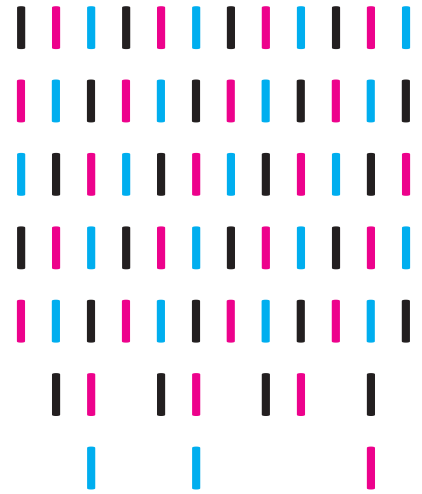




NEPS SURVEY PAPERS

Marie-Ann Sengewald, Inga Hahn
and Jana Kähler

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORTS 4 AND 6 (WAVE 14)

NEPS Survey Paper No. 109
Bamberg, November 2023

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.

Editor-in-Chief: Thomas Bäumer, LIfBi

Review Board: Board of Directors, Heads of LIfBi Departments, and Scientific Management of NEPS Working Units

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 Cohorts 4 and 6 (Wave 14)

Marie-Ann Sengewald¹, Inga Hahn² and Jana Kähler²

¹Leibniz Institute for Educational Trajectories, Bamberg

²Leibniz Institute for Science and Mathematics Education, Kiel

E-mail address of lead author:

marie.sengewald@lifbi.de

Bibliographic data:

Sengewald, M.-A., Hahn, I. & Kähler, J. (2023). *NEPS Technical Report for Science: Scaling Results for Starting Cohorts 4 and 6 (Wave 14)* (NEPS Survey Paper No. 109). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP109:1.0>

Acknowledgements:

This report is an extension to NEPS *Working Paper 23* (Schöps & Sass, 2013), NEPS *Survey Paper 6* (Hahn & Kähler, 2016), and NEPS *Survey Paper 19* (Haschke, Kähler & Hahn, 2017) that presents the scaling results for Science of Starting Cohorts 4 and 6 in the previous waves. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced verbatim to facilitate the understanding of the presented results.

NEPS Technical Report for Science: Scaling Results for Starting Cohorts 4 and 6 (Wave 14)

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. To evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for a test of scientific literacy that was administered in Wave 14 of Starting Cohorts 4 (ninth grade) and 6 (adults). The scientific literacy test included 40 items with different response formats that were administered in a computerized multi-stage test design. The adaptive nature of the test design allowed the administration of different items to each respondent tailored to the individual competence level. The test was administered to a total of 6,215 respondents (50% women) from Starting Cohort 4 ($N = 2,493$) and Starting Cohort 6 ($N = 3,722$). The responses of the participants were scaled using a unidimensional partial credit model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test differentiated well between respondents, exhibited good reliability, and showed a satisfactory fit to the item response model. Limitations of the test pertained to differential item functioning, that some items showed different properties in specific subgroups; especially for gender some items showed different properties as well as for the two starting cohorts that might be referred to as age differences. Overall, the scientific literacy test had good psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

item response theory, scaling, scientific literacy, scientific use file

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span (see Blossfeld & Roßbach, 2019). These include, among others, reading competence, mathematical competence, scientific literacy, computer literacy, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß and colleagues (2021). Most of the administered competence tests are developed specifically for implementation in the NEPS and, thus, are routinely evaluated using psychometric models based on item response theory (IRT; see Pohl & Carstensen, 2012).

In this paper, the results of these analyses are presented for a test of scientific literacy that was administered in Wave 14 of Starting Cohorts 4 (ninth grade) and 6 (adults). The following sections first introduce the main concepts of the administered test and the test design. Then, the competence data of the two starting cohorts are described. Subsequently, we report the results of various psychometric analyses that were performed to estimate competence scores and to check the quality of the test. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is presented.

Please note that the analyses summarized 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 report. However, we do not expect pronounced differences in the presented results.

2. Testing Scientific Literacy

2.1 Conceptual Framework

The framework and test development for the scientific literacy test are described by Weinert and colleagues (2011) as well as Hahn and colleagues (2013) and are depicted in Figure 1. In the following, we briefly describe 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 with two components. These are a) knowledge of science (KOS) and b) knowledge about science (KAS). Using the definition by PISA (OECD, 2006; Prenzel et al., 2007), KOS is specified as knowledge of basic scientific concepts and facts whereas KAS can be regarded as the understanding of scientific processes. KOS is divided into content-related components: matter, system, development, and interaction. KAS is divided in the process-related components scientific enquiry and scientific reasoning. KAS and KOS are implemented in three contexts: health, environment, and technology (see Hahn et al., 2013, and Weinert et al., 2011, for the description of the framework).

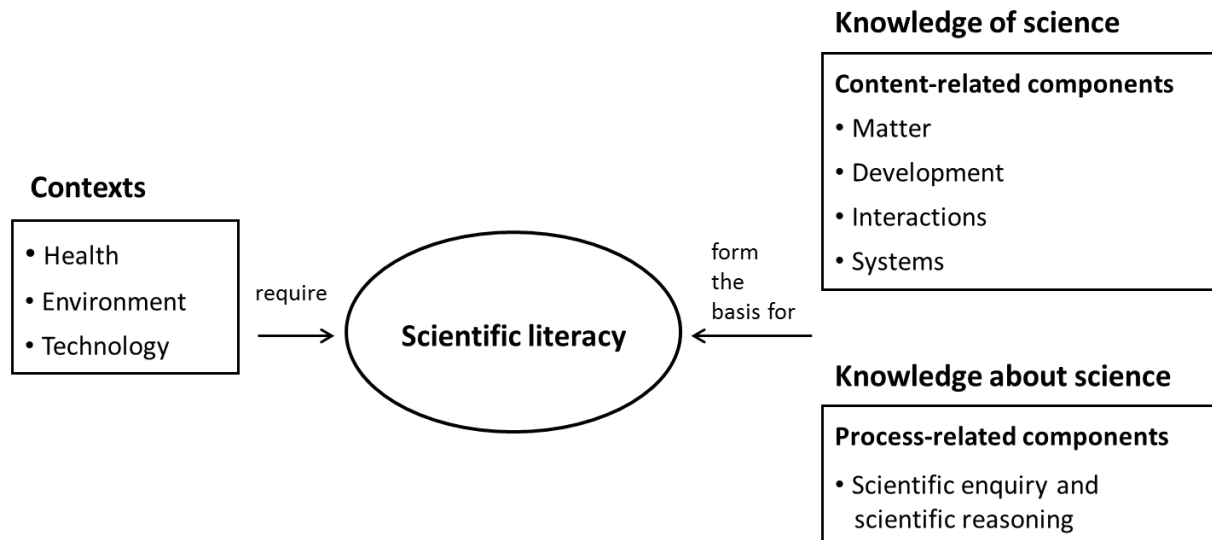


Figure 1. Assessment framework for scientific literacy (Hahn et al., 2013)

The scientific literacy test that was administered in the present study included 40 items. Each item refers to one context-component combination. The distribution of the items across the two components and the three contexts of the assessment framework are summarized in Tables 1 and 2 and details are provided in Appendix A.

Table 1

Number of Items by Different Components

Component	Number of Items
Knowledge of Science (KOS)	28
Knowledge about Science (KAS)	12
Total number of items	40

Table 2

Number of Items by Different Contexts

Context	Number of Items
Health	10
Technology	18
Environment	12
Total number of items	40

2.2 Item Format

The majority of the items are constructed as simple multiple choice (MC) items. In MC items the test taker has to identify the correct answer out of four response options, with one option

being correct and three response items functioning as distractors (i.e., they were incorrect). Two types of more complex response formats were used, too. These are complex multiple choice (CMC) and matching items (MA). In CMC items, four subtasks were presented and the test taker has to decide for each response option whether the answer is correct or not. MA items required the test taker to match a number of responses to a given set of statements. For two MA items (sca14605s_c, sca14121s_c), multiple statements formed one subtask that was only correct, when all statements were matched correctly. For all other MA items, each statement formed one subtask, whereas the number of subtasks ranged from four to six. The number of successful task completions for all CMC and MA items were summed for each item to create a composite score.

Table 3

Number of Items by Different Response Formats

Response format	Number of Items
Simple multiple-choice items	25
Complex multiple-choice items	10
Matching items	5
Total number of items	40

2.3 Test Design

The scientific literacy test administered in the present study adopted a multi-stage design that included a total of 40 items (see Figure 2). However, each respondent received only a subset of 23 items depending on his or her estimated scientific literacy. The multi-stage test (MST) included three distinct stages that were sequentially administered to each respondent. The test time for all three stages was 28 minutes. Each stage consisted of different blocks that could vary in the difficulty of the included items and each respondent received only one of the blocks. The difficulty of each item was estimated from previously published competence data in different starting cohorts and unpublished data from various developmental studies. Stage 1 included two blocks, one with an *easy/medium* item set and one with a *medium/difficult* item set, with 8 items in each block. Stage 2 had two blocks (*easy* or *difficult*) with 8 items in each block and Stage 3 had three blocks (*easy*, *medium*, or *difficult*) with 7 items in each block (see Figure 2). The block at the first stage was assigned to each respondent based on his or her estimated scientific literacy in a previous wave. Respondents with a competence score falling below the mean competence of the sample received the easy to medium block in Stage 1, whereas respondents with competence scores exceeding the mean of the sample were administered the medium to difficult block in Stage 1. Depending on their responses in the previous stage, each respondent received the best fitting block in the subsequent stage.

The MST tried to respect the theoretical testing framework (see above) by distributing the different components and contexts approximately uniformly across the different blocks within each stage. To ensure a common scale for all test takers, the different blocks are linked within each stage by assigning some items to two adjacent blocks. For example, in Stage 3 item

sca14603s_c was assigned to the easy and medium block. In this case, item sca14603s_c acts as link between the two levels and allows for the estimation of a common score for respondents receiving different blocks. There was no multi-matrix design regarding the order of the items within a specific block. All respondents assigned to a specific block received the test items in the same order.

The study assessed different competence domains including, among others, scientific literacy and computer literacy. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the scientific literacy test was either administered as the first or the second test (i.e., after the computer literacy). A detailed description of the study design is available on the NEPS website (<http://www.neps-data.de>).

