NEPS SURVEY PAPERS

Jana Kähler

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 3

NEPS Survey Paper No. 60
Bamberg, November 2019

LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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 2 for Grade 3

Jana Kähler

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

Email address of the lead author:
jkaehler@ipn.uni-kiel.de

Bibliographic data:

Kähler, J. (2019). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 2 for Grade 3* (NEPS Survey Paper No. 60). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP60:1.0

NEPS Technical Report for Science:

Scaling Results of Starting Cohort 2 for Grade 3

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. In order 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 3 of starting cohort 2. The scientific literacy test contained 23 items with different response formats representing different contexts as well as different areas of knowledge. The test was administered to 5,800 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 but one fitted the model in a satisfactory way. 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. Additionally, the design and results of the linking study for the competence scores in grades 1 and 3 are presented.

Key words: scientific literacy, 3rd grade, linking grade 1 and 3, 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 competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication literacy (computer literacy), metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competences 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 3 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 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 one dimensional 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 matter, system, development and interaction. KAS is divided into the process-related components 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.

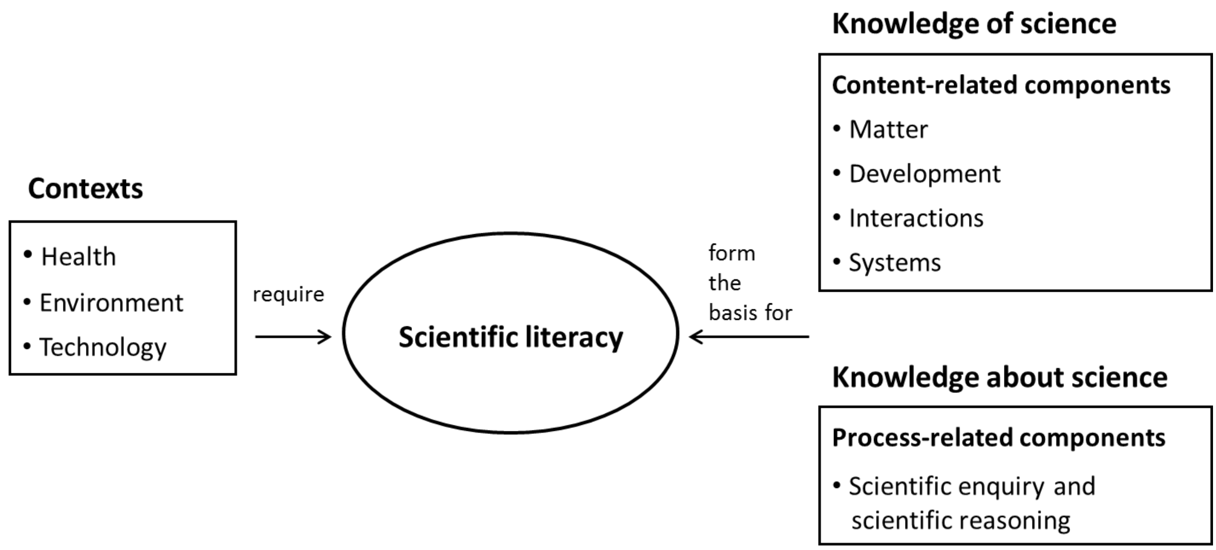

Figure 1. Assessment framework for scientific literacy (Hahn et al., 2013).

In the scientific literacy test for grade 3 of starting cohort 2 (Kindergarten) there were three types of response formats. These were simple multiple choice (MC), complex multiple choice (CMC) in the special form of true false items, and short constructed response (SCR). 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. And in SCR items test takers had to write down a short answer into an empty box.

3 Data

3.1 The design of the study

The study assessed different competence domains including, among others, scientific literacy and computer literacy. The competence tests for these two domains were always presented first within the test battery. For each participant the scientific literacy test was administered as the first test followed by the computer competence test. There was no multi-matrix design regarding the order of the items within a specific test. All students received the test items in the same order. The test consisted of 23 items which were administered in a test time of 30 minutes.

The allocation of the 23 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)	17
Knowledge about Science (KAS)	6
Total number of items	23

Table 2

Number of Items by Different Contexts

Context	**Number of Items**
Health	8
Environment	9
Technology	6
Total number of items	23

Table 3

Number of Items by Response Formats

Response format	**Number of Items**
Simple Multiple-Choice	17
Complex Multiple-Choice (True false items)	4
Short constructed response	2
Total number of items	23

3.2 Sample

A total of 5,800 individuals received the scientific literacy test. For 180 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 was excluded from further analyses (see Pohl

& Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 5,620 individuals (51% 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

A total of 22 items (including all subtasks for the polytomous items) were included in the analyses. One item (scg31610_c) had to be excluded due to insufficient item quality.

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. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. As CMC items 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 gave an indication of how well the persons were coping with the test. We then looked at the occurrence of missing responses per item in order 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. When categories of the polytomous variables had less than N = 200, the categories were collapsed in order to avoid any 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). One of the SCR items (scg30520_c) was scored in three categories (0 = wrong answer, 1 = partly correct answer, 2 = correct answer). None of these categories had to be collapsed. To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 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. In order 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 (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*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. Overall judgment of the fit of an item was based on all fit indicators.

