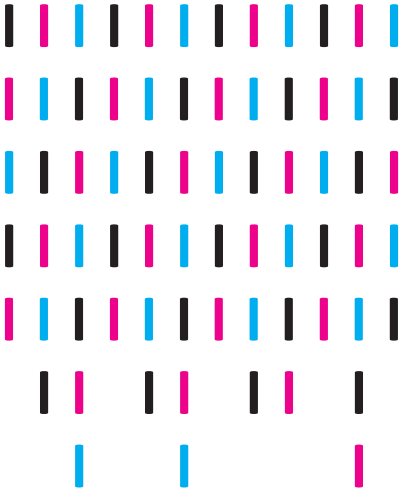

NEPS*SURVEY PAPERS*

Inga Hahn and Jana Kähler

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 11

NEPS*Survey Paper* No. 93
Bamberg, April 2022

LIfBi
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 3 for Grade 11

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

Email address of the lead author:

hahn@leibniz-ipn.de

Bibliographic data:

Hahn, I. & Kähler, J. (2022). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 11* (NEPS Survey Paper No. 93). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP93:1.0

NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 11

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 11 of Starting Cohort 3. The scientific literacy test contained 25 items with different response formats representing different contexts as well as different areas of knowledge. The test was administered to 1,930 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 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. Additionally, the design and results of the linking study for the competence scores in grades 9 and 11 are presented.

Keywords

scientific literacy, 11th grade, linking grade 9 and 11, differential item functioning, item response theory, scaling, scientific use file

Content

1 Introduction 4
2 Testing Scientific Literacy 4
3 Data 5
3.1 The design of the study 5
3.2 Sample 6
4 Analyses 6
4.1 Missing responses 7
4.2 Scaling model 7
4.3 Checking the quality of the test 7
4.4 Software 9
5 Results 9
5.1 Descriptive statistics of the responses 9
5.2 Missing Responses 9
5.2.1 Missing responses per person 9
5.2.2 Missing responses per item 11
5.3 Parameter estimates 13
5.3.1 Item parameters 13
5.3.2 Person parameters 13
5.3.3 Test targeting and reliability 13
5.4 Quality of the test 18
5.4.1 Fit of the subtasks of complex multiple-choice items 18
5.4.2 Distractor analyses 18
5.4.3 Item fit 19
5.4.4 Differential item functioning 19
5.4.5 Rasch-homogeneity 24
5.4.6 Unidimensionality of the test 25
6 Discussion 25
7 Data in the Scientific Use file 26
7.1 Naming conventions 26
7.2 Linking of competence scores 26
7.2.1 Samples 26
7.2.2 The design of the link study 27
7.2.3 Correcting for a change in study design 27
7.2.4 Results 27
7.3 Scientific literacy scores 28
8 References 31

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the lifespan (Blossfeld & Roßbach, 2019). These include, among other things, 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 11 of Starting Cohort 3. First, the main concepts of the scientific literacy test are introduced. Then, the scientific literacy data of Starting Cohort 3 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 the 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 or more items. Each item refers to one context-component-combination.

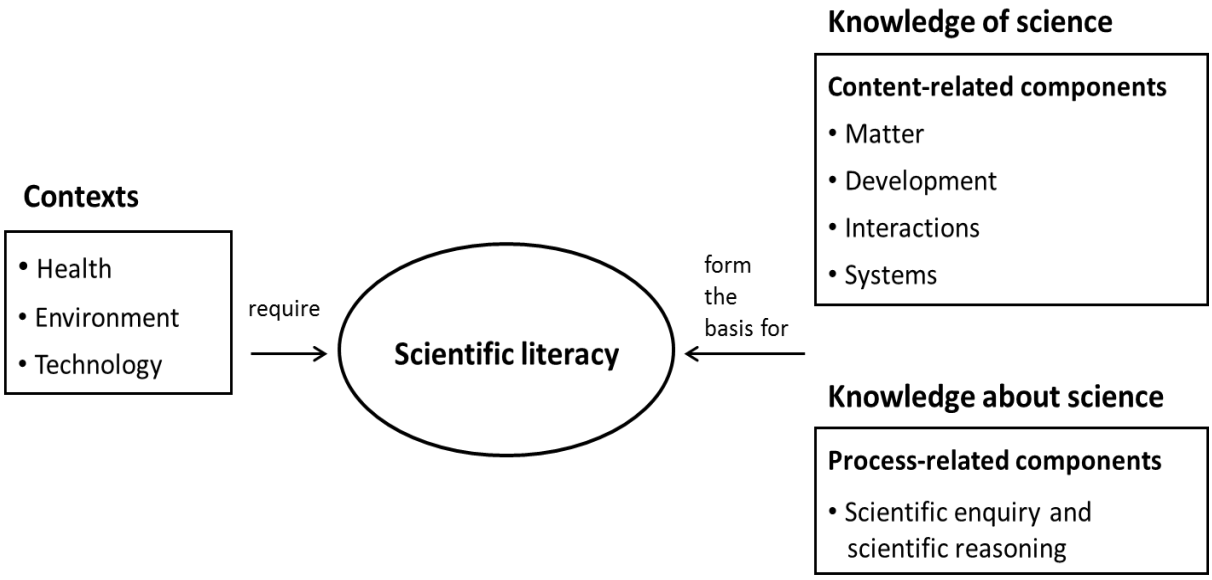

Figure 1. Assessment framework for scientific literacy (Hahn et al., 2013).

In the scientific literacy test for Grade 11 of Starting Cohort 3 (fifth grade), 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

Since scientific literacy was the only competency tested in this study, there was only one testing group who received the science test first and afterwards completed their questionnaire. The test time for the scientific literacy test was 29 minutes, with one additional minute for the procedural metacognition item. There was no multi-matrix design regarding the choice and order of the items within a test. All students got the same test items in the same test order.

The scientific literacy test in Grade 11 originally consisted of 25 items. The characteristics of these 25 items are depicted in Table 1. Table 2 is concerned with the response format whereas Table 3 shows how the items cover the different contents and components of the science framework (see Hahn et al., 2013). One of the 25 items had to be removed from the final analysis presented in this paper due to insufficient item quality. See Appendix B for the detailed assignment of the test items content and process-related components, and to contexts)

Table 1: Classification of Items into Knowledge Domains

Knowledge domains	Number of Items
Knowledge of Science (KOS)	18
Knowledge about Science (KAS)	7
Total number of items	25

Table 2: Number of Items by Different Contexts

Context	Number of Items
Health	7
Environment	8
Technology	10
Total number of items	25

Table 3: Number of Items by Response Formats

Response format	Number of Items
Simple Multiple-Choice	13
Complex Multiple-Choice (True-false items)	12
Total number of items	25

3.2 Sample

A total of 1,930 individuals received the scientific literacy test. The analyses presented in this paper are based on this sample (52.3% 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 25 items (including all subtasks for the polytomous items) were included in the analyses. One item (scs56320_sc3g11_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.