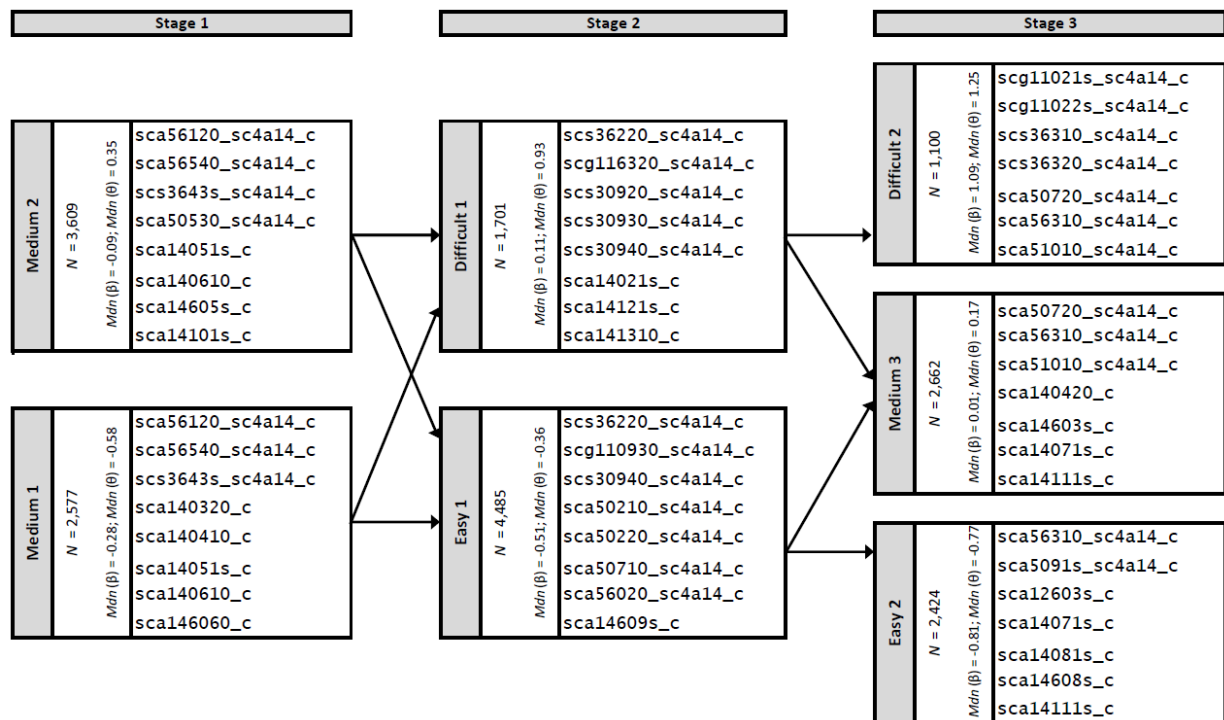


Figure 2. Multi-Stage Test Design with Items, Sample Sizes (N), Median Item Difficulties (β), and Median Respondent Proficiencies (θ) for Each Block

Note. Item names correspond to those from Starting Cohort 4 (ninth grade). An equivalency table for the variable names in Starting Cohort 6 (adults) is given in Appendix A.

3. Data

A total of 6,186 students (50% women) had at least three valid responses on the scientific literacy test and, thus, were used for the psychometric analyses (see Pohl & Carstensen, 2012). Of these, $N = 2,473$ (51% women) were from Starting Cohort 4 and $N = 3,713$ (49% women) were from Starting Cohort 6. The age of the respondents ranged from 25 to 76 years. Basic sociodemographic information of the two samples is summarized in Table 4.

Table 4

Sample Descriptions

	Starting Cohort 4	Starting Cohort 6
Sample size	2,473	3,713
Percentage women	51%	49%
Percentage migration background	9%	7%
Mean age	26.59	58.20
<i>SD</i> age	1.35	10.06

4. Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) items that test takers did not reach within a stage, c) items that were not presented in the subsequent stages because test takers prematurely aborted the test in a previous stage d) items that have not been administered, and, finally, e) missing responses within complex multiple-choice items that are not determined. Omitted items occurred when test takers skipped some items. Due to time limits or lack of motivation, not all persons finished the test. All missing responses after the last valid response within a level of the current stage were coded as not reached, whereas the remaining items in the following stages were coded as missing due to test abortion. Because of the MST design, many items were not administered to any given participant. These items were missing by design. Because two matching items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A complex multiple choice or matching item was coded as missing if at least one subtask contained a missing response. 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). Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

4.2 Scaling Model

Item and person parameters were estimated using a partial credit model (PCM; Masters, 1982) with Gauss-Hermite quadrature (21 nodes). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

Complex multiple-choice and matching items consisted of a set of subtasks that were aggregated to a polytomous variable for each item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the partial credit item was scored as missing. Response categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category. This proceeding differed for polytomous items that were administered in previous tests. For these items we used exactly the same scoring scheme and collapsed the same categories as in the previous tests (i.e., we used `scs3643s_sc4a14_c` as it was with five response categories, a dichotomous scoring for `scg11_021s_sc4a14_c` and `scg11_022s_sc4a14_c` requiring that all subtasks have to be solved, and collapsed the two lowest categories for `sca5091s_sc4a14_c`).

Respondents' scientific literacies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while dichotomous items were scored as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in Section 7.

4.3 Checking the Quality of the Test

The scientific literacy test was specifically constructed for administration in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

The multiple-choice items consisted of one correct response option and multiple distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined using the point-biserial correlation between selecting an incorrect response option for a given item and the total correct score for the remaining items. Negative correlations indicate good distractors, whereas correlations between .00 and .05 were considered acceptable and correlations above .05 were viewed as problematic distractors (Pohl & Carstensen, 2012).

Before aggregating the subtasks of a complex multiple-choice and matching item to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the multiple-choice items in a Rasch (1960) model. The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective t -value, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous variables that were included in the final scaling model. We did not exclude any subtask of the polytomous items that were already in another test, in line with the previous proceeding.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous and polytomous items to the PCM (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective t -value, and the item characteristic curves (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. Overall judgment of the fit of an item was based on all fit indicators.

The scientific literacy test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased. For the present study, measurement invariance was investigated for the variables gender, age, the number of books at home (as a proxy for cultural status), migration background (see Pohl & Carstensen, 2012, for a description of these variables), starting cohort (4 or 6), and test rotation (first versus second). Differential item functioning (DIF) was examined using a multigroup item response 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 considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Minimum hypothesis tests (see Fischer, et al., 2016) were used to statistically test whether the observed differences were significantly larger than 0.4 and, thus, were at least small in size. Additionally, measurement invariance 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) 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 dimensionality of the test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations as indicated by Yen's (1984) Q_3 indicate unidimensionality. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality. Moreover, we examined a multidimensional model based on the two components and three contexts using quasi-Monte Carlo integration (with 2,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

4.4 Software

The item response models were estimated with the TAM package version 4.1-4 (Robitzsch, Kiefer, & Wu, 2022) in R version 4.2.0 (R Core Team, 2023).

5. Results

Preliminary analyses showed that for some polytomous items, specific subtasks exhibited a poor fit to the item response model (i.e., negative or zero discriminations). Therefore, these subtasks (i.e., sca14021d_c, sca14603a_c, sca14071b_c, sca14111_05_c) were removed from the item scoring. The analyses presented in the following sections refer to all 40 items, but the respective polytomous items have fewer response options.

5.1 Missing Responses

5.1.1 Missing responses per person

The number of missing responses when participants omitted items is illustrated in Figure 3. About 80% of respondents did not skip any item, while about 9% skipped one item. However, there were pronounced differences between the two starting cohorts. The adult sample (Starting Cohort 6) exhibited substantially more omitted items; about 62% of respondents in Starting Cohort 6 had no omitted item, whereas 21% exhibited a single omitted item. In contrast, in Starting Cohort 4 about 84% of respondents had no omitted item, while about 11% had a single omitted item.

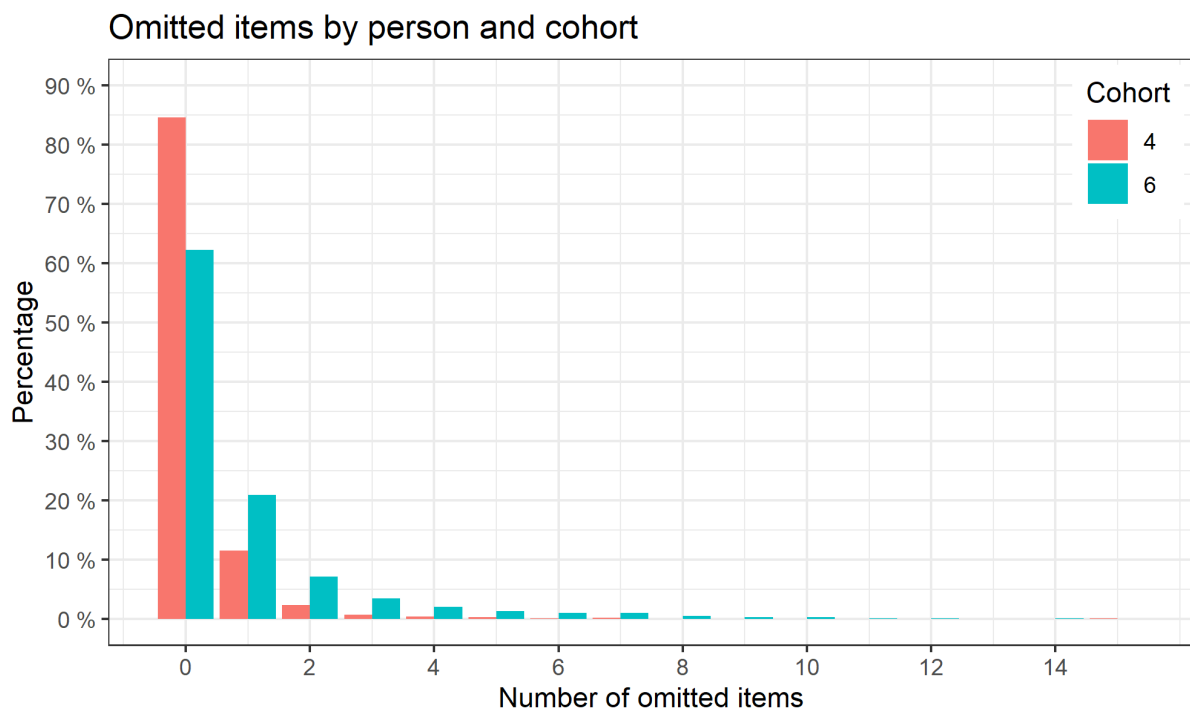


Figure 3. Number of Omitted Items by Starting Cohort

Another source of missing responses is that items were not reached by the respondents because they aborted the test, for example, because the time limit was reached or due to a lack of motivation. These missing values refer to items after the last valid response. The results given in Figure 4 show that around 45% of the sample did not prematurely abort the test. Again, the missing rates were substantially higher in the older sample as compared to the younger sample. In Starting Cohort 4 about 56% of the respondents did not prematurely abort the test, whereas in Starting Cohort 6 this was the case for about 37% of the respondents.

Nearly all respondents received at least 15 of the 23 test items before the test was aborted (indicated by the low frequency of more than 8 not reached items).