The 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 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 scale.

Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) Q3. Because in 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

5.1 Descriptive statistics of the responses

In order 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 15.2% to 73.2% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 19.2% and 49.4%. From a descriptive point of view, the items covered a rather wide range of difficulties.

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. 93.5% of the respondents did not have any invalid response at all; overall, less than 1.0% had more than one invalid response.

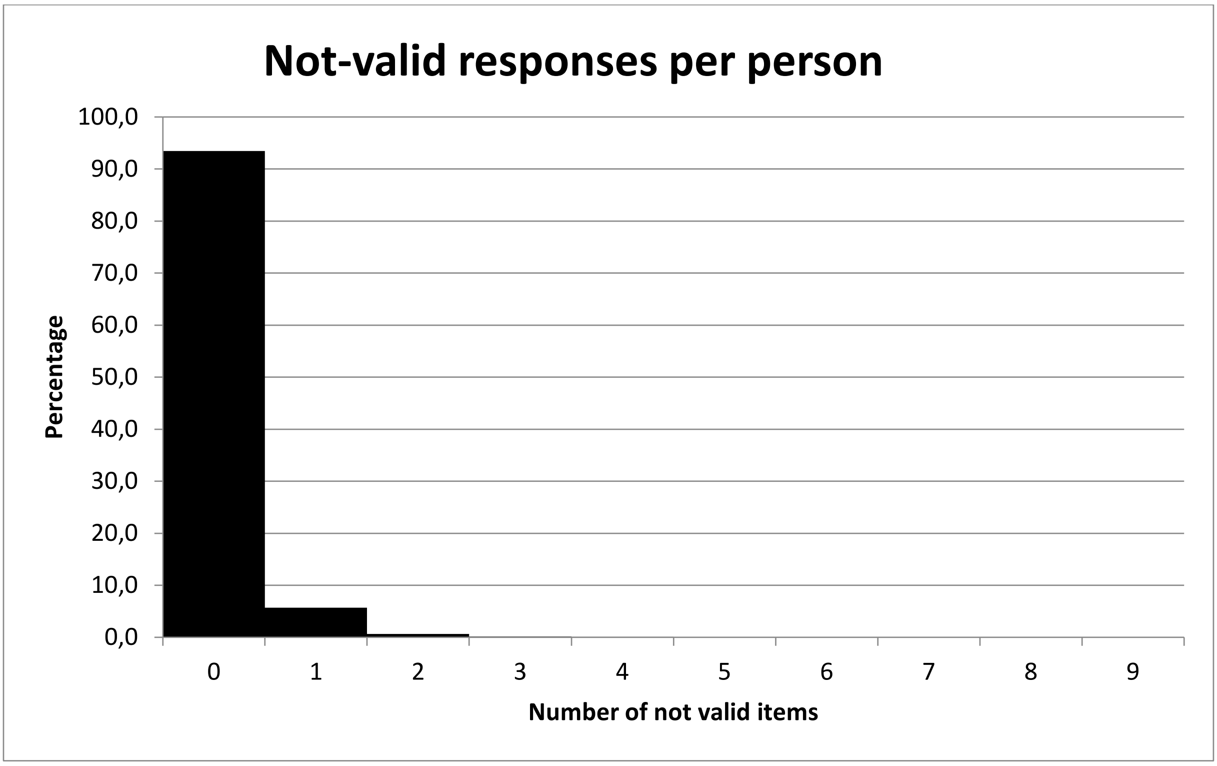

Figure 2. Number of invalid responses per person.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 60.9%, did not skip any item, and less than 3.0% omitted more than three items.