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 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.

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. When categories of the polytomous variables had less than $N = 200$, the categories were collapsed to avoid any possible estimation problems. This usually occurred for the lower categories of polytomous items. For seven of the twelve 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 students. 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), migration background, school type. 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 scale.

Moreover, we examined whether the residuals of the unidimensional 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 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

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 for all items. The percentage of persons correctly responding to an item (relative to all valid responses) ranged from 19.0% to 65.4% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 9.7% and 38.3%. 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: 87.2% of the respondents did not have any invalid response at all. Less than 0.6% had more than three invalid responses.

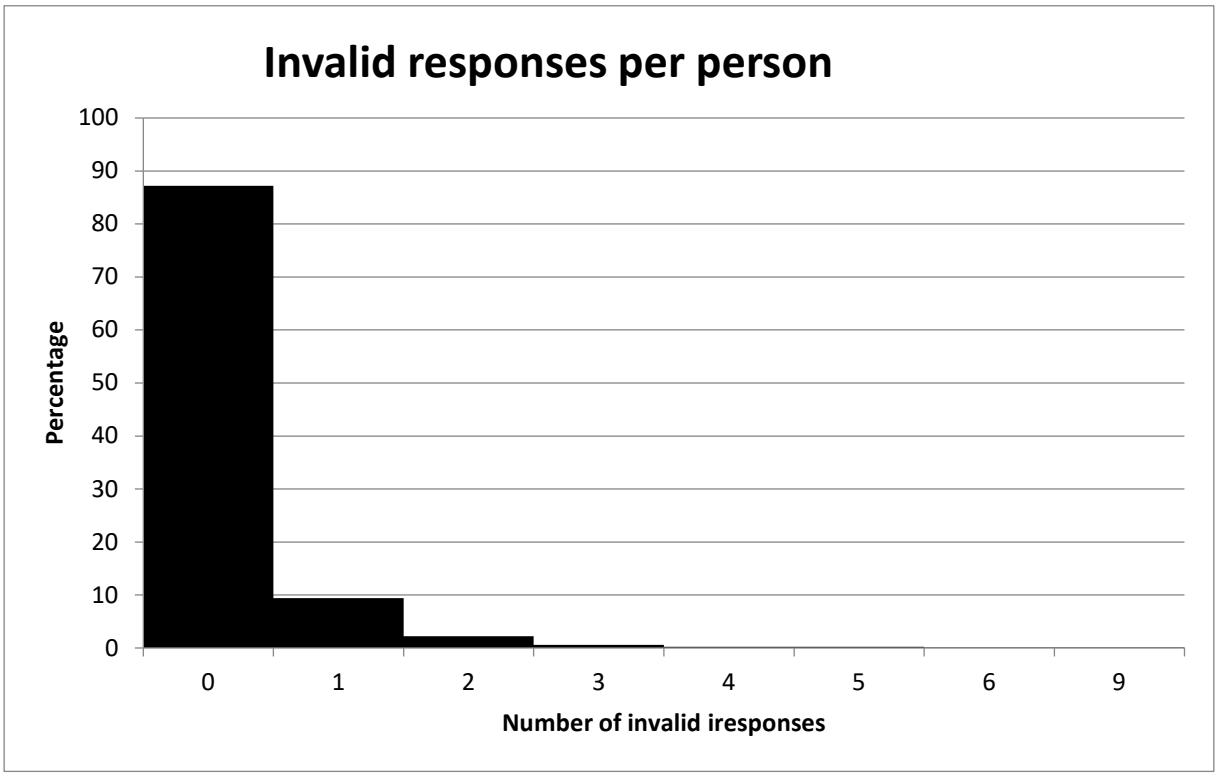

Figure 2. Number of invalid responses per person.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 90.1% did not skip any item, and less than 1.0% omitted more than three items.

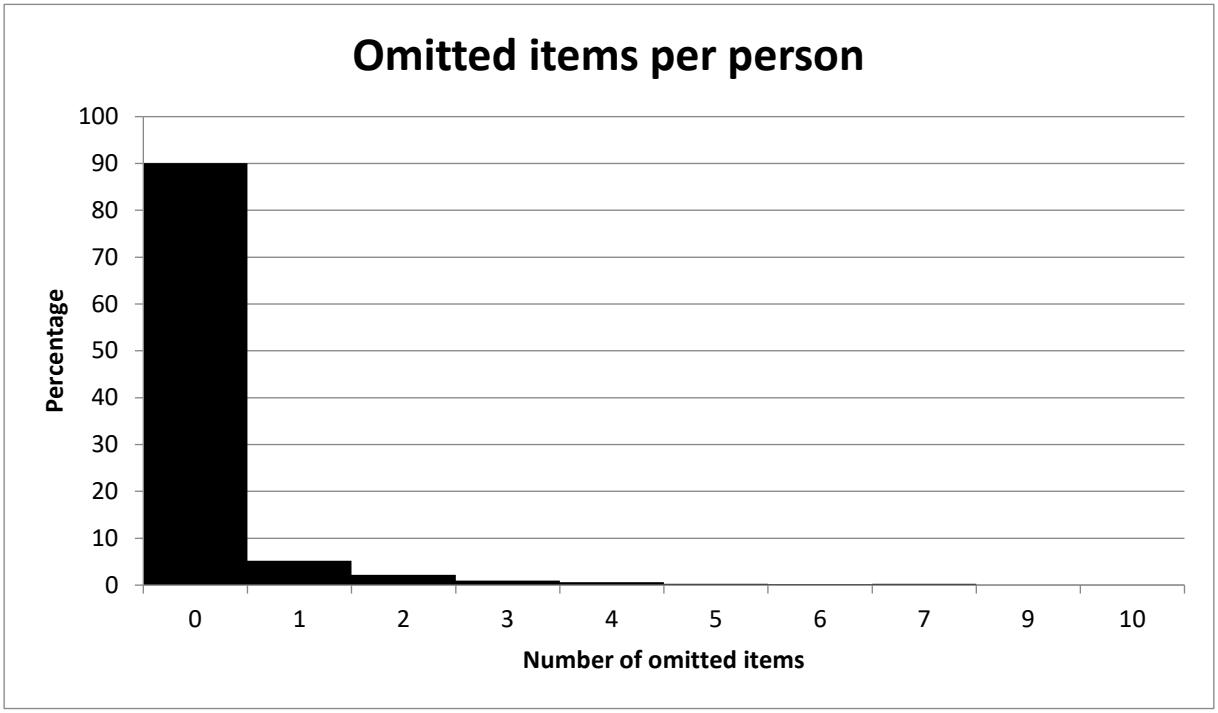

Figure 3. Number of omitted responses per person.