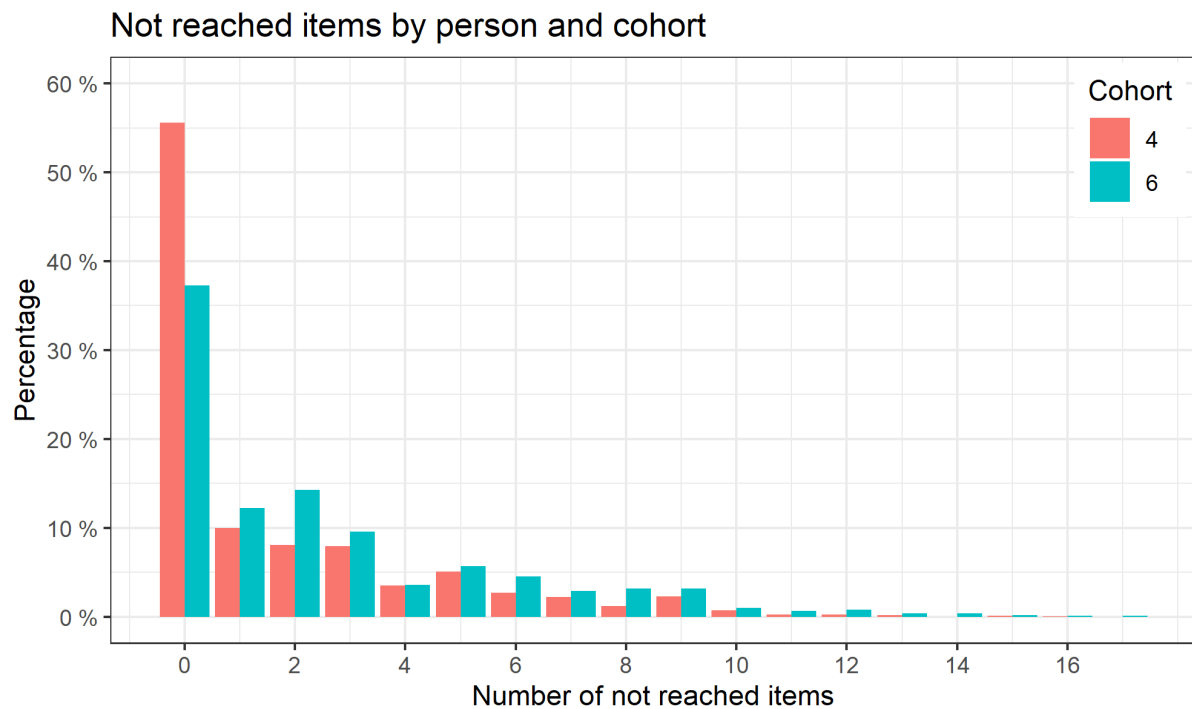


Figure 4. Number of not reached items by sample

The total number of missing responses, aggregated over omitted and not reached missing responses per person, is illustrated in Figure 5. As more than half of the respondents prematurely aborted the test, the number of missing values is substantial. Still, about 38% of the respondents had no missing response at all and the median number of missing responses was 2. Again, in Starting Cohort 6 fewer participants had no missing value (about 25%) as compared to Starting Cohort 4 (about 48%).

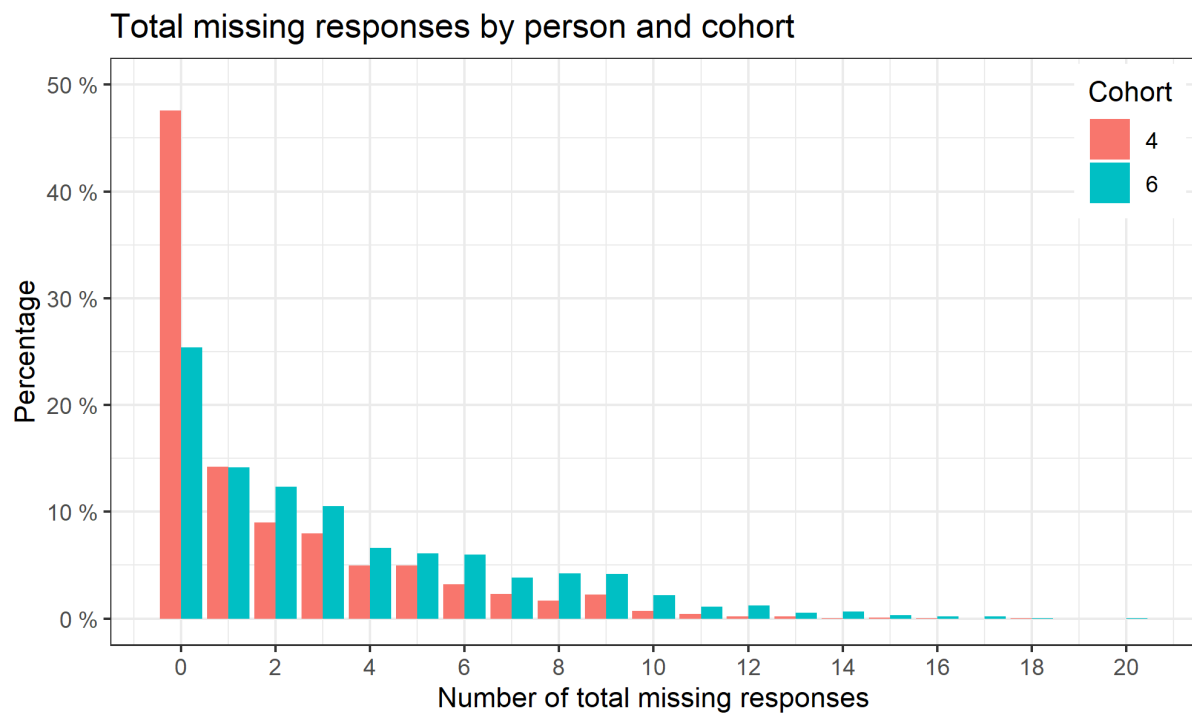


Figure 5. Total number of missing responses by sample

5.1.2 Missing responses per item

Table 5 provides information on the occurrence of different kinds of missing responses per item for the two samples. The percentage of omitted responses varied across items between 0% and 11% ($Mdn = 0.25\%$) in Starting Cohort 4 and between 0% and 35% ($Mdn = 0.80\%$) in Starting Cohort 6. In contrast, the percentage of missing responses because participants did not reach the item within a stage (i.e., excluding missing responses resulting from test abortion) varied across items between 0% and 44% ($Mdn = 1.08\%$) in Starting Cohort 4 and between 0% and 61% ($Mdn = 2.70\%$) in Starting Cohort 6.

Table 5

Percentage of Missing Values by Item

	Item	Stage	Starting Cohort 4				Starting Cohort 6			
			N	N _v	OM	NR	N	N _v	OM	NR
1	sca56120_sc4a14_c	1	2,473	2,473	0.00	0.00	3,713	3,707	0.16	0.00
2	sca56540_sc4a14_c	1	2,473	2,469	0.16	0.00	3,713	3,687	0.70	0.00
3	scs3643s_sc4a14_c	1	2,473	2,443	1.21	0.00	3,713	3,474	6.44	0.00
4	sca50530_sc4a14_c	1	1,783	1,779	0.22	0.00	1,826	1,814	0.66	0.00
5	sca140320_c	1	690	685	0.72	0.00	1,887	1,864	1.22	0.00
6	sca140410_c	1	690	687	0.43	0.00	1,887	1,876	0.58	0.00
7	sca14051s_c	1	2,473	2,420	2.14	0.00	3,713	3,254	12.36	0.00
8	sca140610_c	1	2,473	2,473	0	0.00	3,713	3,701	0.32	0.00
10	sca14101s_c	1	1,783	1,759	1.29	0.06	1,826	1,677	7.78	0.38
11	sca146060_c	1	690	687	0.43	0.00	1,887	1,881	0.32	0.00
12	scs36220_sc4a14_c	2	2,472	2,452	0.73	0.08	3,706	3,529	4.61	0.16
13	scg116320_sc4a14_c	2	880	879	0.00	0.11	821	811	0.49	0.73
14	scg110930_sc4a14_c	2	1,592	1,587	0.25	0.06	2,885	2,832	1.32	0.52
15	scs30920_sc4a14_c	2	880	875	0.00	0.57	821	809	0.24	1.22
16	scs30930_sc4a14_c	2	880	872	0.11	0.8	821	803	0.37	1.83
17	scs30940_sc4a14_c	2	2,472	2,450	0.44	0.44	3,706	3,632	0.67	1.32
18	sca14021s_c	2	880	863	0.57	1.36	821	768	1.83	4.63
19	sca14121s_c	2	880	747	9.89	5.23	821	652	9.62	10.96
20	sca141310_c	2	880	823	0.00	6.48	821	711	0.12	13.28
21	sca50210_sc4a14_c	2	1,592	1,579	0.44	0.38	2,885	2,763	2.46	1.77

		Starting Cohort 4				Starting Cohort 6				
	Item	Stage	N	N _v	OM	NR	N	N _v	OM	NR
22	sca50220_sc4a14_c	2	1,592	1,578	0.25	0.63	2,885	2,782	1.25	2.32
23	sca50710_sc4a14_c	2	1,592	1,564	0.25	1.51	2,885	2,774	0.76	3.08
24	sca56020_sc4a14_c	2	1,592	1,541	0.31	2.89	2,885	2,612	4.06	5.41
25	sca14609s_c	2	1,592	1,523	0.25	4.08	2,885	2,467	5.62	8.87
26	scg11021s_sc4a14_c	3	566	536	0.35	4.95	478	430	2.72	7.32
27	scg11022s_sc4a14_c	3	566	521	0.18	7.77	478	414	0.42	12.97
28	scs36310_sc4a14_c	3	566	488	0.00	13.78	477	380	0.21	20.13
29	scs36320_sc4a14_c	3	566	469	0.00	17.14	477	361	0.21	24.11
30	sca50720_sc4a14_c	3	1,595	1,440	0.25	9.47	1,918	1,713	0.52	10.17
31	sca56320_sc4a14_c	3	1,595	1,378	0.06	13.54	1,918	1,623	0.21	15.17
32	sca51010_sc4a14_c	3	1,595	1,298	0.00	18.62	1,918	1,536	0.16	19.76
33	sca140420_c	3	1,029	895	0.10	12.93	1,441	1,202	0.42	16.17
34	sca56310_sc4a14_c	3	755	748	0.13	0.79	1,422	1,392	0.84	1.27
35	sca5091s_sc4a14_c	3	755	716	2.12	3.05	1,422	1,194	9.21	6.82
36	sca14603s_c	3	1,784	1,428	1.12	18.83	2,863	1,959	7.79	23.79
37	sca14071s_c	3	1,784	1,327	0.73	24.89	2,863	1,858	4.33	30.77
38	sca14081s_c	3	755	619	0.53	17.48	1,422	1,020	2.81	25.46
39	sca14608s_c	3	755	499	3.97	29.93	1,422	513	13.5	50.42
40	sca14111s_c	3	1,784	1,006	0.00	43.61	2,863	1,106	0.00	61.37

Note. Stage = Stage within the test the item was administered in. N = Number of respondents the item was administered to. N_v = Number of valid responses. NR = Percentage of respondents that did not reach the item. OM = Percentage of respondents that omitted the item. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 6 are given in Appendix A.