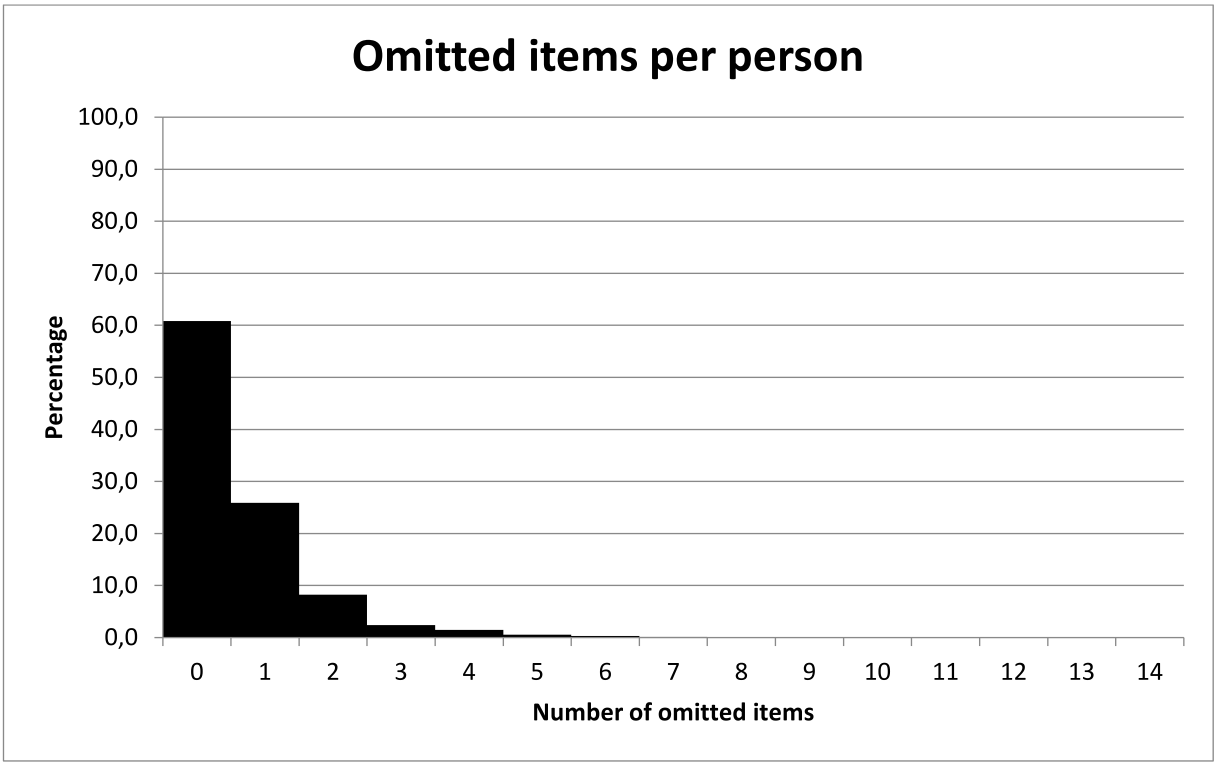

Figure 3. Number of omitted responses per person.

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 rather high, only 48.2% of the respondents were able to finish the test within the allocated time limit (Figure 4). This might be a consequence of some participants taking too much time on some items (particularly, SCR items) and the test administrators neglecting to emphasize that students should skip respective items.

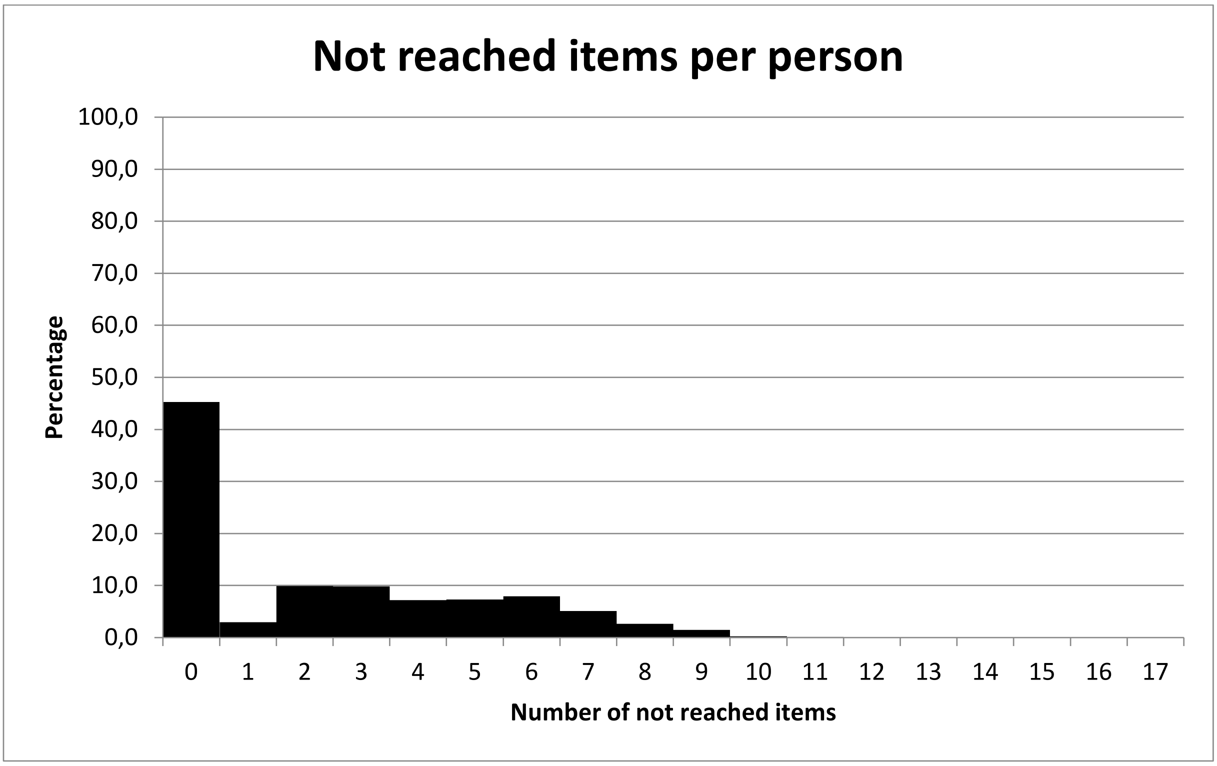

Figure 4. Number of not reached items per person.