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 higher than expected. However, 58.9% of the respondents were able to finish the test within the allocated time limit (Figure 4). Less than 0.2% did not finish more than half of the items.

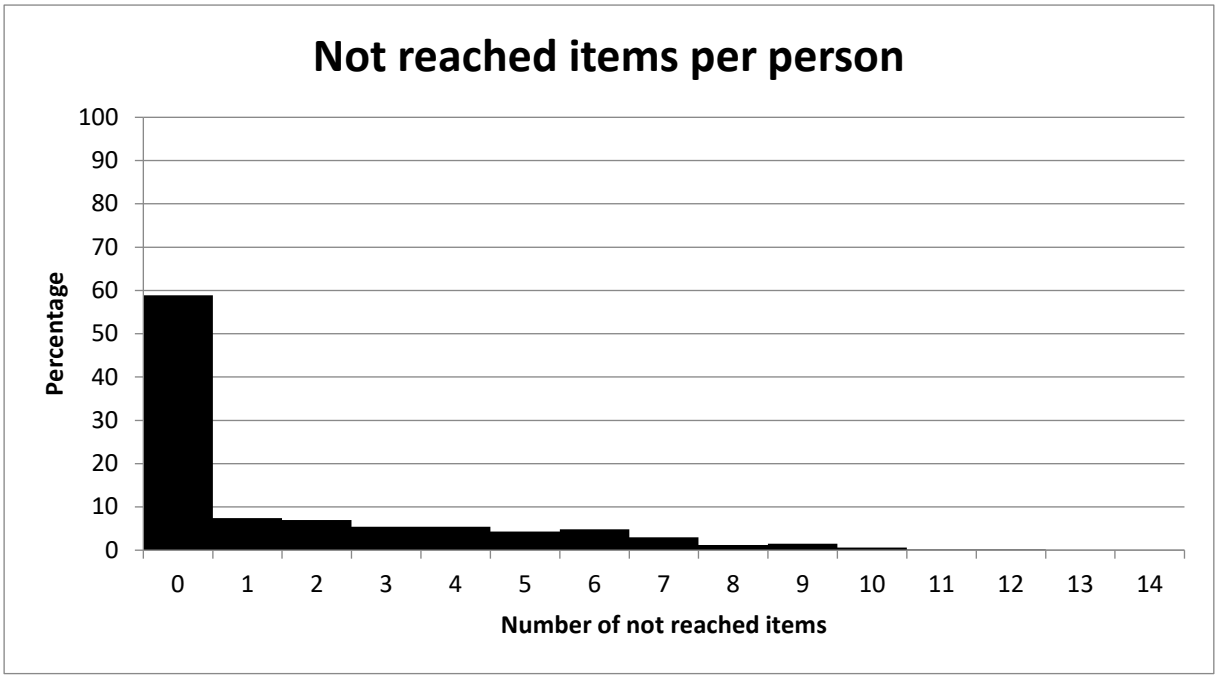

Figure 4. Number of not reached items per person.

The total number of missing responses, aggregated over invalid, omitted and not-reached missing responses, is illustrated in Figure 5. About 48.1% of the respondents had no missing response at all and about 19.3% of the participants had five or more missing responses. Overall, the amount of invalid and omitted items is small, whereas the amount of missing responses due to not-reached items could be smaller.

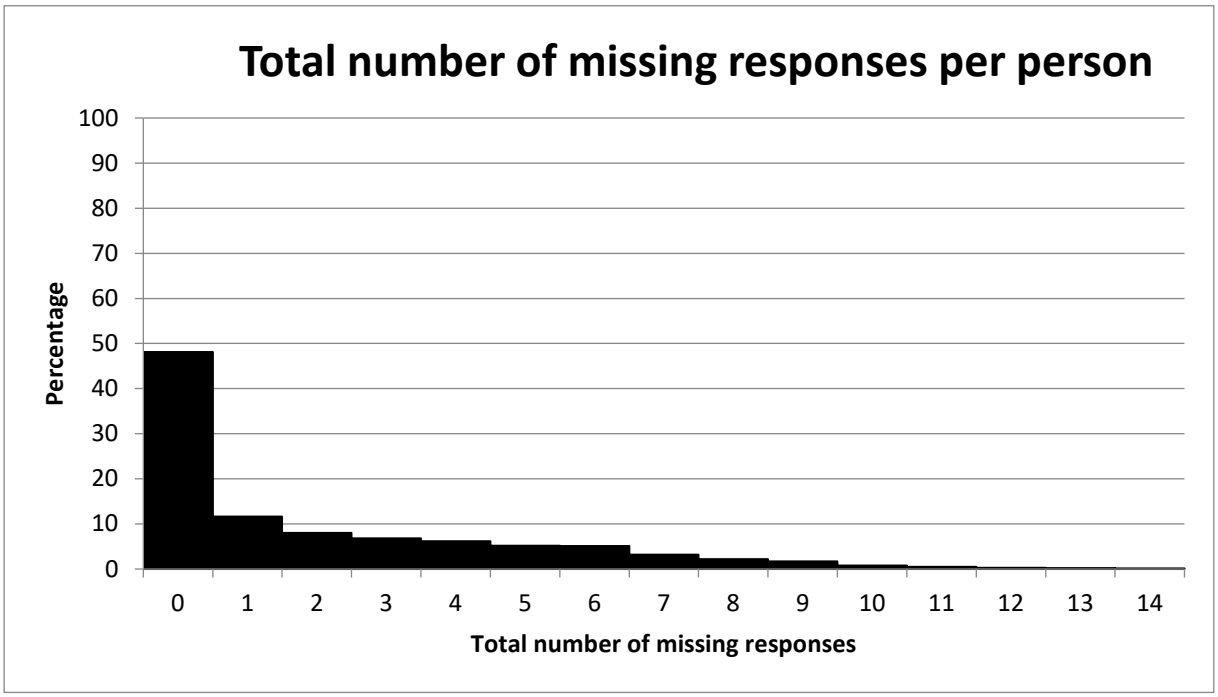

Figure 5. Total number of missing responses per person.

5.2.2 Missing responses per item

Table 4 provides information on the occurrence of different kinds of missing responses per item. The omission rates were rather low, varying across items between 0.0 % and 2.5%. Overall, the missing rates correlated with the item difficulties at about r = .037 (p = .862), indicating that test-takers did not especially miss difficult items. Generally, the percentage of invalid responses per item (column 6) was rather low with the maximum rate being 2.5%. With an item's progressing position in the test, the number of persons that did not reach the item (column 4) rose to a comparatively high amount of 41.4%.