5.2 Parameter Estimates

Preliminary analyses identified pronounced differential item functioning between the two starting cohorts for the items sca56120_sc4a14_c, sca140610_c, and scs36320_sc4a14_c. Therefore, these items were treated as unique in each cohort to estimate different item parameters in Starting Cohorts 4 and 6. Similarly, item sca51010_sc4a14_c was treated as unique for male and female respondents, as this item showed substantial differential item functioning with regard to gender. In this way, DIF does not bias the resulting parameter estimates.

5.2.1 Item Parameters

The fourth column in Table 6 presents the percentage of correct responses in relation to all valid responses for each multiple-choice item. The percentage of correct responses varied between 15% and 72% correct responses. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty. The estimated item difficulties (for dichotomous items) and location parameters (for polytomous items) are given in the fifth column of Table 6. The step parameters for polytomous variables are summarized in Table 7. The item difficulties and location parameters were estimated by constraining the mean of the ability distribution to zero. The standard errors (*SE*) of most estimated parameters (see Tables 6 and 7) were rather small for most items. The estimated item difficulties ranged from -1.626 (item sc1460810s_c) and 2.552 (item scs36320_sc6a14_sc6_c) with a median of -0.12 and, thus, included easy as well as difficult items.

Table 6

Item Parameters for Combined Sample

	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	Discr.	$\alpha Q3$
1	sca56120_sc4a14_c_4	2,473	67.53	-0.705	0.046	0.960	-2.188	1.096	.048
41	sca56120_sc6a14_c_6	3,707	38.25	0.456	0.036	0.945	-4.209	1.157	.047
2	sca56540_sc4a14_c	6,156	55.77	-0.268	0.027	0.969	-3.451	1.004	.026
3	scs3643s_sc4a14_c	5,917	n.a.	0.096	0.013	1.061	3.752	0.285	.038
4	sca50530_sc4a14_c	3,593	65.29	-0.394	0.037	0.977	-1.652	1.024	.028
5	sca140320_c	2,549	34.21	0.280	0.044	1.067	3.912	0.352	.058
6	sca140410_c	2,563	53.1	-0.582	0.042	1.067	5.417	0.373	.043
7	sca14051s_c	5,674	n.a.	-0.287	0.013	0.992	-0.498	0.428	.042
8	sca140610_c_4	2,473	48.32	0.205	0.043	0.964	-2.618	1.034	.045
42	sca140610_c_6	3,701	68.79	-0.993	0.038	0.981	-1.20	0.963	.034
9	sca14605s_c	2,764	n.a.	0.098	0.017	1.037	1.611	0.342	.055
10	sca14101s_c	3,436	n.a.	0.210	0.016	1.052	2.571	0.299	.035
11	sca146060_c	2,568	55.49	-0.69	0.042	1.026	2.106	0.606	.031
12	scs36220_sc4a14_c	5,981	57.68	-0.337	0.028	0.997	-0.30	0.83	.027
13	scg116320_sc4a14_c	1,690	59.17	0.338	0.052	0.948	-3.268	1.515	.047
14	scg110930_sc4a14_c	4,419	51.78	-0.357	0.032	0.952	-5.308	1.276	.028
15	scs30920_sc4a14_c	1,684	65.62	0.036	0.054	0.950	-2.545	1.574	.050
16	scs30930_sc4a14_c	1,675	72.48	-0.312	0.057	0.944	-2.186	1.687	.046
17	scs30940_sc4a14_c	6,082	53.83	-0.183	0.028	1.060	6.560	0.527	.030
18	sca14021s_c	1,631	n.a.	0.520	0.044	0.980	-0.568	0.684	.037
19	sca14121s_c	1,399	n.a.	0.555	0.056	1.027	1.71	0.579	.037

	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	Discr.	<i>aQ3</i>
20	sca141310_c	1,534	62.65	0.179	0.055	1.012	0.638	0.682	.030
21	sca50210_sc4a14_c	4,342	53.41	-0.428	0.032	0.981	-2.106	1.029	.026
22	sca50220_sc4a14_c	4,360	66.77	-1.049	0.034	0.975	-1.887	1.094	.032
23	sca50710_sc4a14_c	4,338	65.33	-0.979	0.033	1.005	0.379	0.766	.038
24	sca56020_sc4a14_c	4,153	62.17	-0.822	0.034	1.009	0.781	0.76	.026
25	sca14609s_c	3,990	n.a.	-0.594	0.017	0.945	-2.800	0.697	.051
26	scg11021s_sc4a14_c	966	n.a.	0.882	0.067	1.002	0.097	1.006	.055
27	scg11022s_sc4a14_c	935	n.a.	2.072	0.077	1.008	0.239	0.919	.050
28	scs36310_sc4a14_c	868	47.24	1.087	0.071	1.004	0.215	0.624	.062
29	scs36320_sc4a14_c_4	469	39.66	1.418	0.098	1.011	0.374	0.576	.075
43	scs36320_sc6a14_c_6	361	18.84	2.552	0.139	1.036	0.457	0.307	.062
30	sca50720_sc4a14_c	3,153	41.93	0.691	0.038	0.946	-4.593	1.315	.039
31	sca56320_sc4a14_c	3,001	55.55	0.077	0.039	1.015	1.375	0.724	.038
32	sca51010_sc4a14_c_m	1,279	15.25	2.016	0.080	0.999	-0.005	0.821	.039
44	sca51010_sc4a14_c_f	1,556	46.53	0.588	0.054	0.975	-1.583	1.100	.046
33	sca140420_c	2,097	54.22	-0.06	0.045	1.038	3.287	0.339	.028
34	sca56310_sc4a14_c	2,140	70.70	-1.575	0.049	0.983	-0.814	1.193	.037
35	sca5091s_sc4a14_c	1,910	n.a.	-0.899	0.026	0.994	-0.224	0.504	.040
36	sca14603s_c	3,387	n.a.	-0.093	0.023	1.022	1.379	0.284	.031
37	sca14071s_c	3,185	n.a.	-0.814	0.020	1.051	2.219	0.218	.025
38	sca14081s_c	1,639	n.a.	-1.626	0.057	0.977	-0.924	1.153	.031
39	sca14608s_c	1,012	n.a.	-0.133	0.039	0.941	-2.320	1.181	.069
40	sca14111s_c	2,112	n.a.	-0.254	0.020	1.054	2.556	0.245	.048

Note. N = Number of observed responses; Difficulty = Item difficulty / location; SE = Standard error of item difficulty / location; WMNSQ = Weighted mean square; t = t -value for WMNSQ; Discr. = Discrimination parameter of a generalized partial credit model; aQ_3 = Average absolute residual correlation for item (Yen, 1983); n.a. = not available. Percent correct scores are not informative for polytomous item scores and, therefore, are not reported. Item names refer to Starting Cohort 4. Items that were different due to DIF effects are highlighted in *italic*. Three items were treated as unique in Starting Cohort 4 (*sca56120_sc4a14_c*, *sca140610_sc4_c*, *scs36320_sc4a14_c*) and 6 (*sca56120_sc6a14_c*, *sca140610_sc4_c*, *scs36320_sc6a14_c*). Similarly, one item was treated as unique for male (*sca51010_sc4a14_m_c*) and female (*sca51010_sc4a14_f_c*) respondents.

Table 7

Step Parameters (with Standard Errors) for Polytomous Items in Combined Sample

	Item	Step 1	Step 2	Step 3	Step 4
3	scs3643s_sc4a14_c	-1.123 (0.0333)	-0.463 (0.0273)	0.481 (0.0323)	1.105
7	sca14051s_c	-1.098 (0.0333)	-0.532 (0.0283)	0.520 (0.0330)	1.110
9	sca14605s_c	-0.592 (0.0443)	-0.568 (0.0393)	0.228 (0.0433)	0.932
10	sca14101s_c	-0.753 (0.0430)	-0.456 (0.0353)	0.374 (0.0423)	0.835
18	sca14021s_c	-1.476 (0.0533)	1.476		
25	sca14609s_c	-0.068 (0.0323)	-0.529 (0.0333)	0.597	
35	sca5091s_sc4a14_c	-0.660 (0.0473)	-0.025 (0.0473)	0.685	
36	sca14603s_c	-0.341 (0.0353)	0.341		
37	sca14071s_c	-0.283 (0.0363)	-0.021 (0.0393)	0.304	
39	sca14608s_c	0.023 (0.0683)	-0.023		
40	sca14111s_c	-0.070 (0.0443)	0.168 (0.0523)	-0.098	

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

5.2.2 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. Because some items in the scientific literacy test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 6, the item difficulties of the scientific literacy items and the ability of the respondents are plotted on the same scale. The distribution of the respondents' estimated abilities is mapped onto the left side whereas the right side shows the distribution of the category thresholds. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.674, which implies some variance restrictions, but still an acceptable differentiation between respondents. The reliability of the test (EAP/PV reliability = .74, WLE reliability = .73) was acceptable. The median of the category threshold distribution was about -0.27 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, on average, the items were slightly too easy for the participants. As a consequence, proficiency estimates in low- and medium-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

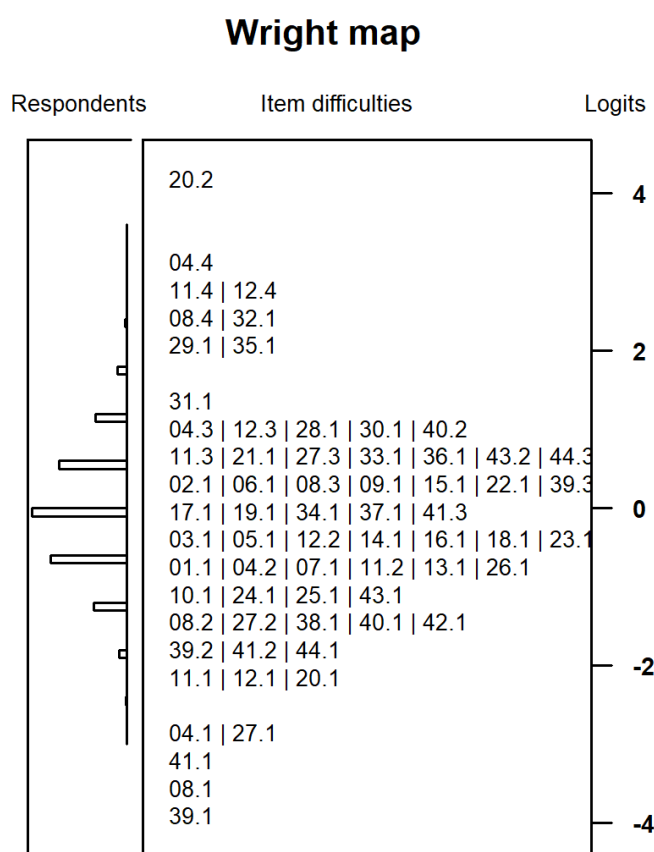


Figure 6. Test targeting. The distribution of the person abilities in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item numbers given in Table 6 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Distractor analyses

To investigate how well the distractors of the multiple-choice items performed the point-biserial correlations between each incorrect response (distractor) and the respondents' total correct scores for the remaining items were calculated. The median point-biserial correlations for the distractors fell at $-.13$ ($Min = -.29$, $Max = .05$). In contrast, the correlations of the correct responses with the total scores varied between $.13$ and $.42$ ($Mdn = .24$). These results indicate that the distractors functioned well.