The total number of missing responses, aggregated over invalid, omitted and not-reached missing responses, is illustrated in Figure 5. 25.8% of the students answered all questions and, consequently, had no missing responses. Only 0.3% of the students had missing responses on more than half of the items. Hence, the amount of missing responses per person can be classified as medium.

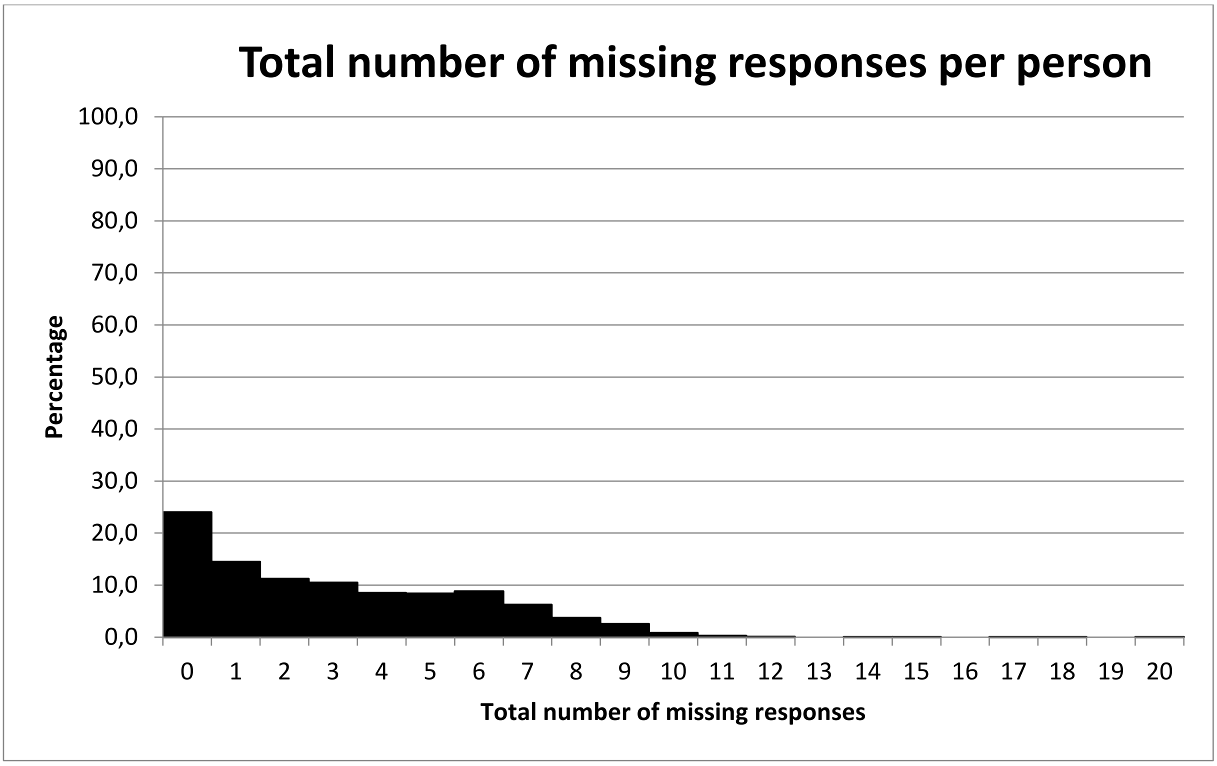

Figure 5. Total number of missing responses per person.

5.2.2 Missing responses per item

Table 4 shows the number of valid responses for each item as well as the percentage of missing responses. Overall, omission rates were rather low, varying across items between 0.3% and 5.3%. There were only two items with an omission rate exceeding 6% (scg37110_c and scg30520_c). The number of missing responses was uncorrelated (r = .08, p = .722) with the difficulty of the item. This result indicates that the test takers did not omit items that are more difficult. Generally, the percentage of invalid responses per item was rather low with the maximum rate being 1.4% (item scg37110_c). The relative frequency of not reached items increased towards the end of the test. Eventually, 51.8% 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 0.8% and 55.4%.