Table 4: Percentage of Missing Values per Item

Item	Position in the test	Number of valid responses	Not reached items (%)	Omitted items (%)	Invalid responses (%)
scg116420_sc3g11_c	1	1923	0.0	0.2	0.2
scg110620_sc3g11_c	2	1896	0.0	1.6	0.2
scg110630_sc3g11_c	3	1912	0.0	0.7	0.2
scg11012s_sc3g11_c	4	1870	0.0	0.6	2.5
scg11083s_sc3g11_c	5	1897	0.0	0.4	1.3
scg110720_sc3g11_c	6	1922	0.0	0.4	0.1
scg11032s_sc3g11_c	7	1908	0.0	0.2	1.5
scg110330_sc3g11_c	8	1914	0.0	0.3	0.6
scg116510_sc3g11_c	9	1895	0.0	0.2	1.6
scg11652s_sc3g11_c	10	1898	0.0	0.1	1.6
scg110510_sc3g11_c	12	1907	0.1	0.8	0.3
scg110520_sc3g11_c	13	1916	0.1	0.5	0.2
scg110540_sc3g11_c	14	1900	0.3	1.1	0.1
scg11123s_sc3g11_c	15	1887	0.5	0.1	1.7
scg11102s_sc3g11_c	16	1877	1.1	0.7	0.9
scg11021s_sc3g11_c	17	1842	2.6	1.0	1.0
scg11022s_sc3g11_c	18	1811	3.8	1.9	0.4
scg11112s_sc3g11_c	19	1775	6.8	0.5	0.7
scg116210_sc3g11_c	20	1647	11.7	2.5	0.5
scg11622s_sc3g11_c	21	1562	16.0	1.4	1.7
scg116320_sc3g11_c	22	1479	21.3	1.9	0.1
scg110930_sc3g11_c	23	1363	26.7	2.5	0.2
scs5131s_sc3g11_c	24	1241	33.7	1.2	0.7
scs5132s_sc3g11_c	25	1126	41.4	0.0	0.6

Note. The item on position 11 was excluded from the scaling procedure due to an inferior model fit.

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 12.8% and 63.9% with an average of 43.8% (*SD* = 17.2) correct responses.

The estimated item difficulties (for dichotomous items, MC items) and location parameters (for polytomous variables, CMC items) are given in Table 5. The step parameters (for polytomous variables) are depicted in Table 6.

For six of the CMC items (scg11652s_sc3g11_c, scg11123s_sc3g11_c, scg11102s_sc3g11_c, scg11021s_sc3g11_c, scg11022s_sc3g11_c, scg11622s_sc3g11_c) the two lowest categories were collapsed, thus, these items were scaled using a scoring of 0, 0.5, 1, and 1.5. For the other six CMC items (scg11012s_sc3g11_c, scg11083s_sc3g11_c, scg11032s_sc3g11_c, scg11112s_sc3g11_c, scs5131s_sc3g11_c, scs5132s_sc3g11_c) the three lowest categories were collapsed, thus, these items were 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 −1.25 (scg11652s_sc3g11_c) and 2.03 (scg11022s_sc3g11_c). In total, the estimated item difficulties had a mean of 0.02 (*SD* = 0.85). Due to the large sample size, the standard errors of the estimated item difficulties were very small ($SE(\beta) \leq 0.09$).

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 0.544, indicating a somewhat limited variability between subjects which can be explained by the fact that the sample only consisted of students from the German "Gymnasium". The reliability of the test (EAP/PV reliability = .704; WLE reliability = .682) was acceptable. Although the items covered a wide range of the ability distribution, few items were covering the upper area around the person ability of 1. As a consequence, person ability in medium and low ability regions will be measured relative precisely, whereas medium high ability estimates will have larger standard errors of measurement.

Table 5: Item parameters

No.	Item	Correct (%)	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discrimination (GPCM)	Q3
1	scg116420_sc3g11_c	60.4	−0.48	0.05	1.05	2.8	0.31	0.48	0.05
2	scg110620_sc3g11_c	49.6	−0.02	0.05	0.97	−2.2	0.44	0.92	0.08
3	scg110630_sc3g11_c	51.7	−0.10	0.05	0.99	−0.6	0.30	0.82	0.08
4	scg11012s_sc3g11_c	45.0	0.16	0.05	0.94	−4.2	0.49	1.18	0.06
5	scg11083s_sc3g11_c	n.a.	−1.04	0.06	0.98	−1.0	0.43	0.87	0.07
6	scg110720_sc3g11_c	62.1	−0.57	0.05	0.98	−0.9	0.40	0.82	0.08
7	scg11032s_sc3g11_c	n.a.	−1.25	0.07	1.00	0.0	0.32	0.76	0.07
8	scg110330_sc3g11_c	59.0	−0.43	0.05	0.97	−2.1	0.44	0.98	0.08
9	scg116510_sc3g11_c	63.9	−0.69	0.05	0.99	−0.6	0.39	0.78	0.08
10	scg11652s_sc3g11_c	n.a.	−0.13	0.06	0.99	−0.3	0.37	0.76	0.10
11	scg110510_sc3g11_c	57.6	−0.38	0.05	0.99	−0.4	0.40	0.75	0.09
12	scg110520_sc3g11_c	56.2	−0.30	0.05	1.04	2.4	0.32	0.49	0.08
13	scg110540_sc3g11_c	19.1	1.58	0.06	1.01	0.4	0.30	0.64	0.07
14	scg11123s_sc3g11_c	n.a.	−0.20	0.06	1.07	2.5	0.33	0.43	0.14

15	scg11102s_sc3g11_c	26.1	1.12	0.06	0.98	−0.6	0.39	0.89	0.07
16	scg11021s_sc3g11_c	27.9	0.99	0.06	0.99	−0.3	0.38	0.80	0.05
17	scg11022s_sc3g11_c	12.8	2.03	0.06	1.00	0.1	0.27	0.67	0.10
18	scg11112s_sc3g11_c	60.2	−0.72	0.07	1.08	4.0	0.23	0.28	0.08
19	scg116210_sc3g11_c	44.8	−0.12	0.06	0.99	−0.5	0.40	0.74	0.08
20	scg11622s_sc3g11_c	15.4	1.59	0.06	1.05	1.3	0.23	0.40	0.08
21	scg116320_sc3g11_c	32.0	0.36	0.07	0.97	−1.8	0.43	0.93	0.10
22	scg110930_sc3g11_c	44.8	−0.66	0.06	0.96	−1.9	0.45	1.03	0.08
23	scs5131s_sc3g11_c	n.a.	0.02	0.08	0.98	−0.7	0.36	0.98	0.08
24	scs5132s_sc3g11_c	n.a.	−0.37	0.09	0.99	−0.4	0.44	0.80	0.14

Note. *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ. r_{it} = point-biserial correlation of the correct response. Percent correct scores are not informative for polytomous CMC item scores. These are 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).