5.3.2 Item fit

The evaluation of item fit was based on the final scaling model presented above. Again, the test quality was examined for the combined sample including respondents from both starting cohorts. Altogether item fit was satisfactory (see Table 6). For all items WMNSQ values were below 1.150 and ranged from 0.941 (item sca14608s_c) to 1.067 (items sca140320_c, sca140410_c) with a median of 0.999. Also a visual inspection of the ICCs showed no pronounced deviation from the expected ICC.

5.3.3 Differential item functioning

DIF was used to evaluate whether the measurement models were comparable for several subgroups. For this purpose, DIF was examined for the variables gender, migration background, the number of books at home (as a proxy for cultural capital), age group, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, we examined DIF effects between the two samples from Starting Cohorts 4 and 6. The differences between the estimated item difficulties (on the logit scale) in the various groups are summarized in Table 8. For example, the column "men vs. women" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for men, whereas a negative value would highlight a lower difficulty for men as opposed to women. Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models which allow for DIF to those that estimate main effects (see Table 9).

Gender: The sample included 3,106 men and 3,074 women. One item (sca51010_sc4a14_c) showed substantial DIF of more than one logit. This item was treated as unique for men and women in the scaling model, thus, accounting for the DIF in parameter estimates. Gender differences in scientific literacy are substantial as indicated by the main effect of 0.432 logits (Cohen's $d = 0.539$). In addition to the one item that was treated as unique, six items exhibited considerable DIF greater than 0.60 logits (sca50720_sc4a14_c, sca50910s_sc4a14_c, sca56120_sc4a14_c, scg11021s_sc4a14_c, scg116320_sc4a14_c, scs30930_sc4a14_c) and five items showed small DIF greater than 0.40 logits (sca56120_sc6a14_c, sca140420_c, sca14605s_c, scs30920_sc4a14_c, scs30940_sc4a14_c). An overall test for DIF (see Table 9) using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) showed that the DIF model exhibited a better fit to the data as compared to a model that only estimated the main effect (but ignored potential DIF). However, since there were some items with DIF in favor of men and some items with DIF in favor of women, the DIF effects are rather balanced and none of the items had to be removed. The main effect when

not allowing for item-level DIF with regard to sex is a bit lower (main effect of 0.213), indicating that parameter estimates are slightly distorted.

Migration background: There were 5,684 respondents without migration background and 484 respondents with a migration background. In comparison to participants with a migration background, participants without a migration background had, on average, a higher scientific literacy (main effect = 0.208 logits, Cohen's $d = 0.253$). Three items exhibited a considerable DIF greater than 0.6 logits (sca140610_sc6_c, scg11021s_sc4a14_c, scs36310_sc4a14_c) and one item showed small DIF greater than 0.40 logits (sca14081s_c). Still, the overall test for DIF using the BIC favored the main effect model that did not include item-level DIF (see Table 9) and the main effects differed only slightly between the models (main effect of 0.138 logits in the model not accounting for DIF). These results indicate that there was no pronounced DIF concerning migration background that might have distorted the parameter estimates.

Books: The number of books at home was used as an indicator of respondents' cultural capital. There were 1,689 respondents with 0 to 100 books at home and 4,382 respondents with more than 100 books at home. On average, participants with lower cultural capital performed -0.435 logits (Cohen's $d = -0.551$) worse on the scientific literacy test as compared to children with higher cultural capital. One item exhibited a noteworthy DIF greater than 0.4 logits (sca56120__sc4a14_c). The overall test for DIF using the BIC favored the main effect model that did not include item-level DIF (see Table 9). Still the main effect when not allowing for item-level DIF with regard to the number of books at home is a bit lower (main effect of -0.262 logits), indicating that parameter estimates are slightly distorted.

Age group: There were 3,232 respondents younger than 50 years and 2,954 respondents that were older. In comparison to younger participants, older participants had, on average, a higher scientific literacy (main effect = 0.369 logits, Cohen's $d = 0.451$). Four items exhibited a considerable DIF greater than 0.6 logits (sca50710_sc4a14_c, sca56120_sc6a14_c, sca140610_sc6_c, scg11022s_sc4a14_c) and three items showed small DIF greater than 0.40 logits (sca56320_sc4a14_c, sca56540_sc4a14_c, scs30940_sc4a14_c). The overall test for DIF using the AIC and BIC also favored the DIF model (see Table 9). However, since there were some items with DIF in favor of younger and some items with DIF in favor of older persons, the DIF effects are rather balanced and none of the items had to be removed. The main effect when not allowing for item-level DIF with regard to age groups is a bit lower (main effect of 0.230 logits), indicating that the DIF effects slightly distorted the parameter estimates.

Rotation: For 3,099 respondents the scientific literacy test was administered before the computer literacy test and for 3,087 respondents it was the other way around. Test rotation showed, on average, no difference on scientific literacy (main effect = 0.006 logits, Cohen's $d = 0.007$). No items exhibited a noteworthy DIF greater than 0.6 logits. The overall test for DIF using the BIC also favored the main effect model that did not include item-level DIF (see Table 9). The main effect when not allowing for item-level DIF with regard to test rotation are very similar (main effect of -0.043 logits). These results indicate that there was no pronounced DIF concerning test rotation that might have distorted the parameter estimates.

Starting Cohort: There were 2,473 respondents in Starting Cohort 4 and 3,713 respondents in Starting Cohort 6. Three items (sca56120_sc4a14_c, sca140610_c, scs36320_sc4a14_c) showed substantial DIF of more than one logit. These items were treated as unique for the

two starting cohorts in the scaling model, thus, accounting for the DIF in parameter estimates. Differences in scientific literacy occurred between the two starting cohorts, that are in line with age differences, whereas participants in Starting Cohort 6 had, on average, a higher literacy than participants in Starting Cohort 4 (main effect = 0.209 logits, Cohen's $d = 0.253$). In addition to the three items that were treated as unique for the starting cohorts, two items exhibited considerable DIF greater than 0.60 logits (sca50710_sc4a14_c, scg11022s_sc4a14_c) and two item showed small DIF greater than 0.40 logits (scs30940_sc4a14_c, scs36310_sc4a14_c). The overall test for DIF (see Table 9) using the AIC and BIC favored also the DIF model. However, the DIF effects are rather balanced between the starting cohorts and none of the items had to be removed. The main effect when not allowing for item-level DIF with regard to the starting cohorts is a bit lower (main effect of 0.197 logits), indicating that the DIF effects slightly distorted the parameter estimates.

Table 8

Differential Item Functioning

Item		Gender	Migration	Books	Age	Rotation	Starting Cohort
		boys vs. girls	without vs. with	low vs. high	younger vs. older	first vs. second	sc4 vs. sc6
1	sca56120_sc4a14_c_4	0.638 (0.796)	-0.146 (-0.178)	0.532 (0.674)	0.035 (0.043)	-0.408 (-0.496)	0.988 (1.197)
41	sca56120_sc6a14_c_6	0.553 (0.690)	-0.218 (-0.266)	0.409 (0.518)	-0.734 (-0.898)	-0.469 (-0.570)	0.947 (1.147)
2	sca56540_sc4a14_c	-0.028 (-0.035)	0.035 (0.043)	0.141 (0.179)	-0.521 (-0.638)	0.010 (0.012)	-0.544 (-0.659)
3	scs3643s_sc4a14_c	-0.399 (-0.498)	-0.135 (-0.165)	0.260 (0.330)	-0.054 (-0.066)	-0.015 (-0.018)	0.123 (0.149)
4	sca50530_sc4a14_c	0.007 (0.009)	-0.037 (-0.045)	0.087 (0.110)	-0.347 (-0.425)	0.041 (0.050)	-0.306 (-0.371)
5	sca140320_c	0.277 (0.346)	0.416 (0.508)	-0.052 (-0.066)	-0.185 (-0.226)	0.097 (0.118)	-0.107 (-0.130)
6	sca140410_c	0.162 (0.202)	-0.065 (-0.079)	0.092 (0.117)	0.112 (0.137)	0.112 (0.136)	0.066 (0.080)
7	sca14051s_c	-0.280 (-0.350)	0.034 (0.041)	-0.257 (-0.326)	0.482 (0.590)	0.123 (0.150)	0.376 (0.455)
8	sca140610_c_4	-0.436 (-0.544)	-0.044 (-0.054)	-0.136 (-0.172)	0.066 (0.081)	0.109 (0.133)	1.010 (1.223)
42	sca140610_c_6	-0.060 (-0.075)	-0.643 (-0.785)	0.341 (0.432)	0.715 (0.875)	0.094 (0.114)	0.923 (1.118)
9	sca14605s_c	0.585 (0.730)	0.148 (0.181)	-0.217 (-0.275)	-0.203 (-0.248)	0.041 (0.050)	-0.267 (-0.323)
10	sca14101s_c	0.316 (0.394)	0.125 (0.153)	-0.202 (-0.256)	0.204 (0.250)	0.043 (0.052)	0.042 (0.051)
11	sca146060_c	-0.143 (-0.178)	-0.331 (-0.404)	0.221 (0.280)	-0.026 (-0.032)	-0.152 (-0.185)	-0.170 (-0.206)
12	scs36220_sc4a14_c	0.176 (0.220)	-0.083 (-0.101)	0.263 (0.333)	0.086 (0.105)	-0.101 (-0.123)	0.031 (0.038)
13	scg116320_sc4a14_c	-0.635 (-0.793)	0.339 (0.414)	-0.086 (-0.109)	-0.144 (-0.176)	0.194 (0.236)	-0.262 (-0.317)
14	scg110930_sc4a14_c	-0.110 (-0.137)	-0.255 (-0.311)	0.259 (0.328)	-0.077 (-0.094)	0.037 (0.045)	-0.134 (-0.162)
15	scs30920_sc4a14_c	-0.582 (-0.726)	0.036 (0.044)	-0.251 (-0.318)	-0.323 (-0.395)	0.115 (0.140)	-0.278 (-0.337)
16	scs30930_sc4a14_c	-0.772 (-0.964)	-0.201 (-0.245)	-0.144 (-0.183)	-0.013 (-0.016)	-0.049 (-0.060)	-0.137 (-0.166)
17	scs30940_sc4a14_c	-0.521 (-0.650)	0.039 (0.048)	-0.212 (-0.269)	0.539 (0.660)	-0.035 (-0.043)	0.558 (0.676)
18	sca14021s_c	-0.165 (-0.206)	0.114 (0.139)	-0.131 (-0.166)	-0.177 (-0.217)	0.184 (0.224)	-0.300 (-0.363)
19	sca14121s_c	-0.076 (-0.095)	0.227 (0.277)	-0.157 (-0.199)	0.425 (0.520)	-0.031 (-0.038)	0.198 (0.240)
20	sca141310_c	0.451 (0.563)	0.365 (0.445)	-0.151 (-0.191)	-0.042 (-0.051)	0.188 (0.229)	-0.198 (-0.240)
21	sca50210_sc4a14_c	-0.083 (-0.104)	-0.056 (-0.068)	0.059 (0.075)	-0.264 (-0.323)	0.011 (0.013)	-0.235 (-0.285)