Table 4

Valid Responses and Missing Values

Item	Position in the test	Number of valid responses	Not reached items (%)	Omitted items (%)	Invalid responses (%)
scg30109_c	1	5561	0.0	1.0	0.1
scg33510_c	2	5573	0.0	0.8	0.1
scg3181s_c	3	5307	0.0	5.3	0.3
scg37110_c	4	4496	0.0	18.6	1.4
scg31010_c	5	5552	0.0	0.7	0.5
scg36710_c	6	5561	0.0	0.8	0.2
scg3131s_c	7	5346	0.0	4.6	0.2
scg34010_c	8	5558	0.0	0.9	0.2
scg32220_c	9	5472	0.0	2.4	0.2
scg33710_c	10	5547	0.0	1.2	0.1
scg36210_c	11	5529	0.0	1.3	0.3
scg36920_c	12	5572	0.0	0.7	0.1
scg32620_c	13	5517	0.1	1.4	0.3
scg31510_c	14	5556	0.4	0.5	0.2
scg30310_c	15	5484	1.9	0.3	0.2
scg3641s_c	16	5181	4.5	3.0	0.2
scg30520_c	17	4406	9.6	10.8	1.2
scg37410_c	18	4486	17.6	2.5	0.1
scg33310_c	19	4181	24.9	0.5	0.2
scg3091s_c	20	3682	32.1	2.2	0.2
scg33610_c	21	3223	41.9	0.5	0.2
scg32910_c	22	2509	51.8	2.5	1.0

Note. The item on position 23 (scg31610_c) was excluded from the analyses due to insufficient item quality (see section 4).

Table 5

Item parameters

Item	Percentage correct	Difficulty/location parameter	*SE* (difficulty/ location parameter)	WMNSQ	*t*-value for WMNSQ	Pt. bis. Corr. of correct response	Discrimination (2PL)	Yens Q3
scg30109_c	73.2	−1.130	0.032	1.03	1.7	.33	0.62	.06
scg33510_c	71.5	−1.038	0.031	0.98	−1.2	.40	0.87	.07
scg3181s_c	n.a.	−0.362	0.030	1.00	−0.2	.34	1.21	.09
scg37110_c	58.9	−0.371	0.032	0.99	−0.8	.41	0.80	.06
scg31010_c	67.6	−0.832	0.030	1.03	2.8	.33	0.57	.06
scg36710_c	50.9	−0.043	0.029	1.03	3.1	.37	0.65	.04
scg3131s_c	n.a.	−0.354	0.032	1.00	−0.3	.28	1.21	.06
scg34010_c	49.3	0.031	0.029	0.93	−7.4	.51	1.22	.07
scg32220_c	41.1	0.409	0.029	0.99	−1.0	.43	0.86	.06
scg33710_c	58.2	−0.374	0.029	0.99	−1.4	.42	0.85	.05
scg36210_c	20.0	1.556	0.035	0.99	−0.6	.33	0.75	.06
scg36920_c	48.4	0.072	0.029	1.06	6.3	.32	0.49	.04
scg32620_c	66.1	−0.758	0.030	0.97	−2.4	.43	0.91	.04
scg31510_c	37.7	0.569	0.029	1.04	3.6	.34	0.56	.05
scg30310_c	36.8	0.609	0.030	0.96	−3.8	.44	0.91	.05
scg3641s_c	n.a.	−1.194	0.028	0.92	−5.0	.50	2.00	.09
scg30520_c	15.2	1.682	0.042	0.95	−2.4	.36	1.35	.07
scg37410_c	51.0	−0.054	0.032	1.00	0.1	.40	0.76	.07
scg33310_c	66.7	−0.803	0.035	1.02	1.2	.37	0.69	.04
scg3091s_c	n.a.	−0.585	0.034	1.03	1.5	.26	0.98	.07
scg33610_c	59.6	−0.466	0.038	1.02	1.4	.37	0.67	.04
scg32910_c	36.3	0.594	0.044	1.08	4.9	.27	0.37	.06

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

5.3 Parameter estimates

5.3.1 Item parameters

Column 2 in Table 5 shows the percentage of correct responses in relation to all valid responses for each item. Note that since there 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 15.2% and 73.2% with an average of 46.8% (*SD* = 17.8) correct responses.

The estimated item difficulties (for dichotomous items, MC items) and location parameters (for polytomous variables, CMC and SCR items) are given in Table 5. The step parameters (for polytomous variables) are depicted in Table 6. For all of the CMC items the two lowest categories were collapsed, thus, these items were scaled using a scoring of 0, 0.5, 1, and 1.5. One SCR item (scg30520_c) 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 –1.19 (scg3641s_c) and 1.68 (scg30520_c). In total, the estimated item difficulties had a mean of –0.13 (*SD* = 0.78). Due to the large sample size, the standard errors of the estimated item difficulties were very small (*SE*(ß) ≤ 0.044). Overall, the item difficulties were rather low; the test did not include items with a high difficulty (above 2 logits).

Table 6

Step parameters for the CMC items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3
scg3181s_c	−0.371 (0.028)	−0.470 (0.028)	0.841
scg3131s_c	−0.384 (0.029)	−0.849 (0.028)	1.232
scg3641s_c	−0.313 (0.029)	0.594 (0.035)	−0.281
scg30520_c	0.599 (0.041)	−0.599	
scg3091s_c	−0.647 (0.034)	0.331 (0.037)	0.316

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, 2012a). 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).

5.3.3 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 5, 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 0.613, indicating a somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .703; WLE reliability = .679) was acceptable. Although the items covered a wide range of the ability distribution, there were no items covering the lower and upper peripheral ability areas. As a consequence, person ability in medium ability regions will be measured relative precisely, whereas lower and higher ability estimates will have larger standard errors of measurement.