Table 6: Step parameters for the CMC items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3 (*SE*)	Step 4
scg11083s_sc3g11_c	−0.59 (0.07)	0.33 (0.07)	0.26	
scg11032s_sc3g11_c	0.04 (0.54)	−0.04		
scg11652s_sc3g11_c	−1.27 (0.06)	0.40 (0.06)	0.86	
scg11123s_sc3g11_c	−1.08 (0.08)	−0.55 (0.07)	0.82 (0.07)	0.80
scs5131s_sc3g11_c	0.42 (0.06)	0.42		
scs5132s_sc3g11_c	−1.75 (0.11)	0.37 (0.09)	−0.15 (0.09)	1.53

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

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

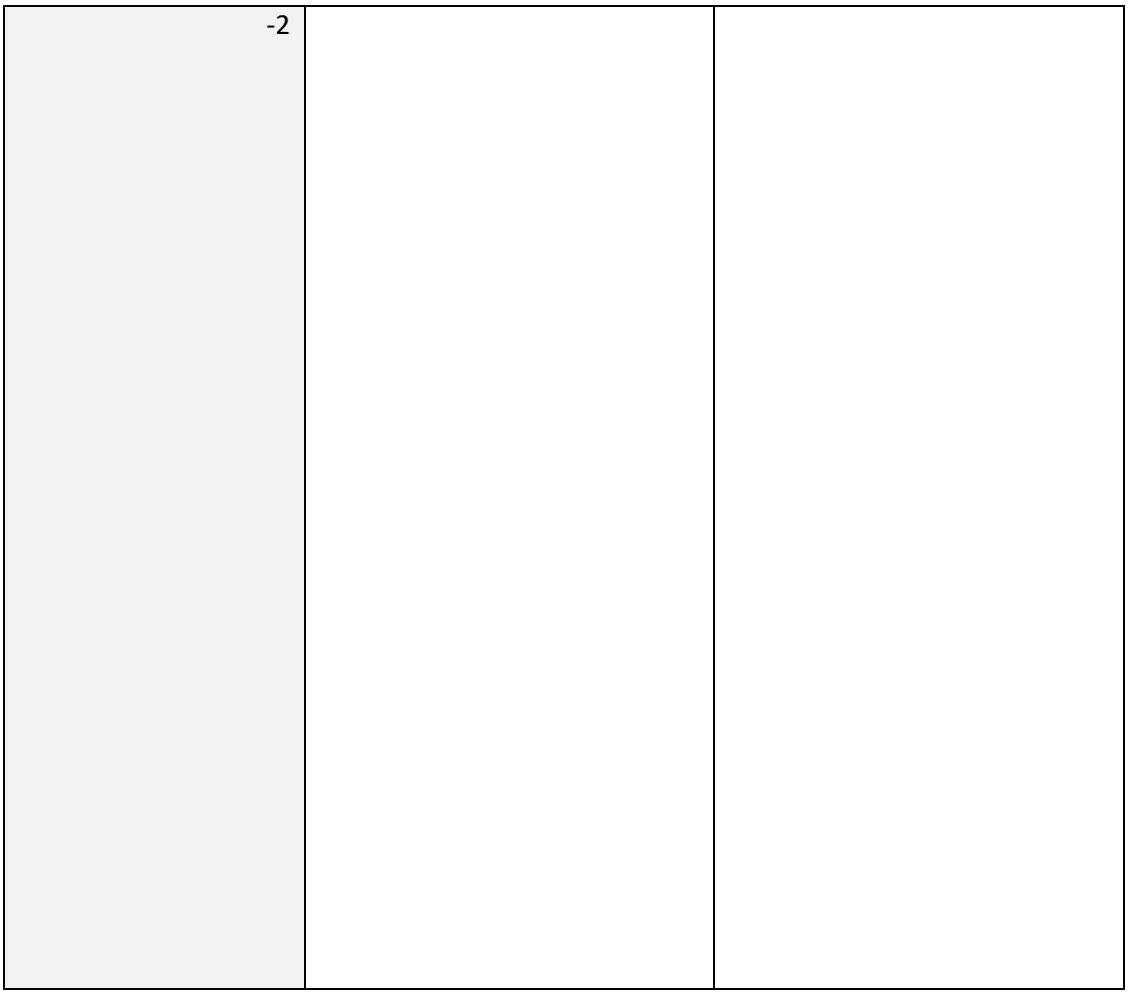

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 11.1 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 5).

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 61 items. The percentage of a correct response ranged from 19.4% to 97.0% across all items (*Mdn* = 66.2%). 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 0.89 to 1.10, the respective *t*-value from −8.8 to 6.8, 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. Most distractors had a point-biserial

correlation with the total scores below zero. However, there were some CMC Items which had distractors showing noticeable positive correlations (scg11652s_sc3g11_c, scg11622s_sc3g11_c, scs5131s_sc3g11_c, scs5132s_sc3g11_c). Since these items were already part of the Grade 11 test of Starting Cohort 4 we had to score them according to this starting cohort in order to keep them comparable. We kept them in the analyses because there were no noticeable deviations of the empirically estimated probabilities from the model-implied item characteristic curves and there were no anomalies in item fit and differential item functioning.

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 0.94 (item scg11012s_sc3g11_c) to 1.08 (scg11112s_sc3g11_c). No item exhibited a *t*-value of the WMNSQ greater than 6. The highest t-value was 4.0 (scg11112s_sc3g11_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 .11 (item scg11112s_sc3g11_c) to .39 (scg11102s_sc3g11_c) and had a mean of .25. 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), migration background and school type (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. Also, Table 8 shows the main effect for the examined subgroups (inclusive Cohen's d).

