or fewer books (highest DIF = −0.392). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to 0.636 logits.

Migration background

There were 2,084 (70.2%) participants without a migration background (coded 0) and 507 (17.1%) participants with a migration background (for 0.5% students neither their mother, father or themselves were born in Germany, for 5.6% only the participants were born in Germany and both of their parents were born abroad, and for 11.0% of the participants only one of their parents was born abroad, coded 1). A total of 376 (12.7%) students could not be allocated to either group. These groups were used for investigating DIF of migration. There was a considerable difference in the average performance of participants with or without migration background. Participants without a migration background showed higher scientific literacy scores than participants with a migration background (main effect = 0.306 logits, Cohen's d = 0.382) and also higher scores than students with an unknown background on migration (main effect = 0.336 logits, Cohen's *d* = 0.415). Furthermore, students with a migration background scored higher than those with an unknown background on migration (main effect = 0.030 logits, Cohen's *d* = 0.036). There was no considerable DIF comparing participants with and without a migration background (highest DIF = 0.350). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to 0.394 logits.

Type of School

DIF was also investigated for the type of secondary school. At the end of primary school, children in Germany will be mainly allocated for secondary school to one of the following types: "Hauptschule", a secondary general school for Grades five through nine or ten, "Realschule", a more practical secondary school for Grades five through ten, or "Gymnasium", a more academic secondary school for Grades five through twelve/thirteen. There were 1,572 (53.0%) students visiting "Gymnasium" (coded 1), and 1,063 (35.8%) students from lower schools (coded 0), such as "Hauptschule" or "Realschule". A total of 332 (11.2%) students could not be allocated to either group (coded 9). On average, students visiting "Gymnasium" had distinctly higher scores in scientific literacy than students from other school types (main effect = −0.728 logits, Cohen's d = −0.980), or students without valid responses (main effect = 0.568 logits, Cohen's d = 0.752). Students from lower schools showed lower scientific literacy than students from other school types (main effect = −0.164 logits, Cohen's d = −0.222). There was no considerable DIF comparing students visiting "Gymnasium" and students from other school types (highest DIF = 0.580). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to −0.458 logits.

Table 8: Main effects and Cohen's d of the examined subgroups

Variables	Subgroups	Main effect	Cohen`s d
Rotation	Science first (0) Science second (1)	−0.024	−0.029
Gender	Male (0) Female (1)	0.048	0.058
Books	0 to 100 books at home (0) More than 100 books at home (1)	−0.480	−0.613
	0 to 100 books at home (0) Invalid response (9)	0.130	0.165
	More than 100 books at home (1) Invalid response (9)	0.612	0.769
Migration background	Without migration background (0) With migration background (1)	0.306	0.382
	Without migration background (0) Invalid response (9)	0.336	0.415
	With migration background (1) Invalid response (9)	0.030	0.036
School type	Other school types (0) Gymnasium (1)	−0.728	−0.980
	Other school types (0) Invalid response (9)	−0.164	−0.222
	Gymnasium (1) Invalid response (9)	0.568	0.752

Note. The numbers behind the subgroups display their coding.

Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models that allow for DIF with those that allow only for main effects. In Table 9, the models including only the main effects are compared with those that additionally estimate DIF. For these models, we used the valid responses from the participants. For example, the variable books represents the comparison of the participants with less than 100 books and those with more than 100 books. Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC, Schwarz, 1978) were used for comparing the models. The AIC favored the model considering DIF for four DIF variables (rotation, gender, books, and school type). Only for the migration background, the AIC favored the model which allows only for main effects. 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 four of the three DIF variables (gender, books, migration background, and school type). Only for gender, the BIC preferred the model which allows DIF.

Table 9: Comparison of models with and without DIF

DIF variable	Model	Deviance	N	Number of parameters	AIC	BIC
Rotation	main effect	105549.37	2967	45	105639.37	105909.16
	DIF	105336.98	2967	71	105478.98	105904.65
Gender	main effect	100419.17	2828	45	100509.17	100776.80
	DIF	100349.21	2828	71	100491.21	100913.47
Books	main effect	92247.38	2591	45	92337.38	92601.07
	DIF	92185.66	2591	71	92327.66	92743.71
Migration background	main effect	105550.72	2967	45	105640.72	105910.51
	DIF	105520.04	2967	71	105662.04	106087.71
School type	main effect	93040.98	2635	45	93130.98	93395.43
	DIF	92889.33	2635	71	93031.33	93448.57

Note. The results of the variables books, migration background, and school type display main effect and DIF between the valid responses.

5.4.5 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 5), ranging from 0.40 (item scg91110_c) to 2.02 (item scg9621s_c). The average discrimination parameter fell at 1.01. Model fit indices suggested a better model fit of the GPCM (AIC = 105,314.51, BIC = 105,728.18) as compared to the PCM model (AIC = 105,639.22, BIC = 105,903.02). Despite the empirical preference for the GPCM, the PCM model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.4.6 Unidimensionality of the test

The dimensionality 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: the process-related components (knowledge about science – KAS) and the content-related components (knowledge of science – KOS; for more details see Hahn et al., 2013). For estimating a two-dimensional model Gauss' Hermite quadrature estimation in ConQuest was used (nodes were chosen in such a way that stable parameter estimation was obtained). The unidimensional model (BIC = 105,903.02, number of parameters = 44) fitted the data slightly better than the two-dimensional model (BIC = 105,916.63, number of parameters = 46). Also, the correlation between the two dimensions was very high (r = .94). So the one-dimensional measurement model was used to estimate a single competence score for scientific literacy.