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

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each ‘X’ represents 35.9 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 4).

5.4 Quality of the test

5.4.1 Fit of the subtasks of complex multiple-choice items

Before the 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 35 items. The percentage of a correct response ranged from 15.2% to 90.7% across all items (*Mdn* = 59.2%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks of the CMC items showed a satisfactory item fit. WMNSQ ranged from 0.92 to 1.12, the respective *t*-value from −5.7 to 12.7, and there were no noticeable deviations of the empirical 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. All accept one distractor had a point-biserial correlation with the total scores below zero. One distractor had a point-biserial correlation of 0.02 (scg32910). Besides that, the results indicate that the distractors worked well.

5.4.3 Item fit

The evaluation of the item fit was performed on the basis of 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 0.92 (item scg3641s_c) to 1.08 (scg32910_c). Only one item exhibited a *t*-value of the WMNSQ greater than 6. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .26 (item scg3091s_c) to .51 (item scg34010_c) and had a mean of .37. 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 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). 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.

Gender

The sample included 2,866 (51%) female test takers and 2,753 (49%) male test takers. One test taker did not report the gender. On average, male students had slightly higher scores in scientific literacy than female students (main effect = 0.104 logits, Cohen's *d* = 0.133). But there was no item with a considerable gender DIF. The highest difference in difficulties between the two groups was −0.440 logits.

Books

The number of books at home was used as a proxy for socio-economic status. There were 1,627 (29.0%) test takers with 0 to 100 books at home, 3,105 (55.2%) test takers with more than 100 books at home, and 888 (15.8%) test takers did not give a valid response. DIF was investigated using these three groups. There were considerable average differences between these three groups. Participants with 100 or less books at home performed showed lower scientific literacy scores than participants with more than 100 books (main effect = −0.524 logits, Cohen's *d* = −0.723). Participants without a valid response on the variable 'books at home' performed lower than participants with up to 100 (main effect = −0.170 logits, Cohen's *d* = −0.240) and also lower than participants with more than 100 books at home (main effect = −0.692 logits, Cohen's *d* = −0,943). There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.412). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to −0.442 logits.

Migration background

There were 3,396 (60.4%) participants without a migration background and 889 (15.8%) participants with a migration background (for 0.8% students neither their mother, father or themselves were born in Germany, for 6.7% only the participants were born in Germany and both of their parents were born abroad, and for 8.3% of the participants only one of their parents was born abroad). A total of 1,335 (23.8%) 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 = 0.372 logits, Cohen's *d* = 0.494) and also higher scores than students with an unknown background on migration (main effect = 0.474 logits, Cohen's *d* = 0.628). Furthermore, students with a migration background performed better than those with an unknown background on migration (main effect = 0.100 logits, Cohen's *d* = 0.134). There was no considerable DIF comparing participants with and without a migration background (highest DIF = 0.400). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to −0.292 logits.

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 comparing the models. The AIC favored the model considering DIF for all three DIF variables. The Bayesian information criterion (BIC, Schwarz, 1978) 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 all three DIF variables.

Table 7

Differential item functioning (absolute differences between difficulties)

Item	Gender	Books			Migration status		
	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing
scg30109_c	−0.074	−0.054	−0.116	−0.068	−0.208	−0.104	0.104
scg33510_c	−0.118	0.242	−0.022	−0.270	−0.124	−0.188	−0.064
scg3181s_c	0.064	−0.104	0.090	0.188	0.030	0.130	0.100
scg37110_c	0.222	−0.182	0.030	0.208	−0.044	0.026	0.068
scg31010_c	0.028	−0.050	−0.060	−0.016	0.232	0.076	−0.158
scg36710_c	−0.124	0.070	0.282	0.208	−0.076	0.060	0.136
scg3131s_c	0.186	0.112	0.006	−0.122	−0.158	−0.188	−0.030
scg34010_c	−0.112	0.228	0.000	−0.234	−0.042	−0.138	−0.096
scg32220_c	−0.220	−0.068	0.020	0.082	0.024	0.122	0.098
scg33710_c	−0.282	0.162	0.000	−0.168	−0.108	−0.170	−0.062
scg36210_c	0.244	0.046	0.264	0.214	0.342	0.224	−0.118
scg36920_c	0.088	−0.324	−0.130	0.188	0.122	0.174	0.052
scg32620_c	−0.048	−0.024	0.002	0.018	−0.090	0.042	0.132
scg31510_c	0.184	−0.038	0.040	0.072	0.008	0.000	−0.010
scg30310_c	0.092	0.162	0.018	−0.150	−0.138	−0.194	−0.056
scg3641s_c	0.092	0.032	−0.122	−0.148	−0.094	−0.088	0.008
scg30520_c	−0.440	0.412	−0.032	−0.442	−0.400	−0.292	0.110
scg37410_c	0.054	−0.020	−0.008	0.006	0.240	0.174	−0.066
scg33310_c	0.180	−0.176	−0.048	0.122	−0.010	0.122	0.132
scg3091s_c	−0.038	−0.030	0.046	0.072	0.234	0.090	−0.128
scg33610_c	−0.018	−0.134	−0.274	−0.146	0.114	−0.088	−0.202
scg32910_c	−0.044	−0.280	0.062	0.336	0.174	0.188	0.014