Table 7: Differential item functioning (differences between difficulties)

Item	Gender	Books			Migration status			School type
	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing	Other vs. Gymnasium
scg116420_sc3g11_c	0.288	−0.004	0.100	0.106	−0.172	−0.056	0.116	0.050
scg110620_sc3g11_c	0.042	0.038	0.084	0.048	0.102	0.108	0.004	−0.078
scg110630_sc3g11_c	0.150	0.006	0.034	0.030	−0.196	−0.080	0.116	−0.012
scg11012s_sc3g11_c	−0.216	0.314	0.374	0.058	−0.218	0.068	0.288	−0.032
scg11083s_sc3g11_c	−0.342	0.026	0.018	−0.002	−0.130	0.068	0.194	0.144
scg110720_sc3g11_c	0.310	0.090	−0.084	−0.172	0.110	0.016	−0.096	0.106
scg11032s_sc3g11_c	−0.150	0.076	−0.008	−0.086	−0.004	−0.120	−0.116	−0.200
scg110330_sc3g11_c	−0.040	0.144	0.064	−0.078	−0.034	−0.092	−0.060	0.200
scg116510_sc3g11_c	0.504	0.118	−0.008	−0.124	0.100	−0.152	−0.254	0.192
scg11652s_sc3g11_c	−0.290	0.192	0.198	0.020	−0.246	−0.106	0.136	0.112
scg110510_sc3g11_c	−0.072	0.080	0.270	0.192	−0.172	0.066	0.240	−0.326
scg110520_sc3g11_c	0.158	−0.196	−0.214	−0.014	0.128	0.112	−0.018	0.348

scg110540_sc3g11_c	0.146	0.278	0.310	0.032	−0.084	0.132	0.218	0.634
scg11123s_sc3g11_c	0.106	0.030	0.052	0.020	0.502	0.052	−0.468	−0.274
scg11102s_sc3g11_c	−0.050	−0.138	−0.154	−0.014	−0.142	−0.062	0.080	−0.066
scg11021s_sc3g11_c	−0.526	−0.138	−0.180	−0.038	0.014	−0.058	−0.074	−0.074
scg11022s_sc3g11_c	−0.448	−0.328	−0.276	0.056	−0.030	0.086	0.116	0.086
scg11112s_sc3g11_c	0.388	−0.330	−0.258	0.076	−0.002	0.114	0.116	−0.512
scg116210_sc3g11_c	0.386	0.066	−0.140	−0.204	−0.174	−0.108	0.066	0.180
scg11622s_sc3g11_c	0.010	−0.426	−0.404	0.028	0.462	−0.098	−0.566	−0.356
scg116320_sc3g11_c	−0.420	−0.180	−0.220	−0.036	0.046	0.000	−0.048	−0.190
scg110930_sc3g11_c	0.286	−0.020	0.064	0.088	0.024	0.084	0.058	0.306
scs5131s_sc3g11_c	−0.170	−0.202	−0.324	−0.128	0.146	−0.188	−0.332	−0.638
scs5132s_sc3g11_c	−0.582	0.048	0.168	0.106	0.106	0.092	−0.014	0.094

Gender

The sample included 921 (47.7%) male test-takers (coded 0) and 1,009 (52.3%) female test-takers (coded 1). On average, male students had slightly higher scores in scientific literacy than female students (main effect = 0.374 logits, Cohen's *d* = 0.524). However, the items showed no considerable DIF. The highest difference in difficulties between the two groups was −0.582 logits.

Books

The number of books at home was used as a proxy for socioeconomic status. There were 307(15.9%) test takers with 0 to 100 books at home (coded 0), 1,088 (56.4%) test takers with more than 100 books at home (coded 1), and 535 (27.7%) 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 performed showed lower scientific literacy scores than participants with more than 100 books (main effect = −0.338 logits, Cohen's *d* = −0.477). Participants with up to 100 books performed lower than participants without a valid response on the variable 'books at home' (main effect = −0.086 logits, Cohen's *d* = −0.115). Participants with more than 100 books at home performed better than participants without a valid response on the variable 'books at home' (main effect = 0.254 logits, Cohen's *d* = 0.351). There was no considerable DIF comparing participants with many or fewer books (highest DIF = -0.426). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to −0.404 logits (Participants with 100 or fewer books at home vs. Participants without a valid response).

Migration background

There were 1,134 (58.8%) participants without a migration background (coded 0) and 223 (11.6%) participants with a migration background (coded 1; for 5.0% of the students neither their mother nor their father was born in Germany and for 6.5% of the participants only one of their parents was born abroad). A total of 573 (29.7%) students could not be allocated to either group (coded 9). These groups were used for investigating DIF. There was a small 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 = 0.188 logits, Cohen's *d* = 0.263) and also higher scores than students with an unknown background on migration (main effect = 0.178 logits, Cohen's *d* = 0.245). Furthermore, students with a migration background scored slightly lower than those with an unknown background on migration (main effect = −0.012 logits, Cohen's *d* = −0.016). There was no considerable DIF comparing participants with and without a migration background (highest DIF = 0.502). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to -0.566 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 1,752 (90.8%) students visiting "Gymnasium" (coded 1), and 178 (9.2%) students from lower schools (coded 0), such as "Hauptschule" or "Realschule". On average, students visiting "Gymnasium" had distinctly higher scores in scientific literacy than students from other school types (main effect = −0.424 logits, Cohen's d = −0.583). There were two items with a considerable DIF of 0.634 (item scg110540_sc3g11_c) and -0.638 (item scs5131s_sc3g11_c). Since both items didn`t display problems in other areas they remained in the analysis.

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

Variables	Subgroups	Main effect	Cohen`s d
Gender	Male (0) Female (1)	0.374	0.524
Books	0 to 100 books at home (0) More than 100 books at home (1)	−0.338	−0.477
	0 to 100 books at home (0) Invalid response (9)	−0.086	−0.115
	More than 100 books at home (1) Invalid response (9)	0.254	0.351
Migration background	Without migration background (0) With migration background (1)	0.188	0.263
	Without migration background (0) Invalid response (9)	0.178	0.245
	With migration background (1) Invalid response (9)	−0.012	−0.016
School type	Other school type (0) Gymnasium (1)	−0.424	-0.583

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 (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 one DIF variable (gender). For the variables books, migration background and school type, the AIC favored the model which only allows 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 all variables.

Table 9: Comparison of models with and without DIF

DIF variable	Model	Deviance	N	Number of parameters	AIC	BIC
Gender	main effect	63743.90	1930	38	63819.90	64031.38
	DIF	63563.02	1930	62	63687.02	64032.07
Books	main effect	46169.64	1395	38	46245.64	46444.79
	DIF	46139.72	1395	62	46263.72	46588.64
Migration background	main effect	45032.83	1357	38	45108.83	45306.92
	DIF	45003.01	1357	62	45127.01	45450.22
School type	main effect	63795.86	1930	38	63871.86	64083.34
	DIF	63754.71	1930	62	63878.71	64223.76

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 0.28 (item scg11112s_sc3g11_c) to 1.18 (item scg11012s_sc3g11_c). The average discrimination parameter fell at 0.76. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 63742.66, BIC = 63819.79) as compared to the PCM model (AIC = 63907.22, BIC = 63954.78). 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 = 63,954.78, number of parameters = 37) fitted the data slightly better than the two-dimensional model (BIC = 63,955.48, number of parameters = 39). Additionally, the correlation between the two dimensions was r = .95 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 11 of Starting Cohort 3 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.