6 Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the science test administered in Grade 7 of starting cohort 2 and at describing how scientific literacy was estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We checked item fit statistics for simple MC items, subtasks of CMC items, as well as the polytomous CMC items and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence.

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number of missing responses was reasonably small.

The test had acceptable reliability and distinguished well between test-takers. The test's variance was acceptable.

Indicated by various fit criteria – WMNSQ, *t*-value of the WMNSQ – the items exhibited a good item fit. Also, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with total score) were acceptable. Different variables were used for testing measurement invariance across various subgroups. Only a few items showed considerable DIF for the examined variables, indicating that the test was fair to the considered subgroups.

Fitting a two-dimensional partial credit model (the dimensions being the "content-related components" and the "process-related components") yielded no better model fit than the unidimensional partial credit model. Also, the high correlation between the two dimensions indicates that a unidimensional model describes the data reasonably well.
Summarizing the results, the test had good psychometric properties that facilitated the estimation of a unidimensional scientific literacy score.

7 Data in the Scientific Use file

7.1 Naming conventions

There are 26 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 '_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 was scored with 0.5 points. Manifest scale scores are provided in form of WLE estimates (scg7_sc1) including the respective standard error (scg7_sc2). Please note that when categories of the polytomous variables had less than 200 valid responses, the categories were collapsed. For the science test, this concerned the two lowest categories of all of the polytomous items (see section 5.3.1**Fehler! Verweisquelle konnte nicht gefunden werden.** on the aggregation of CMC items). In the scaling model, the collapsed polytomous item was scored in steps of 0, 0.5, 1.0, and 1.5 (denoting the highest). Additionally, for one of the CMC items (scg6144_c) one of the subtasks was eliminated due to negative discrimination. Thus, this item was scaled using a scoring of 0, 0.5, and 1. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. Students who did not take part in the test or those who did not have enough valid responses to estimate a scale score have a non-determinable missing value on the WLE score for scientific literacy.

7.2 Scientific literacy scores

In the SUF manifest science literacy scores are provided in the form of WLEs (scg7_sc1) including their respective standard error (scg7_sc2).

The estimated WLE scores were corrected for differences in the test position because the science test was either presented as the first or the second test within the test battery.

Unlike the previous competence measurements (Kindergarten, Grade 1 and Grade 3), there are no competence scores which can be used for longitudinal comparisons. The scientific literacy tests which were administered in Grade 3 and 7, included different items that were constructed in such a way as to allow for an accurate measurement of scientific literacy within each age group. As a consequence, the competence scores derived in the different Grades cannot be directly compared. Differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across Grades, all items from the Grade 3 and the Grade 7 scientific literacy tests would have to be administered in an independent link sample – including students from Grade 7 that were not part of Starting Cohort 3 – within a single measurement occasion. However, the two measurement points (Grade 3 and 7) are too far apart to provide reliable data from a sample of students from Grade 7 participating on the scientific literacy test of Grade 3. Therefore, the previous competence scores can only serve as predictors for the competence scores in Grade 7.

The ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the science test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

8 References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software (Version 4) [Computer software]. Camberwell: Australian Council for Educational Research.

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

Blossfeld, H.-P., Roßbach, H.-G., & Maurice, J. v. (2011). Education as a Lifelong Process − The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 14*.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6).* Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

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

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

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

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 14*, 67–86.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Appendix

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

```
Title G7 Science analysis, Partial Credit Model;
data filename.dat;
format id 1–7 responses 8–33;

labels << filename_with_labels.txt;

recode (0,1,2,3)      (0,0,1,2)        !item (17);
recode (0,1,2,3,4)    (0,0,1,2,3)      !item (1,7,8,10,12,16,22,25);

codes 0,1,2,3;
score (0,1)           (0,1)            !item (2-6,9,11,13-15,18-21,23-24,26);
score (0,1,2)         (0,0.5,1)        !item (17);
score (0,1,2,3)       (0,0.5,1,1.5)    !item (1,7,8,10,12,16,22,25);

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

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

Appendix B: Assignment of items to the content and process-related components and contexts

Variable name	Position in the test	Component	Context
scg9611s_c	1	KAS	Health
scg96120_c	2	KAS	Health
scg91030_c	3	KOS	Technology
scg91040_c	4	KOS	Technology
scg91050_c	5	KOS	Technology
scg96420_c	6	KAS	Technology
scg9042s_c	7	KOS	Environment
scg9043s_c	8	KOS	Environment
scg90110_c	9	KOS	Health
scg9012s_c	10	KOS	Health
scg90510_c	11	KOS	Environment
scg9052s_c	12	KOS	Environment
scg91110_c	13	KOS	Technology
scg91120_c	14	KAS	Technology
scg97410_c	15	KOS	Technology
scg6142s_c	16	KOS	Technology
scg6144s_c	17	KOS	Technology
scg90320_c	18	KOS	Technology
scg90330_c	19	KOS	Technology
scg9061s_c	20	KOS	Health
scg90630_c	21	KOS	Health
scg9651s_c	22	KAS	Environment
scg96530_c	23	KAS	Environment
scg90930_c	24	KOS	Environment
scg9621s_c	25	KAS	Environment
scg96220_c	26	KAS	Environment

Note. KOS = knowledge of science (content-related components); KAS = knowledge about science (process-related components)