Table 8

Comparison of models with and without DIF

DIF variable	Model	Deviance	N	Number of parameters	AIC	BIC
Gender	main effect	161995.69	5619	33	162061.69	162280.61
	DIF	161858.60	5619	55	161968.60	162333.47
Books	main effect	135900.65	4732	33	135966.65	136179.90
	DIF	135795.00	4732	55	135905.00	136260.41
Migration background	main effect	123135.37	4285	33	123201.37	123411.34
	DIF	123069.31	4285	55	123179.31	123529.27

5.4.5 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (GPCM; Muraki, 1982) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 6), ranging from 0.37 (item scg32910_c) to 2.00 (item scg3641s_c). The average discrimination parameter fell at 0.88. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 161,620.49, BIC = 161,972.10) as compared to the PCM model (AIC = 162,116.13, BIC = 162,328.42). 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 = 162,328.42, number of parameters = 32) fitted the data slightly better than a two-dimensional model (BIC = 162,342.96, number of parameters = 34). As the correlation between the two dimensions was r = .94 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 3 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 amount of not-reached items was rather high, indicating that the test was too long for the allocated testing time. Other types of missing responses were reasonably small.

The test had an acceptable reliability and distinguished well between test takers of average and low scientific literacy, but not as well for high performers. Very difficult items as well as very easy items were missing. Hence, test targeting was somewhat suboptimal. The test measured the scientific literacy of high-performing and very low performing students a little less accurately. This was depicted by the test's variance which, ideally, should be higher.

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 2PL model or as a correlation of the item score with total score) were acceptable. Different variables were used for testing measurement invariance across various subgroups. No considerable DIF became evident for any of these 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. Moreover, the high correlation between the two dimensions indicate that a unidimensional model describes the data reasonably well.

Summarizing the results, the test had good psychometric properties that facilitate the estimation of a unidimensional scientific literacy score.

7 Data in the Scientific Use file

7.1 Naming conventions

There are 23 items in the data set that are either scored as dichotomous variables (MC or SCR items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '_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 (scg3_sc1) including the respective standard error (scg3_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 four polytomous items (see section 5.3.1) 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). 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 Linking of competence scores

In starting cohort 2, the scientific literacy tests which were administered in grade 1 and 3, 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, I adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Following an anchor-group-design, all items from the grade 1 and the grade 3 scientific literacy tests were administered in an independent link sample – including students from grade 3 that were not part of starting cohort 2 – within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 2 across the two grades.

7.2.1 Samples

In starting cohort 2, a subsample of 5,330 students participated at both measurement occasions, in grade 1 and also in grade 3. Two participants had less than three valid item responses in one of the tests and therefore were excluded from the linking. Consequently, N = 5,328 students were used to link the two tests across both grades (see Fischer et al., 2016.). Moreover, an independent link sample of *N* = 286 students (50% female) from grade 3 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The test administered in grade 1 included 25 items, whereas the test administered in grade 3 included 23 items (see above). The science tests were administered in a random order. Half of the sample received the grade 1 test before working on the grade 3 test, whereas the other half received the grade 3 test before the grade 1 test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the science items in the same order.

7.2.3 Results

To examine whether the two tests administered in the link sample measured the same construct, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria favored the two-dimensional model, AIC = 16,264.67 BIC = 16,509.62, over the one-dimensional model, AIC =16,282.40, BIC = 16,520.04. However, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale – the average absolute residual correlation was M = .00 (SD = .07). This indicates that the scientific literacy tests administered in grades 1 and 3 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 2 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 9.

Table 9

Differential Item Functioning Analyses between the Starting Cohort and the Link Sample