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. There could have been more items covering the medium upper area. Hence, test targeting could have been better. The test measured the scientific literacy of high-performing students a little less accurately. This was depicted by the test's variance which, ideally, should be higher but which was presumably limited due to the fact that the sample only consisted of students from the German "Gymnasium".

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. No considerable DIF became evident for any of these variables, indicating that the test was mainly 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 indicates 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 25 items in the data set that are either scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '_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 (scg11_sc1) including the respective standard error (scg11_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 two CMC items (scg11083s_sc3g11_c, scg11652s_sc3g11_c), and the three lowest categories of three CMC items (scg11032s_sc3g11_c, scs5131s_sc3g11_c; see section 5.3.1). In the scaling model, the collapsed polytomous item was scored in steps of 0.0, 0.5, 1.0, and 1.5 (denoting the highest) for items with the two lowest categories collapsed, and steps of 0.0, 0.5, and 1.0 (denoting the highest) for items with the three lowest categories collapsed. Six of the CMC items (scg11012s_sc3g11_c, scg11102s_sc3g11_c, scg11021s_sc3g11_c, scg11022s_sc3g11_c, scg11112s_sc3g11_c, scg11622s_sc3g11_c) were treated as MC-items using a scoring of 0 and 1 (right answer on all subtasks). 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 3, the scientific literacy tests which were administered in Grades 9 and 11, 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 competencies across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Following an anchor-group-design, all items from the Grade 9 and the Grade 11 scientific literacy tests were administered in an independent link sample – including students from Grade 11 that were not part of Starting Cohort 3 – within a single measurement occasion. These responses were used to link the two tests administered in Starting Cohort 3 across the two grades.

7.2.1 Samples

In Starting Cohort 3, a subsample of 1,860 students participated at both measurement occasions, in Grade 9 and also in Grade 11. Consequently, *N* = 1,860 students were used to link the two tests across both grades (Fischer et al., 2016). Moreover, an independent link

sample of *N* = 178 students (60.7% female) from Grade 11 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The test administered in the linking sample for Grade 9 included 28 items from the easy test version (see Kähler, 2020). Item scg9611s_sc3g11_c was excluded from the linking since it was also excluded from the main study analyses due to insufficient item quality. The test administered for Grade 11 included 25 items. Item scs56320_sc3g11_c had to be excluded from the analyses due to insufficient item quality. Thus, this item was also excluded from the linking. The science tests were administered in random order. Half of the sample received the Grade 9 test before working on the Grade 11 test, whereas the other half received the Grade 11 test before the Grade 9 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 Correcting for a change in study design

The design of the link study was identical to the test design of the main study in Grade 11, but different from the test design in Grade 9. In Grade 9 the science test was either administered in first or second position. This design changed for the main study in Grade 11; here, the science test was the only competence test that was administered. In order to correct for this change in test position, we used an approximation. We adjusted the estimated correction term by subtracting half of the position effect of Grade 9 (0.088/2; see Kähler, 2020). Additionally, we corrected the linked WLEs as follows: we added half of the position effect to the WLEs of participants receiving the Grade 9 test in first position and we subtracted half of the position effect from the WLEs of participants receiving the Grade 9 test in second position.

7.2.4 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 one-dimensional model (AIC = 10,597.09, BIC = 10,823.00), over the two-dimensional model (AIC = 10,600.70, BIC = 10,832.97). An examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) confirmed a unidimensional scale – the average absolute residual correlation was *M* = 0.00 (*SD* = 0.09). This indicates that the scientific literacy tests administered in Grades 9 and 11 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 3 and the respective tests for measurement invariance based on the Wald statistic (Fischer et al., 2016) are summarized in Table 10.

Measurement invariance for Grade 9 and Grade 11 showed three items with *F*-statistics exceeding the critical value of $F_{.0154}(1, 2{,}038) = 53.37$. Consequently, these three items had to be excluded from the estimation of the correction term.

Moreover, analyses of differential item functioning between the link sample and Starting Cohort 3 in Grade 9 showed no DIF greater than 0.40 for 19 items of the test (difference in logits: *Min* = −0.36, *Max* = 0.37). However, eight items (scg90510_sc3g11_c, scg9052s_sc3g11_c, scg96120_sc3g11_c, scg90810_sc3g11_c, scg9043s_sc3g11_c, scg96530_sc3g11_c, scg9621s_sc3g11_c, scg91120_sc3g11_c) showed a DIF greater than 0.40. These items were therefore excluded from the estimation of the correction term. For Grade 11 (difference in logits: *Min* = −0.34, *Max* = 0.33) there were also eight items with a DIF greater than 0.40 (scg110620_sc3g11_c, scg110520_sc3g11_c, scg11123s_sc3g11_c, scg11021s_sc3g11_c, scg11112s_sc3g11_c, scg116320_sc3g11_c, scg110930_sc3g11_c, scs5131s_sc3g11_c). Therefore, the scientific literacy tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (Fischer et al., 2016).

The correction term was calculated as *c* = 0.7734 (with a link error of 0.071). This correction term was adjusted by subtracting half of the position effect of Grade 9 resulting in a correction term of c = 0.7294 which was subsequently added to each difficulty parameter estimated in Grade 11 (see Table 5). Finally, the correction term of Grade nine c = 0.8782 was added to derive the linked item parameters.

7.3 Scientific literacy scores

In the SUF manifest scientific literacy scores are provided in the form of two different WLEs (scg11_sc1 and scg11_sc1u), including their respective standard error (scg11_sc2 and scg11_sc2u).For scg11_sc1u, person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scored were corrected for the change in test design (see 7.2.3). As a result, the WLE scores provided in scg11_sc1u can be used for longitudinal comparisons between Grades 9 and 11. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in "scg11_sc1" are not linked to the underlying reference scale of grade 9. 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. Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. Plausible values for competence tests administered in the NEPS can be estimated using the R package *NEPSscaling*[1] (Scharl, Carstensen, & Gnambs, 2020).