	Item	Gender	Migration	Books	Age	Rotation	Starting Cohort
		boys vs. girls	without vs. with	low vs. high	younger vs. older	first vs. second	sc4 vs. sc6
22	sca50220_sc4a14_c	0.096 (0.120)	-0.215 (-0.262)	0.188 (0.238)	-0.383 (-0.469)	0.157 (0.191)	-0.406 (-0.492)
23	sca50710_sc4a14_c	-0.280 (-0.350)	-0.076 (-0.093)	0.172 (0.218)	0.774 (0.947)	0.075 (0.091)	0.684 (0.828)
24	sca56020_sc4a14_c	0.191 (0.238)	0.012 (0.015)	0.045 (0.057)	0.095 (0.116)	-0.012 (-0.015)	0.063 (0.076)
25	sca14609s_c	0.434 (0.542)	-0.154 (-0.188)	-0.001 (-0.001)	-0.120 (-0.147)	0.018 (0.022)	-0.192 (-0.233)
26	scg11021s_sc4a14_c	-0.611 (-0.763)	-0.631 (-0.770)	-0.025 (-0.032)	-0.377 (-0.461)	0.070 (0.085)	-0.386 (-0.468)
27	scg11022s_sc4a14_c	-0.383 (-0.478)	0.310 (0.378)	0.253 (0.321)	0.735 (0.899)	-0.118 (-0.143)	0.851 (1.031)
28	scs36310_sc4a14_c	-0.007 (-0.009)	0.713 (0.870)	-0.490 (-0.621)	-0.448 (-0.548)	-0.191 (-0.232)	-0.500 (-0.606)
29	scs36320_sc4a14_c	-0.104 (-0.130)	-0.147 (-0.179)	-0.348 (-0.441)	0.064 (0.078)	-0.016 (-0.019)	-0.614 (-0.744)
43	scs36320_sc6a14_c	0.059 (0.074)	-0.168 (-0.205)	-0.164 (-0.208)	-0.290 (-0.355)	-0.443 (-0.539)	-0.678 (-0.821)
30	sca50720_sc4a14_c	-0.617 (-0.770)	0.321 (0.392)	0.354 (0.449)	0.401 (0.491)	0.112 (0.136)	0.289 (0.350)
31	sca56320_sc4a14_c	0.060 (0.075)	-0.139 (-0.170)	0.082 (0.104)	-0.571 (-0.699)	-0.093 (-0.113)	-0.548 (-0.664)
32	sca51010_sc4a14_c_m	-0.445 (-0.555)	0.311 (0.380)	0.131 (0.166)	-0.226 (-0.277)	-0.058 (-0.071)	-0.326 (-0.395)
44	sca51010_sc4a14_c_f	0.226 (0.282)	-0.164 (-0.200)	-0.243 (-0.308)	-0.025 (-0.031)	0.036 (0.044)	-0.053 (-0.064)
33	sca140420_c	0.533 (0.665)	0.042 (0.051)	-0.089 (-0.113)	0.121 (0.148)	-0.018 (-0.022)	0.118 (0.143)
34	sca56310_sc4a14_c	-0.237 (-0.296)	-0.122 (-0.149)	0.275 (0.349)	-0.179 (-0.219)	0.007 (0.009)	-0.336 (-0.407)
35	sca5091s_sc4a14_c	0.673 (0.840)	0.008 (0.010)	0.057 (0.072)	0.044 (0.054)	0.075 (0.091)	-0.029 (-0.035)
36	sca14603s_c	0.242 (0.302)	0.116 (0.142)	-0.223 (-0.283)	0.177 (0.217)	0.095 (0.115)	0.052 (0.063)
37	sca14071s_c	0.038 (0.047)	0.087 (0.106)	-0.298 (-0.378)	0.217 (0.266)	0.090 (0.109)	0.025 (0.030)
38	sca14081s_c	-0.036 (-0.045)	-0.518 (-0.632)	0.340 (0.431)	0.290 (0.355)	0.221 (0.269)	0.104 (0.126)
39	sca14608s_c	0.482 (0.602)	0.259 (0.316)	0.120 (0.152)	-0.359 (-0.439)	-0.178 (-0.216)	-0.425 (-0.515)
40	sca14111s_c	0.012 (0.015)	0.217 (0.265)	-0.285 (-0.361)	0.400 (0.489)	0.000 (0.000)	0.229 (0.277)
Main effects							
	DIF model	0.432 (0.539)	0.208 (0.253)	-0.435 (-0.551)	0.369 (0.451)	0.006 (0.007)	0.209 (0.253)
	Main effect model	0.213 (0.270)	0.138 (0.169)	-0.262 (-0.332)	0.230 (0.287)	-0.043 (-0.053)	0.197 (0.243)

Table 9

Comparisons of Models with and without DIF

DIF variable	Model	N	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	6180	207373	69	207511	207975
	DIF model	6180	206037	112	<u>206261</u>	<u>207015</u>
Migration	Main effect	6168	207142	69	207280	<u>207745</u>
	DIF model	6168	207034	112	<u>207258</u>	208011
Books	Main effect	6071	203582	69	203720	204183
	DIF model	6071	203200	112	<u>203424</u>	<u>204176</u>
Age group	Main effect	6186	207558	69	207696	208160
	DIF model	6186	206432	112	<u>206656</u>	<u>207410</u>
Test position	Main effect	6186	207771	69	207909	<u>208373</u>
	DIF model	6186	207634	112	<u>207858</u>	208612
Starting cohort	Main effect	6186	207618	69	207756	208221
	DIF model	6186	206742	112	<u>206966</u>	<u>207720</u>

Note. The best-fitting model according to each information criterion is underlined.

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) 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 discrimination parameters differed moderately between items (see Table 6). The median discrimination parameter fell at 0.799 (*Min* = 0.218, *Max* = 1.687). Model fit indices suggested a better model fit of the GPCM (AIC = 207120, BIC = 207867, number of parameters = 111) as compared to the PCM model (AIC = 207915, BIC = 208372, number of parameters = 68). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.5 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted *Q3* statistics (see last column in Table 6) were quite low, the largest individual residual correlation was 0.075 (item *scs36320_sc4a14_c*). Overall, these results indicate an essentially unidimensional test.

We also examined the dimensionality of the test by specifying two multi-dimensional model based on the two components (*KOS/KAS*) as well as on the three contexts (*health/ technology/ environment*) see Section 2.1. Each item was assigned to one of two or three latent factors (between-item multidimensionality). The multidimensional models were estimated using Quasi Monte Carlo method with 5,000 nodes. The variances and correlations of the three dimensions reflecting the process components are summarized in Table 10. All dimensions

exhibited substantial variances. As expected, the correlation between the two components was with .932 rather high, but a bit lower than a perfect correlation (i.e., $r = .95$, see Carstensen, 2013). Accordingly, the two-dimensional model fit the data slightly better (AIC = 207889, BIC = 208360, number of parameters = 70) than the unidimensional model (AIC = 207915, BIC = 208372, number of parameters = 68). These results indicate that the different components measure a common construct, although they are not perfectly unidimensional. Because the scientific literacy test was constructed to measure a single dimension, a unidimensional competence score was estimated.

Table 10

Results of Two-Dimensional Scaling for the Components

	KOS	KAS
KOS	0.650	
KAS	.932	0.840

Note. Variances of the dimensions are given in the diagonal; correlations are given in the off-diagonal.

6. Discussion

The analyses in the previous sections reported information on the quality of scientific literacy test that was administered in Starting Cohorts 4 and 6 of the NEPS. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, some items exhibited profound differential item functioning with regard to the two starting cohorts and gender. These items were treated as unique for the respective groups, to control for the DIF effects in the estimation of the scientific literacy scores. However, overall, the test had a good reliability and distinguished well between test-takers. In summary, the test had satisfactory psychometric properties that allowed the estimation of a unidimensional scientific literacy score for the participants in both cohorts.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUFs for the two starting cohorts contain 40 items, of which 25 were scored dichotomously (simple multiple-choice items) with 0 indicating an incorrect response and 1 indicating a correct response. The remaining items were scored as polytomous variables (CMC and matching items) that are marked with a 's_c' at the end of the variable names. For further details on the naming conventions of the variables see Fuß and colleagues (2021).

7.2 Linking of Competence Scores

In both starting cohorts, scientific literacy was measured in the current wave in the computer based MST assessment and also in a previous wave in a paper-pencil based assessment. The tests in the different waves included different items and were constructed in such a way as to

allow for an accurate measurement of scientific literacy within the respective age group. As a consequence, the competence scores derived in the different waves 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 waves, the anchor-group linking procedure described in Fischer et al. (2016) was adopted. Following an anchor-group design, an independent link sample including respondents that were not part of the present cohorts were administered a set of items from the previous and the current test of scientific literacy within a single measurement occasion. These responses were used to link the current test to the test that was administered in the previous wave of each starting cohort.

7.2.1 Samples

In Starting Cohort 4, a subsample of 2,390 respondents participated at both measurement occasions, in Wave 3 (i.e., Grade 9; see Schöps & Sass, 2013) and also in Wave 14 (see above) and had at least three valid responses on the test. In Starting Cohort 6, a subsample of 2,426 respondents participated at both measurement occasions, in Wave 3 (i.e., see Haschke et al., 2017) and also in Wave 14 (see above). Consequently, these respondents that participated at both measurement occasions were used to link the two tests across both waves. Moreover, an independent link sample of $N = 501$ students received items from both tests within a single measurement occasion and had at least three valid responses on the tests.

7.2.1.1 Design of the Link Study

The paper-based test administered in Wave 3 of Starting Cohort 4 included 28 items (Schöps & Sass, 2013) and the paper-based test administered in Wave 3 in Starting Cohort 6 included 26 items (Haschke et al., 2017). The computerized MST in Wave 14 of both starting cohorts had 40 items, of which each respondent could receive 23 items. All MST items were administered in the link study. In the link sample, the three tests were administered to all participants. But, because of time limits, only 14 items of the paper-based test in Starting Cohort 4 and 10 of the Starting Cohort 6 were administered. The MST included two items that were also included in the paper-based test of Starting Cohort 6, these items were replaced by item twins with similar item content and item difficulty as the original items, in order to ensure identical item positions of all items. Because no information on the respondents' competence level was available from previous assessments, the easy to medium and medium to difficult level in the first stage of the MST was randomly selected for each respondent.