	Grade 1				Grade 3			
	Item	**Δσ**	***SE*Δσ**	***F***	**Item**	**Δσ**	***SE*Δσ**	***F***
1.	scg10820_c	0.16	0.21	0.57	scg30109_c	−0.08	0.14	0.35
2.	scg10840_c	0.30	0.17	3.05	scg33510_c	−0.10	0.14	0.47
3.	scg11510_c	−0.16	0.13	1.43	scg3181s_c	0.33	0.13	6.31
4.	scg10650_c	0.54	0.37	2.16	scg37110_c	−0.39	0.14	7.30
5.	scg16510_c	0.24	0.15	2.70	scg31010_c	−0.14	0.14	1.00
6.	scg1652s_c	−0.50	0.10	26.41	scg36710_c	0.18	0.13	1.95
7.	scg16110_c	−0.54	0.16	11.19	scg3131s_c	−0.25	0.14	3.44
8.	scg1091s_c	0.00	0.15	0.00	scg34010_c	0.11	0.13	0.75
9.	scg10920_c	−0.10	0.17	0.32	scg32220_c	−0.30	0.14	4.93
10.	scg1011s_c	−0.53	0.14	14.03	scg33710_c	0.11	0.13	0.73
11.	scg10120_c	−0.52	0.14	13.16	scg36210_c	−0.07	0.16	0.17
12.	scg11210_c	0.43	0.21	4.23	scg36920_c	0.10	0.13	0.59
13.	scg11110_c	0.39	0.25	2.36	scg32620_c	−0.28	0.13	4.60
14.	scg11130_c	−0.04	0.18	0.05	scg31510_c	0.15	0.13	1.31
15.	scg16530_c	−0.24	0.13	3.24	scg30310_c	−0.07	0.13	0.28
16.	scg16020_c	−0.42	0.14	8.40	scg3641s_c	−0.02	0.13	0.03
17.	scg16030_c	0.22	0.17	1.67	scg3052s_c	0.69	0.16	19.03
18.	scg11610_c	−0.01	0.15	0.00	scg37410_c	0.01	0.14	0.00
19.	scg11710_c	−0.14	0.14	1.14	scg33310_c	−0.16	0.14	1.24
20.	scg10310_c	−0.64	0.14	22.45	scg3091s_c	0.16	0.14	1.30
21.	scg10520_c	−0.04	0.20	0.04	scg33610_c	0.03	0.15	0.05
22.	scg16310_c	0.84	0.28	9.12	scg32910_c	−0.02	0.16	0.01
23.	scg16220_c	−0.11	0.17	0.42				
24.	scg11440_c	0.22	0.15	2.12				
25.	scg10410_c	0.66	0.21	9.38				

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 1 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is $F_{.0154}$ (1, 5,614) = 121.59. A non−significant test indicates measurement invariance.

Measurement invariance for grade 1 and grade 3 showed no item with *F*-statistics exceeding the critical value of $F_{.0154}(1, 5,614) = 121.59$. Thus, no item had to be excluded from the estimation of the correction term.

Moreover, analyses of differential item functioning between the link sample and starting cohort 2 showed in grade 1 for 24 items of the test no DIF greater than 0.40 (difference in logits: *Min* = −0.35, *Max* = 0.32). But, one item (scg16310_c) showed a DIF greater than 0.40. This item was therefore excluded from the estimation of the correction term. For grade 3 (difference in logits: *Min* = −0.19, *Max* = 0.30) there were no items with a DIF greater than 0.40. Therefore, the scientific literacy tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 1.225$ (with a link error of 0.08). This correction term was subsequently added to each difficulty parameter estimated in grade 3 (see Table 5) to derive the linked item parameters. Moreover, the correction term calculated for the linking between kindergarten and grade 1 ($c = 1.445$, link error = 0.10) was also added to each difficulty parameter estimated in grade 3 (see Kähler, in press).

7.3 Scientific literacy scores

In the SUF manifest scientific literacy scores are provided in the form of two different WLEs, "scg3_sc1" and "scg3_sc1u", including their respective standard error, "scg3_sc2" and "scg3_sc2u".

For "scg3_sc1u", person abilities were estimated using the linked item difficulty parameters. As a result the WLE scores provided in "scg3_sc1u" can be used for longitudinal comparisons between kindergarten and 3. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in "scg3_sc1" are not linked to the underlying reference scale of kindergarten. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. 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).

References

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

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.). *Research Outcomes of the PISA Research Conference 2009* (pp. 199–214). New York, NY: Springer.

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

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

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

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.

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, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H., & 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, Germany: Waxmann.

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*, 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, Germany: VS Verlag für Sozialwissenschaften.

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

Appendix

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

```
Title G3 Science analysis, Partial Credit Model;
data filename.dat;
format id 1–7 responses 8–29;

labels << filename_with_labels.txt;

codes 0,1,2,3;
score (0,1)         (0,1)            !item (1,2,4–6,8–15,18,19,21–22);
score (0,1,2)       (0,0.5,1)        !item (17);
score (0,1,2,3)     (0,0.5,1,1.5)    !item (3,7,16,20);

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 test items to the content and process related components and to contexts

Variable name	Position in the test	Component	Context
scg30109_c	1	KOS	Health
scg33510_c	2	KOS	Environment
scg3181s_c	3	KOS	Health
scg37110_c	4	KAS	Health
scg31010_c	5	KOS	Environment
scg36710_c	6	KAS	Technology
scg3131s_c	7	KOS	Health
scg34010_c	8	KOS	Technology
scg32220_c	9	KOS	Environment
scg33710_c	10	KOS	Technology
scg36210_c	11	KAS	Health
scg36920_c	12	KAS	Technology
scg32620_c	13	KOS	Environment
scg31510_c	14	KOS	Health
scg30310_c	15	KOS	Health
scg3641s_c	16	KAS	Environment
scg30520_c	17	KOS	Technology
scg37410_c	18	KAS	Environment
scg33310_c	19	KOS	Environment
scg3091s_c	20	KOS	Environment
scg33610_c	21	KOS	Technology
scg32910_c	22	KOS	Environment
scg31610_c	23	KOS	Health

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