[1] https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values

Table 10: Differential Item Functioning Analyses between the Main Sample and the Link Sample

	Grade 9				Grade 11			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1.	scg90110_sc3g9_c	-0.27	0.20	1.84	scg116420_sc3g11_c	0.16	0.18	0.79
2.	scg9012s_sc3g9_c	-0.03	0.29	0.01	scg110620_sc3g11_c	-0.69	0.19	13.53
3.	scg90510_sc3g9_c	-0.72	0.19	13.64	scg110630_sc3g11_c	-0.21	0.18	1.42
4.	scg9052s_sc3g9_c	0.94	0.36	6.90	scg11012s_sc3g11_c	-0.26	0.18	1.96
5.	scg90920_sc3g9_c	0.01	0.18	0.00	scg11083s_sc3g11_c	-0.14	0.20	0.52
6.	scg90930_sc3g9_c	-0.31	0.25	1.49	scg110720_sc3g11_c	-0.07	0.18	0.17
7.	scg96120_sc3g9_c	-0.62	0.21	9.07	scg11032s_sc3g11_c	0.38	0.25	2.33
8.	scg96410_sc3g9_c	-0.23	0.32	0.54	scg110330_sc3g11_c	0.14	0.18	0.66
9.	scg96420_sc3g9_c	-0.27	0.21	1.69	scg116510_sc3g11_c	-0.33	0.18	3.53
10.	scg9061s_sc3g9_c	0.00	0.20	0.00	scg11652s_sc3g11_c	-0.04	0.20	0.04
11.	scg90630_sc3g9_c	0.10	0.27	0.13	scg110510_sc3g11_c	0.43	0.18	5.97
12.	scg90810_sc3g9_c	0.93	0.73	1.62	scg110520_sc3g11_c	-1.80	0.18	96.81
13.	scg9083s_sc3g9_c	0.34	0.31	1.27	scg110540_sc3g11_c	2.27	0.23	100.31
14.	scg91030_sc3g9_c	-0.39	0.18	4.65	scg11123s_sc3g11_c	-1.34	0.20	47.38

	Grade 9				Grade 11			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
15.	scg91040_sc3g9_c	-0.05	0.29	0.03	scg11102s_sc3g11_c	-0.28	0.21	1.70
16.	scg91050_sc3g9_c	-0.23	0.22	1.13	scg11021s_sc3g11_c	-0.69	0.23	8.77
17.	scg9042s_sc3g9_c	-0.16	0.28	0.31	scg11022s_sc3g11_c	3.19	0.26	152.01
18.	scg9043s_sc3g9_c	0.23	0.35	0.42	scg11112s_sc3g11_c	-0.67	0.22	9.68
19.	scg9651s_sc3g9_c	-0.15	0.30	0.26	scg116210_sc3g11_c	-1.51	0.21	50.11
20.	scg96530_sc3g9_c	0.57	0.24	5.70	scg11622s_sc3g11_c	0.82	0.28	8.78
21.	scg90320_sc3g9_c	0.20	0.24	0.68	scg116320_sc3g11_c	1.59	0.24	43.75
22.	scg90330_sc3g9_c	-0.17	0.19	0.87	scg110930_sc3g11_c	-0.21	0.25	0.72
23.	scg9621s_sc3g9_c	1.18	0.37	10.52	scs5131s_sc3g11_c	-0.82	0.31	7.15
24.	scg96220_sc3g9_c	0.23	0.23	1.06	scs5132s_sc3g11_c	0.09	0.28	0.11
25.	scg91110_sc3g9_c	-0.48	0.19	6.36				
26.	scg91120_sc3g9_c	-0.60	0.20	8.87				
27.	scg91130_sc3g9_c	-0.06	0.19	0.10				

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 9 and 11 and the link sample (positive values indicate easier items in the link sample); *SE*$_{Δσ}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F.0154 (1, 2,038) = 53.37. A non−significant test indicates measurement invariance.

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. http://doi.org/10.1007/978-1-4612-1694-0_16

Blossfeld, H.-P. & Roßbach, H.-G. (Hrsg.). (2019). Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition ZfE (2. Auflage). Springer VS.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *NEPS Survey Paper No. 1, 2016 Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_I.pdf

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. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/Overview_NEPS_Competence-Data.pdf

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. https://doi: 10.25656/01:8427

Kähler, J. (2020). NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 9 (NEPS Survey Paper No. 79). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP79:1.0

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1002/j.2333-

8504.1992.tb01436.x

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. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

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. https://www.pedocs.de/volltexte/2013/8430/pdf/JERO_2013_2_Pohl_Carstensen_Scaling_of_competence_tests.pdf

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

Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6* (NEPS Survey Paper No. 71). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP71:1.0

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi/10.1214/aos/1176344136

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. https://doi.org/10.1007/s11618-011-0182-7

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. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in Starting Cohort III

```
Title G11 Science analysis, Partial Credit Model;
data filename.dat;
format id 1–7 responses 8–31;

labels << filename_with_labels.txt;

recode (0,1,2,3,4)    (0,0,0,0,1)        !item (4,16-19,21)
recode (0,1,2,3,4)    (0,0,0,1,2)        !item (7,24);
recode (0,1,2,3,4)    (0,0,1,2,3)        !item (5,10);

codes 0,1,2,3,4;
score (0,1)           (0,1)              !item (1-4,6,8,9,11-13,15-22);
score (0,1,2)         (0,0.5,1)          !item (7,23);
score (0,1,2,3)       (0,0.5,1,1.5)      !item (5,10);
score (0,1,2,3,4)     (0,0.5,1,1.5,2)    !item (14,24);

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

Items	Position in the test	Component	Context
scg116420_sc3g11_c	1	KAS	Environment
scg110620_sc3g11_c	2	KOS	Technology
scg110630_sc3g11_c	3	KOS	Technology
scg11012s_sc3g11_c	4	KOS	Technology
scg11083s_sc3g11_c	5	KOS	Technology
scg110720_sc3g11_c	6	KOS	Technology
scg11032s_sc3g11_c	7	KOS	Environment
scg110330_sc3g11_c	8	KOS	Environment
scg116510_sc3g11_c	9	KAS	Health
scg11652s_sc3g11_c	10	KAS	Health
scs56320_sc3g11_c	11	KAS	Health
scg110510_sc3g11_c	12	KOS	Health
scg110520_sc3g11_c	13	KOS	Health
scg110540_sc3g11_c	14	KOS	Health
scg11123s_sc3g11_c	15	KOS	Environment
scg11102s_sc3g11_c	16	KOS	Environment

scg11021s_sc3g11_c	17	KOS	Technology
scg11022s_sc3g11_c	18	KOS	Technology
scg11112s_sc3g11_c	19	KOS	Health
scg116210_sc3g11_c	20	KAS	Environment
scg11622s_sc3g11_c	21	KAS	Environment
scg116320_sc3g11_c	22	KAS	Technology
scg110930_sc3g11_c	23	KOS	Environment
scs5131s_sc3g11_c	24	KOS	Technology
scs5132s_sc3g11_c	25	KOS	Technology

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