In the link sample, the items from the different waves were administered at different positions in the test battery. A random sample of 243 respondents received the MST before working on the paper-based tests, whereas 258 respondents worked on the paper-based tests first. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the scientific literacy items in the same order.

The data from the link study was prepared in the same way as in the main studies. As such, the same scoring schemes are used and the same response categories are collapsed for partial credit items. For some items in the link sample the response frequencies were very low, especially in lower response categories. Items with less than 20 responses in a category were excluded from the linking. From the MST test ten items had to be excluded: scs36310_sc4a14_c, scs36320_sc4a14_c, scg11022s_sc4a14_c, sca5091s_sc4a14c, sca14071s_c, sca14071s_c, sca14081s_c, sca14121s_c, sca14608s_c, sca14111s_c, sca51010_sc4a14_c) and from the paper-based test in Starting Cohort 4 three items had to be

excluded (i.e., scg9012s_c, scg9611s_c, scg9083s_c). No item had to be excluded from the paper-based test in Starting Cohort 6.

7.2.1.2 Results

To examine whether the two tests administered in the link sample measured a common scale, 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 computerized MST and each paper-based tests. In Starting Cohort 4, the information criteria favored the two-dimensional model (AIC = 20481, BIC = 20755, number of parameters = 65) in comparison to the unidimensional model (AIC = 20511, BIC = 20777, number of parameters = 63). Similarly, in Starting Cohort 6, the two-dimensional model (AIC = 19799, BIC = 20069, number of parameters = 64) had a slightly better fit than the unidimensional model (AIC = 19815 BIC = 20077, number of parameters = 62). Because the differences in the information criteria between the unidimensional model and the two-dimensional model were small and, therefore, negligible, the results indicate that the scientific literacy tests administered in the two waves 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 cohorts. The differences in item difficulties between the longitudinal sample in Starting Cohort 4 and the link sample for the two tests from Waves 7 and 14 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 11. The minimum effects hypothesis test identified significant ($\alpha = .05$) DIF for eight items. Moreover, the root means squared deviation (RMSD) showed non-negligible DIF with $\text{RMSD} \geq .08$ for twelve additional items. Also, some items exhibited noticeable differences in item difficulties greater than 0.40 logits. Therefore, we selected 6 items from the test administered in Wave 3 with DIF effects that did not exceed 0.40 on the logit scale and 16 items from the MST administered in Wave 14 for the linking

The respective results of the DIF analyses for Starting Cohort 6 are presented in Table 12. Seven of the item exhibited significant DIF according to the minimum effect hypothesis test. In addition, twelve items showed non-negligible DIF with $\text{RMSD} \geq .08$. Therefore, these items were not considered for the linking. As a result, 4 items from the paper-based test administered in Wave 3 and 17 items from the MST administered in Wave 14 were used as common items to link the two tests.

The scientific literacy tests administered in the two waves of each starting cohort were linked using the “mean/mean” method for the anchor-group design using these common items (see Fischer et al., 2016). For Starting Cohort 4, the correction term that was used to link the two waves was calculated as 0.455. The link error reflecting the uncertainty in the linking process was calculated according to Equation 2 in Fischer et al. (2016) as 0.078. The respective calculations for Starting Cohort 6 gave a correction term of -0.343 and a link error of 0.140.

Table 11

Differential Item Functioning Analyses between Starting Cohort 4 and the Link Sample

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Paper based test from Wave 13					
scg90110_c	-0.15	0.11	1.823	0.035	0.026
scg90510_c	-0.05	0.12	0.200	0.007	0.015
scg9052s_c†	0.09	0.07	1.722	0.027	0.094
scg90920_c†	0.66	0.12	30.514*	0.062	0.127
scg90930_c	-0.30	0.12	6.026	0.030	0.045
scg96120_c†	-1.01	0.11	80.172*	0.041	0.162
scg96410_c	-0.01	0.16	0.002	-0.005	-0.009
scg96420_c	0.08	0.12	0.422	0.024	0.015
scg9061s_c	0.01	0.12	0.002	0.016	0.026
scg90630_c†	1.12	0.17	42.228*	0.021	0.126
scg90810_c†	-0.43	0.20	4.596	0.034	0.049
Computerized MST from Wave 14					
sca56120_sc4a14_c†	-0.49	0.11	20.802*	0.034	0.090
sca56540_sc4a14_c	-0.09	0.11	0.643	0.021	0.020
scs3643s_sc4a14_c†	0.30	0.05	37.531*	0.011	0.028
sca50530_sc4a14_c	-0.30	0.16	3.402	0.011	0.079
sca140320_c†	0.62	0.15	16.727	0.044	0.075
sca140410_c†	0.41	0.15	7.406	0.059	0.045
sca14051s_c†	0.37	0.05	54.092*	0.025	0.046
sca140610_c†	0.55	0.11	26.809*	0.053	0.102
sca14605s_c	0.20	0.07	7.307	0.027	0.022
sca14101s_c	0.25	0.07	11.902	0.029	0.047
sca146060_c	0.22	0.15	1.995	0.037	-0.013
scs36220_sc4a14_c	0.01	0.11	0.003	0.010	-0.014
scg116320_sc4a14_c†	-0.31	0.23	1.889	-0.052	-0.189
scg110930_sc4a14_c	-0.07	0.12	0.343	0.032	0.012
scs30920_sc4a14_c†	-0.13	0.23	0.303	-0.075	-0.229
scs30930_sc4a14_c†	-0.01	0.25	0.000	-0.066	-0.266
scs30940_sc4a14_c	0.30	0.11	7.794	0.063	0.059
sca14021s_c†	0.42	0.19	4.621	-0.107	-0.407
sca141310_c†	-0.15	0.24	0.407	-0.040	-0.226
sca50210_sc4a14_c	-0.20	0.12	2.559	0.022	0.040
sca50220_sc4a14_c	-0.28	0.13	4.849	0.044	0.059
sca50710_sc4a14_c	0.40	0.12	10.538	0.015	0.059

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Computerized MST from Wave 14					
sca56020_sc4a14_c	0.10	0.13	0.667	-0.009	-0.023
sca14609s_c	0.02	0.05	0.130	0.034	0.050
scg11021s_sc4a14_c†	-1.60	0.32	25.261*	-3.388	-22.525
sca50720_sc4a14_c	-0.02	0.17	0.014	0.044	0.019
sca56320_sc4a14_c	-0.28	0.16	2.825	0.043	-0.044
sca140420_c†	0.01	0.19	0.002	0.063	0.093
sca56310_sc4a14_c†	-0.43	0.17	6.308	-0.196	0.073
sca14603s_c	0.17	0.07	5.025	0.014	0.019

Note. $\Delta\sigma$ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 7); $SE_{\Delta\sigma}$ = 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 $F_{0154}(1, 499) = 19,865$. A non-significant test indicates measurement invariance. $RMSD$ = Root mean squared deviation for main sample in Starting Cohort 4 (*ms*) or link sample (*ls*).

† items that were used for estimating the link constant, * $p < .05$

Table 12

Differential Item Functioning Analyses between Starting Cohort 6 and the Link Sample

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Paper based test from Wave 3					
sca56130_c†	0.49	0.13	13.669	0.025	0.035
sca51110_c	0.21	0.14	2.339	0.032	0.005
sca51140_c	0.11	0.12	0.839	0.029	0.054
sca50410_c†	0.54	0.13	16.820	0.026	0.040
sca50420_c	0.38	0.11	12.191	0.079	0.047
sca5652s_c†	0.68	0.06	126.512*	0.083	0.235
sca50120_c†	-1.16	0.24	24.174*	0.014	0.085
sca50140_c†	-0.60	0.17	12.011	0.034	0.083
sca51430_c	-0.24	0.14	2.767	0.031	0.074
sca51440_c†	-0.40	0.13	9.030	0.050	0.177
Computerized MST from Wave 14					
sca56120_sc6a14_c†	0.57	0.11	28.016*	0.037	0.109
sca56540_sc6a14_c	0.39	0.11	13.018	0.025	0.07
scs3643s_sc6a14_c†	0.39	0.05	62.531*	0.036	0.049
sca50530_sc6a14_c	-0.05	0.16	0.104	-0.012	0.053
sca140320_sc6a14_c†	0.60	0.14	17.685	0.086	0.091
sca140410_sc6a14_c†	0.23	0.14	2.702	0.086	0.037
sca14051s_sc6a14_c	-0.01	0.05	0.053	0.022	0.032
sca140610_sc6a14_c†	-0.73	0.11	45.041*	0.028	0.136

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Computerized MST from Wave 14					
sca14605s_sc6a14_c†	0.39	0.08	25.239*	0.033	0.067
sca14101s_sc6a14_c	0.17	0.07	5.022	0.049	0.050
sca146060_sc6a14_c	0.23	0.14	2.704	0.013	0.026
scs36220_sc6a14_c	-0.08	0.11	0.598	0.021	0.013
scg116320_sc6a14_c†	0.05	0.23	0.052	-0.052	-0.213
scg110930_sc6a14_c	0.00	0.12	0.000	0.051	-0.016
scs30920_sc6a14_c†	0.30	0.24	1.583	-0.021	-0.248
scs30930_sc6a14_c†	0.18	0.25	0.532	-0.086	-0.309
scs30940_sc6a14_c	-0.36	0.11	10.726	0.021	0.065
sca14021s_sc6a14_c†	0.21	0.20	1.148	-0.329	-0.465
sca141310_sc6a14_c†	-0.00	0.25	0.000	-0.142	-0.264
sca50210_sc6a14_c	-0.06	0.12	0.224	0.017	-0.008
sca50220_sc6a14_c	-0.03	0.12	0.058	0.010	0.018
sca50710_sc6a14_c	-0.39	0.12	10.031	0.015	0.075
sca56020_sc6a14_c	-0.07	0.12	0.269	0.005	-0.018
sca14609s_sc6a14_c	0.11	0.05	5.743	0.024	0.022
scg11021s_sc6a14_c†	-1.50	0.32	21.879*	-672.351	-48.519
sca50720_sc6a14_c	-0.29	0.17	3.012	0.027	0.035
sca56320_sc6a14_c	0.18	0.17	1.142	-0.035	-0.052
sca140420_sc6a14_c†	-0.16	0.19	0.778	-0.031	0.085
sca56310_sc6a14_c	-0.28	0.16	3.015	-0.020	0.041
sca14603s_sc6a14_c	0.03	0.08	0.182	0.014	0.025

Note. $\Delta\sigma$ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 7); $SE_{\Delta\sigma}$ = 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 $F_{0.154}(1, 499) = 19.865$. A non-significant test indicates measurement invariance. $RMSD$ = Root mean squared deviation for main sample in Starting Cohort 4 (ms) or link sample (ls).

† items that were used for estimating the link constant, * $p < .05$

7.3 Scientific Literacy Scores

In the SUF, manifest scientific literacy scores are provided in the form of WLEs (sca14_sc1) including their respective standard error (sca14_sc2). Because the responses for the two starting cohorts were concurrently scaled, these WLEs can be used to compare scientific literacy across starting cohorts in Wave 14. For sca14_sc1u person abilities were estimated using the linked item difficulty parameters. As a result, these WLE scores can be used for longitudinal comparisons between different waves within a starting cohort. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in sca14_sc1 are not linked to the underlying reference scale of the previous wave. However, they are corrected for the position of the scientific literacy test within the test battery. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions.

The R Syntax for estimating the WLEs is provided in Appendix B. In the IRT scaling model, all polytomous variables were scored as 0.5 for each category. For respondents who did not take part in the scientific literacy test or who did not give enough valid responses no WLEs were 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. Plausible values for the scientific literacy test can be estimated using the R package *NEPSscaling* (Scharl et al., 2020; Scharl & Zink, 2022).

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- Blossfeld, H.-P. & Roßbach, H.-G. (2019). Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition *ZfE* (2nd ed.). Springer VS. <https://doi.org/10.1007/978-3-658-23162-0>
- 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). Springer.
- Fischer, L., Rohm, T., Gnamb, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Leibniz Institute for Educational Trajectories. Retrieved from Leibniz Institute for Educational Trajectories website: <https://doi.org/10.5157/NEPS:SP01:1.0>
- Fuß, D., Gnamb, T., Lockl, K., & Attig, M. (2021). *Competence data in NEPS: Overview of measures and variable naming conventions* (Starting Cohorts 1 to 6). Leibniz Institute for Educational Trajectories, National Educational Panel Study.
- Hahn, I. & Kähler, J. (2016): NEPS Technical Report for Science: Scaling Results of Starting Cohort 4 in 11th Grade (NEPS Survey Paper No. 6). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP06:1.0>
- 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. <https://doi.org/10.25656/01:8427>
- Haschke, L. I., Kähler, J. & Hahn, I. (2017): NEPS Technical Report for Science: Scaling Results of Starting Cohort 6 for adults (NEPS Survey Paper No. 19). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP19:1.0>

- 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>
- OECD (2006). Assessing Scientific, Reading and Mathematical Literacy. A Framework for PISA 2006. Paris: OECD. <https://doi.org/10.1787/9789264026407-en>
- Pohl, S., & Carstensen, C. H. (2012). NEPS Technical Report: Scaling the data of the competence tests. NEPS Working Paper 14. University of Bamberg.
<https://doi.org/10.5157/NEPS:WP14:1.0>
- 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://doi.org/10.25656/01:8430>
- Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich [An international comparison of scientific literacy]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme et al. (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63–105). Münster: Waxmann.
- R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute of Education Research.
- Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test analysis modules*. R package version 2.12-18. URL: <https://CRAN.R-project.org/package=TAM>
- Scharl, A., Carstensen, C. H., & Gnamb, 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>

- Scharl, A., & Zink, E. (2022). NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-scale Assessments in Education*, 10:28. <https://doi.org/10.1186/s40536-022-00145-5>
- Schöps, K. & Sass, S. (2013): NEPS Technical Report for Science – Scaling results of Starting Cohort 4 in ninth grade (NEPS Working Paper No. 23). Bamberg: University of Bamberg, National Educational Panel Study.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450. <https://doi.org/10.1007/BF02294627>
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *ZfE*, 14, 67–86. <https://doi.org/10.1007/s11618-011-0182-7>
- Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0*. Acer Press.
- 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. <https://doi.org/10.1177/01466216840080020>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>

Appendix

Appendix A: Variable names and assignment of items to content and process-related component and contexts

	Variable name Starting Cohort 4	Variable name Starting Cohort 6	Stage	Component	Context
1	sca56120_sc4a14_c	sca56120_sc6a14_c	1	KOS	Health
2	sca56540_sc4a14_c	sca56540_sc6a14_c	1	KAS	Health
3	scs3643s_sc4a14_c	scs3643s_sc6a14_c	1	KOS	Technology
4	sca50530_sc4a14_c	sca50530_sc6a14_c	1	KOS	Health
5	sca140320_c	sca140320_sc6a14_c	1	KOS	Environment
6	sca140410_c	sca140410_sc6a14_c	1	KOS	Technology
7	sca14051s_c	sca14051s_sc6a14_c	1	KOS	Technology
8	sca140610_c	sca140610_sc6a14_c	1	KOS	Environment
9	sca14605s_c	sca14605s_sc6a14_c	1	KAS	Environment
10	sca14101s_c	sca14101s_sc6a14_c	1	KOS	Health
11	sca146060_c	sca146060_sc6a14_c	1	KAS	Technology
12	scs36220_sc4a14_c	scs36220_sc6a14_c	2	KAS	Environment
13	scg116320_sc4a14_c	scg116320_sc6a14_c	2	KAS	Technology
14	scg110930_sc4a14_c	scg110930_sc6a14_c	2	KOS	Environment
15	scs30920_sc4a14_c	scs30920_sc6a14_c	2	KOS	Technology
16	scs30930_sc4a14_c	scs30930_sc6a14_c	2	KOS	Technology
17	scs30940_sc4a14_c	scs30940_sc6a14_c	2	KOS	Technology
18	sca14021s_c	sca14021s_sc6a14_c	2	KOS	Technology
19	sca14121s_c	sca14121s_sc6a14_c	2	KOS	Environment
20	sca141310_c	sca141310_sc6a14_c	2	KOS	Health
21	sca50210_sc4a14_c	sca50210_sc6a14_c	2	KOS	Health
22	sca50220_sc4a14_c	sca50220_sc6a14_c	2	KOS	Health
23	sca50710_sc4a14_c	sca50710_sc6a14_c	2	KOS	Environment
24	sca56020_sc4a14_c	sca56020_sc6a14_c	2	KAS	Technology
25	sca14609s_c	sca14609s_sc6a14_c	2	KAS	Technology
26	scg11021s_sc4a14_c	scg11021s_sc6a14_c	3	KOS	Technology
27	scg11022s_sc4a14_c	scg11022s_sc6a14_c	3	KOS	Technology
28	scs36310_sc4a14_c	scs36310_sc6a14_c	3	KOS	Health
29	scs36320_sc4a14_c	scs36320_sc6a14_c	3	KOS	Health
30	sca50720_sc4a14_c	sca50720_sc6a14_c	3	KOS	Environment
31	sca56320_sc4a14_c	sca56320_sc6a14_c	3	KAS	Technology
32	sca51010_sc4a14_c	sca51010_sc6a14_c	3	KOS	Technology
33	sca140420_c	sca140420_sc6a14_c	3	KOS	Technology
34	sca56310_sc4a14_c	sca56310_sc6a14_c	3	KAS	Technology
35	sca5091s_sc4a14_c	sca5091s_sc6a14_c	3	KOS	Health
36	sca14603s_c	sca14603s_sc6a14_c	3	KAS	Environment
37	sca14071s_c	sca14071s_sc6a14_c	3	KOS	Technology
38	sca14081s_c	sca14081s_sc6a14_c	3	KOS	Environment
39	sca14608s_c	sca14608s_sc6a14_c	3	KAS	Environment
40	sca14111s_c	sca14111s_sc6a14_c	3	KOS	Environment

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

Appendix B: R-Syntax for estimating WLEs in Starting Cohorts 4 and 6

```
# load packages
library(rio)    # to import SPSS files
library(doby)   # recode variables
library(TAM)    # for IRT analyses

# load competence data
dat <- rio::import("SUF for competencies.sav")

# 40 items of the science test
items <- c("sca56120_sc4a14_c", "sca56540_sc4a14_c",
          "scs3643s_sc4a14_c", "sca50530_sc4a14_c",
          ...)

# polytomous items
poly <- c("scs3643s_sc4a14_c", "sca14051s_c",
          "sca14605s_c", "sca14101s_c",
          "sca14021s_c", "sca14121s_c",
          "sca14609s_c", "scg11021s_sc4a14_c",
          "scg11022s_sc4a14_c", "sca5091s_sc4a14_c",
          "sca14603s_c", "sca14071s_c",
          "sca14081s_c", "sca14608s_c",
          "sca14111s_c")

# collapse response categories
dat$sca14021s_c <- doBy::recodeVar(
  dat$sca14021s_c,
  c(0, 1, 2, 3),
  c(0, 0, 1, 2)
)

dat$sca14121s_c <- doBy::recodeVar(
  dat$sca14121s_c,
  c(0, 1, 2),
  c(0, 1, 1)
)

dat$sca14609s_c <- doBy::recodeVar(
  dat$sca14609s_c,
  c(0, 1, 2, 3, 4),
  c(0, 1, 2, 3, 3)
)
```

```
dat$scg11021s_sc4a14_c <- doBy::recodeVar(  
  dat$scg11021s_sc4a14_c,  
  c(0, 1, 2, 3, 4),  
  c(0, 0, 0, 0, 1)  
)
```

```
dat$scg11022s_sc4a14_c <- doBy::recodeVar(  
  dat$scg11022s_sc4a14_c,  
  c(0, 1, 2, 3, 4),  
  c(0, 0, 0, 0, 1)  
)
```

```
dat$sca5091s_sc4a14_c <- doBy::recodeVar(  
  dat$sca5091s_sc4a14_c,  
  c(0, 1, 2, 3, 4),  
  c(0, 0, 1, 2, 3)  
)
```

```
dat$sca14603s_c <- doBy::recodeVar(  
  dat$sca14603s_c,  
  c(0, 1, 2, 3),  
  c(0, 0, 1, 2)  
)
```

```
dat$sca14081s_c <- doBy::recodeVar(  
  dat$sca14081s_c,  
  c(0, 1, 2, 3, 4),  
  c(0, 0, 0, 0, 1)  
)
```

```
dat$sca14608s_c <- doBy::recodeVar(  
  dat$sca14608s_c,  
  c(0, 1, 2, 3, 4, 5, 6),  
  c(0, 0, 0, 0, 1, 2, 2)  
)
```

```
dat$sca14111s_c <- doBy::recodeVar(  
  dat$sca14111s_c,  
  c(0, 1, 2, 3, 4),  
  c(0, 0, 1, 2, 3)  
)  
  
# define Q-matrix for 0.5 scoring of PCM  
Q <- matrix(1, nrow = length(items), ncol = 1)  
Q[items %in% poly, 1] <- 0.5    # score of 0.5  
  
# estimate partial credit model  
mod <- TAM::tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2",  
  pid = dat$ID_t)  
  
summary(mod)  
  
# item fit  
TAM::tam.fit(mod)  
  
# WLE  
TAM::tam.wle(mod)
```